Uncertainty Quantification in Data-Driven Simulation and Optimization: Statistical and Computational Efficiency

Huajie Qian

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

ABSTRACT

Uncertainty Quantification in Data-Driven Simulation and Optimization: Statistical and Computational Efficiency

Huajie Qian

Models governing stochasticity in various systems are typically calibrated from data, therefore are subject to statistical errors/uncertainties which can lead to inferior decision making. This thesis develops statistically and computationally efficient data-driven methods for problems in stochastic simulation and optimization to quantify and hedge impacts of these uncertainties.

The first half of the thesis focuses on efficient methods for tackling input uncertainty which refers to the simulation output variability arising from the statistical noise in specifying the input models. Due to the convolution of the simulation noise and the input noise, existing bootstrap approaches consist of a two-layer sampling and typically require substantial simulation effort. Chapter 2 investigates a subsampling framework to reduce the required effort, by leveraging the form of the variance and its estimation error in terms of the data size and the sampling requirement in each layer. We show how the total required effort is reduced, and explicitly identify the procedural specifications in our framework that guarantee relative consistency in the estimation, and the corresponding optimal simulation budget allocations. In Chapter 3 we study an optimization-based approach to construct confidence intervals for simulation outputs under input uncertainty. This approach computes confidence bounds from simulation runs driven by probability weights defined on the data, which are obtained from solving optimization problems under suitably posited averaged divergence constraints. We illustrate how this approach offers benefits in computational efficiency and finite-sample performance compared to the bootstrap and the delta method. While resembling distributionally robust optimization, we explain the procedural design and develop tight statistical guarantees via a generalization of the empirical likelihood method.

The second half develops uncertainty quantification techniques for certifying solution feasibility and optimality in data-driven optimization. Regarding optimality, Chapter 4 proposes a statistical method to estimate the optimality gap of a given solution for stochastic optimization as an assessment of the solution quality. Our approach is based on bootstrap aggregating, or bagging, resampled sample average approximation (SAA). We show how this approach leads to valid statistical confidence bounds for non-smooth optimization. We also demonstrate its statistical efficiency and stability that are especially desirable in limited-data situations. We present our theory that views SAA as a kernel in an infinite-order symmetric statistic. Regarding feasibility, Chapter 5 considers data-driven optimization under uncertain constraints, where solution feasibility is often ensured through a "safe" reformulation of the constraints, such that an obtained solution is guaranteed feasible for the oracle formulation with high confidence. Such approaches generally involve an implicit estimation of the whole feasible set that can scale rapidly with the problem dimension, in turn leading to over-conservative solutions. We investigate validation-based strategies to avoid set estimation by exploiting the intrinsic low dimensionality of the set of all possible solutions output from a given reformulation. We demonstrate how our obtained solutions satisfy statistical feasibility guarantees with light dimension dependence, and how they are asymptotically optimal and thus regarded as the least conservative with respect to the considered reformulation classes.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This thesis concludes my Ph.D. research over the past five years. I am indebted to many outstanding scholars and individuals who have guided me through the rewarding journey.

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Henry Lam, for his enlightening, persistent, and comprehensive guidance throughout these years. He has been constantly offering insightful advice with his enduring patience and broad knowledge in simulation, optimization, statistics and beyond. As a statistician as well as an operations researcher, his diverse perspectives on research questions have greatly helped me transition from a mathematical background to operations research. He has also been extremely approachable and considerate, which has made my Ph.D. study such a smooth and enjoyable experience. I have learnt from him a lot on how to do research and be a scholar. All these have set him a role model for me in pursuit of academic excellence.

I would like to thank Professor Jing Dong, Garud Iyengar, David Yao, and Xunyu Zhou for serving on my thesis committee. Their constructive feedback and comments have led to great improvement of this thesis in terms of both presentation and material. I have learnt much from the insightful discussions with them.

I also want to extend my appreciation to many exceptional professors with whom I have taken courses that lay the foundation for the research works in this thesis. I would like to thank Professor Jinho Baik and David Yao for probability and stochastics, Professor Long Nguyen and Ambuj Tewari for statistical and machine learning, Professor John Duchi and Ruiwei Jiang for theories and algorithms for stochastic and robust optimizaton, and Professor Roman Vershynin for high dimensional probability theory. I also thank Professor Kristen Moore and Virginia Young for their guidance on stochastic control problems in mathematical finance, and Professor Donald Lee and

David Yao for introducing impactful applications of statistical learning.

My sincere thanks also go to many friends and fellow students at Columbia University and University of Michigan, and those I've luckily met at conferences. I thank all of you for the help I've received in both research and daily life, and for the many casual chats and academic discussions that have contributed to this thesis in one way or another.

Lastly and most importantly, I thank my parents, Huifa Qian and Jinmei Fan, and my wife, Xiaochen Wang, for their unconditional love and support. Besides emotional support, I have learnt much about statistics from my wife, who recently earned her Ph.D. in biostatistics.

Dedicated to my father, Huifa Qian, and my wife, Xiaochen Wang.

In loving memory of my mother, Jinmei Fan.

Chapter 1

Introduction

In the data-rich era, decision making under uncertainty often relies on inference of unknown stochasticity from real-world data. A common concern, however, is that the model and statistical errors from the data may not be properly controlled when integrating into the downstream simulation and optimization tasks, thus leading to inferior decisions. Therefore a quantitative understanding of the statistical uncertainties is crucial in guarding against catastrophic decision making. Broadly speaking, this has stimulated interests across multiple research communities, and various approaches have been proposed to handle statistical uncertainties for different kinds of problems, such as uncertainty sets in (distributionally) robust opotimization, penalties in regularized risk minimization, and upper confidence bound (UCB) algorithms in reinforcement learning. This thesis instead investigates uncertainty quantification methods for two commonly used tools in operations research, i.e., stochastic simulation (in Chapters 2 and 3) and optimization under uncertainty (in Chapters 4 and 5), and focuses on statistical and/or computational efficiencies of these methods.

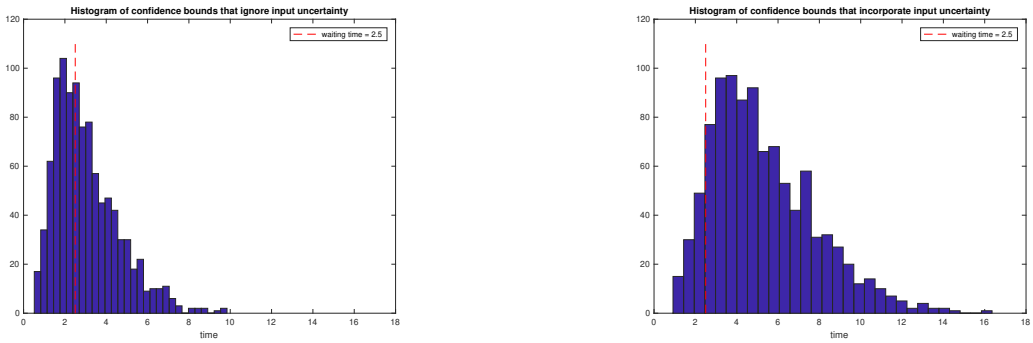
1.1 Stochastic Simulation under Input Uncertainty

The first part (Chapters 2 and 3) of the thesis is on efficient methods for tackling *input uncertainty* in stochastic simulation. Stochastic simulation has been used routinely to assess and optimize performances of stochastic operational systems. In conventional simulation output analysis, the

underlying input models are assumed completely known or given by expert opinions, and simulation outputs generated from these input models are used to make statistical inference on the performance metric of interest. In a data-driven setting, however, input models are estimated from data to drive simulation, and input uncertainty arises due to the propagation of the input estimation errors to the output. Therefore, statistically valid inference and performance prediction require careful incorporation of model errors on top of the stochastic computation noises in the Monte Carlo simulation.

To further illustrate the necessity of tackling input uncertainty, consider an M/M/1 queue with arrival rate 0.8 and service rate 1.0, and the performance measure of interest is the mean waiting time of the first 20 arrivals (true value ≈ 2.57). Suppose that the true arrival and service rates are unknown and can only be estimated from data of inter-arrival times and service times, each of size 50, therefore input uncertainty is present. Suppose that a criterion in designing the queuing system is that the mean waiting time must be no longer than 2.5 units of time (the current design is infeasible). We compare two approaches to assessing feasibility of the current design, both involving the construction of upper confidence bounds for the target quantity. In the first approach, an arrival rate and a service rate are estimated from the data, and then treated as the truth to drive the simulation to obtain a 95%-level performance bound based on 500 replications. The experiment is then repeated on 1000 independent input data sets, and the distribution of the obtained performance bounds are shown in Figure 1.1a. The second approach, however, acknowledges the statistical errors in the estimated input models, and incorporate them in constructing the performance bounds. The results are in Figure 1.1b. We observe that when input uncertainty is ignored the obtained bounds often (44%) fall below the threshold 2.5, rendering a substantial chance of incorrect feasibility assessment, whereas after incorporating input uncertainty misassessment happens much less frequently (10%). Quantification of input uncertainty therefore is essential for correctly hedging the total risk in the output.

There are several challenges in quantifying input uncertainty. The first is the computational demand in disentangling the statistical noise in calibrating the input model from the Monte Carlo noise. Previous approaches to this problem such as the bootstrap (Barton and Schruben (1993,



(a) Input uncertainty ignored.

(b) Input uncertainty taken into account.

Figure 1.1: Distribution of upper confidence bounds, relative to the threshold 2.5.

2001), Cheng and Holland (1997)) require a substantial computation effort because of the need to conduct multi-layer nested simulation and consequently a multiplicatively growing size of simulation replications. Secondly, approaches based-on the delta method (e.g., Chapter 3 in Asmussen and Glynn (2007) and Cheng and Holland (1997, 1998)) construct interval estimates from a linearization of the performance metric and an estimation of the standard error term, which tend to undercover the true performance metric under small input data. Chapters 2 and 3, respectively, are devoted to addressing these challenges.

Chapter 2 develops a *subsampling* technique that significantly reduces the order of computation in each layer of the nested simulation, by leveraging and properly rescaling the standard error arising from input uncertainty according to the subsample size parameters. The proposed method provably allows the simulation cost to grow independently of the data size, in contrast to the standard bootstrap where the required simulation burden has to grow linearly, thus making our method more attractive when each simulation run is computationally expensive or simulation resources are limited. We also derive the optimal algorithmic configurations, regarding choices of the subsample size and the simulation sizes to allocate to each layer, that achieve the minimum error in estimating the input uncertainty under a fixed simulation budget, by balancing a trade-off between a Monte Carlo (computational) error and a statistical error.

In Chapter 3 we propose an optimization-based approach that computes interval estimates as the optimal values of suitably posited optimization problems which do not rely on linearization.

Our formulation is based on an "empirical" version of *distributionally robust optimization* (DRO). The latter is a decision-making framework for stochastic problems where the underlying distribution is not fully known, which advocates the search of the best solution over the worst-case scenario. Our formulation constructs interval estimates by optimizing the performance metric over a set of distributions that are supported on the input data, satisfying a suitably weighted Kullback-Leibler divergence constraint. We demonstrate how our approach can conform naturally to the numerical boundary of the performance metric and leads to better finite-sample coverage than linearization-based interval estimates. Moreover, we develop tight coverage guarantees via a generalization of the *empirical likelihood* theory, in contrast to potentially loose confidence guarantees in previous data-driven DRO formulations.

1.2 Uncertainty Quantification in Data-Driven Optimization

In the second part (Chapters 4 and 5) we switch focus to uncertainty quantification for data-driven optimization. Stochastic optimization has been extensively used for decision making under uncertainty in both operations research and machine learning, where the decision maker optimizes a certain expected performance measure, potentially subject to uncertain constraints. In the context where the governing distributions are estimated from data, Chapters 4 and 5 investigate statistically efficient methodologies to assess and improve solution performances in terms of optimality and feasibility.

Chapter 4 presents a novel method based on *bagging* or *bootstrap aggregating*, an ensemble method in machine learning, to compute bounds for the optimality gap of a given solution. The motivation is that data-driven solutions to stochastic optimization can be suboptimal due to contamination from statistical and model errors, and a quantitative assessment of solution quality can help with screening out inferior solutions. The goal here is to assess solution performance by using data; this is in contrast to the common analyses of stochastic optimization algorithms that reveal the convergence rate, which are based on the worst-case and could be over-conservative for a given particular problem instance. Existing methods based on data batching (Mak et al. (1999)) tend

to generate unnecessarily loose bounds due to the inefficient use of the data, while those based on *sample average approximation* (SAA) asymptotics (Shapiro et al. (2014), Bayraksan and Morton (2006)) require Lipschitz smoothness from the optimization and can perform poorly in practice due to the instability in estimating the standard error. The proposed bagging method reduces the estimation variance of optimality gap bounds and stabilizes estimation of the standard error by averaging a large number of resampled estimates, and at the same time extends the SAA asymptotic theories to non-smooth problems by smoothing the SAA optimal values. Mathematically, we established the asymptotic performance of our bagging approach by utilizing the so-called infinite-order symmetric statistics, in which the SAA optimal value can be viewed as the kernel of the corresponding statistics.

Chapter 5 focuses on improving data-driven solutions for optimization under uncertain constraints, such as probabilistic or expectation constraints. When these constraints are only observable via data, feasibility can only be guaranteed at best with high confidence, and a data-driven procedure needs to strike a balance between optimality and feasibility. Common data-driven formulations, such as DRO, SAA, and *robust optimization*, ensure feasibility guarantees via a feasible set estimation, or in other words, an implicit simultaneous estimation problem of the noisy constraint over the whole decision space. This could subsequently lead to over-conservative solutions especially for high dimensional problems. To address this issue, we develop a general constraint-validation framework that allows one to examine feasibility only on a low dimensional solution path that is intrinsic to these common data-driven optimization formulations. We establish both asymptotic and finite-sample performance guarantees of our framework, and dissect our results to various formulations, by using recently developed *high-dimensional Berry-Esseen theorem* and *empirical process* theory.

In the remainder of the thesis, Chapters 2-5 present in detail the four projects mentioned above, and Appendices A-D contain technical proofs for each of the chapters respectively. As an effort to improve the manageability of the notation system, mathematical symbols will be made self-contained within each chapter, in other words, a symbol that refers to a certain object in one chapter may be used to represent a different object in another chapter. The thesis is based on Lam

and Qian (2018c, 2017, 2018b, 2019a), for which preliminary versions have appeared in Lam and Qian (2018d, 2016, 2018a, 2019b).

Chapter 2

Subsampling to Enhance Efficiency in Input Uncertainty Quantification

2.1 Introduction

Stochastic simulation is one of the most widely used analytic tools in operations research. It provides a flexible means to approximate complex models and to inform decisions. See, for instance, Law et al. (2000) and Banks et al. (2005) for applications in manufacturing, revenue management, service and operations systems etc. In practice, the simulation platform relies on input models that are typically observed or calibrated from data. These statistical noises can propagate to the output analysis, leading to significant errors and suboptimal decision-making. In the literature, this problem is commonly known as input uncertainty or extrinsic uncertainty.

In conventional simulation output analysis where the input model is completely pre-specified, the statistical errors come solely from the Monte Carlo noises, and it suffices to account only for such noises in analyzing the output variability. When input uncertainty is present, such an analysis will undermine the actual variability. One common approach to quantify the additional uncertainty is to estimate the variance in the output that is contributed from the input noises (e.g., Song et al. (2014)); for convenience, we call this the *input variance*. This quantity acts as an uncertainty measure which, when added together with the Monte Carlo variance, gives rise to the

overall variance in the outputs. A refined decomposition of input variance across multiple input sources can be used to identify models that are overly ambiguous and flag the need of more data collection (e.g., Song et al. (2014)). Input variance also provides a building block to construct valid output confidence intervals (CIs) that account for combined input and simulation errors (e.g., Cheng and Holland (2004)). Motivated by its central role in quantifying input uncertainty, this chapter aims to study the efficient estimation of input variance.

In the literature, bootstrap resampling is a common approach for the above purpose. This applies most prominently in the nonparametric regime, namely when no assumptions are placed on the input parametric family. It could also be used in the parametric case (where more alternatives are available). For example, Cheng and Holland (1997) proposes the variance bootstrap, and Song and Nelson (2015) studies the consistency of this strategy on a random-effect model that describes the uncertainty propagation. A bottleneck with using bootstrap resampling in estimating input variances, however, is the need to "outwash" the simulation noise, which often places substantial burden on the required simulation effort. More precisely, to handle both the input and the simulation noises, the bootstrap procedure typically comprises a two-layer sampling that first resamples the input data (i.e., *outer* sampling), followed by running simulation replications using each resample (i.e., *inner* replications). Due to the reciprocal relation between the magnitude of the input variance and the input data, the input variance becomes increasingly small as the input data size increases. This deems the control of the relative estimation error increasingly expensive, and requires either a large outer bootstrap size or inner replication size to extinguish the effect of simulation noises.

The main goal of this chapter is to investigate subsampling as a simulation saver for input variance estimation. This means that, instead of creating distributions by resampling a data set of the full size, we only resample (with or without replacement) a set of smaller size. We show that a judicious use of subsampling can reduce the total simulation effort from an order bigger than the data size in the conventional two-layer bootstrap to an order independent of the data size, while retaining the estimation accuracy. This approach leverages the interplay between the form of the input variance and its estimation error, in terms of the data size and the sampling

effort in each layer of the bootstrap. On a high level, the subsample is used to estimate an input variance as if less data are available, followed by a correction of this discrepancy in the data size by properly rescaling the input variance. We call this approach *proportionate subsampled variance bootstrap*. We explicitly identify the procedural specifications in our approach that guarantee estimation consistency, including the minimally required simulation effort in each layer. We also study the theoretical behavior of our estimation error, in relation to the simulation effort allocation in these layers as well as the input data and subsample sizes, which in turn reveals the optimal configurations and provides implementation guidance.

In the statistics literature, subsampling has been used as a remedy for situations where the full-size bootstrap does not apply, due to a lack (or undeterminability) of uniform convergence required for its statistical consistency, which relates to the functional smoothness or regularity of the estimators (e.g., Politis and Romano (1994)). Subsampling has been used in time series and dependent data (e.g., Politis et al. (1999), Hall et al. (1995), Datta and McCormick (1995)), extremal estimation (e.g., Bickel and Sakov (2008)), shape-constrained estimation (e.g., Sen et al. (2010)) and other econometric contexts (e.g., Abadie and Imbens (2008), Andrews and Guggenberger (2009, 2010)). In contrary to these works, our subsampling approach is introduced to reduce the simulation effort faced by the two-layer sampling necessitated from the presence of both the input and simulation noises. In other words, we are not concerned about the issue of uniform convergence, but instead, we aim to distort the relation between the required simulation effort and data size in a way that allows more efficient deconvolution of the effects of the two noises. We also note that, as we will use resampling with replacement (instead of without replacement), our approach is closer to the so-called m out of n bootstrap (Bickel et al. (1997), Bickel and Sakov (2008)). For coherence, throughout the chapter we use the term subsampling broadly to indicate a bootstrap with a smaller resample size than the original data size.

We close this introduction with a brief review of other related work in input uncertainty. In the nonparametric regime (the focus of this chapter), besides Cheng and Holland (1997) and Song and Nelson (2015) that study bootstrap-based estimation of the input variance, Barton and Schruben (1993) and Barton and Schruben (2001) investigate the percentile bootstrap to construct CIs (i.e.,

the CI limits are determined from the quantiles of the bootstrap distributions). Like variance bootstrap, percentile bootstrap also encounters two-layer sampling that requires substantial simulation efforts. Yi and Xie (2017) investigates adaptive budget allocation policies based on ranking and selection to reduce simulation cost in the percentile bootstrap, and empirically shows the computational advantage of their approach. On the other hand, contrary to this work, they do not investigate the required simulation efforts in relation to the input data size. Lam and Qian (2016, 2017) study the use of empirical likelihood as an optimization-based alternative to the percentile bootstrap, which requires simulation efforts to estimate the gradient information that remain substantial. Beyond the frequentist regime considered in this chapter, Xie et al. (2018) studies nonparametric Bayesian methods based on Dirichlet process mixtures to estimate the variance contributed from input uncertainty and construct CIs. Glasserman and Xu (2014), Hu et al. (2012), Lam (2016b) and Ghosh and Lam (2019) study input uncertainty from a robust optimization viewpoint, where they compute worst-case bounds subject to constraints or so-called uncertainty sets that represent partial beliefs on unknown distributions. In the parametric regime, Barton et al. (2013) and Xie et al. (2016) investigate the basic bootstrap with a metamodel built in advance, a technique known as the metamodel-assisted bootstrap. Cheng and Holland (1997) studies the delta method, and Cheng and Holland (1998, 2004) reduce its computation burden via the so-called two-point method. Lin et al. (2015) and Song and Nelson (2019) study regression approaches to estimate sensitivity coefficients which are used to apply the delta method, generalizing the gradient estimation method in Wieland and Schmeiser (2006). Zhu et al. (2020) studies risk criteria and computation to quantify parametric uncertainty. Finally, Chick (2001), Zouaoui and Wilson (2003), Zouaoui and Wilson (2004) and Xie et al. (2014) study variance estimation and interval construction from a Bayesian perspective. We comment that although the exposition in this chapter focuses on the nonparametric setting, the same idea of subsampling can be adapted naturally to the parametric setting, with similar advantages in computational efficiency. For general surveys on input uncertainty, readers are referred to Barton et al. (2002), Henderson (2003), Chick (2006), Barton (2012), Song et al. (2014), Lam (2016a), and Nelson (2013) Chapter 7.

The remainder of this chapter is as follows. Section 2.2 introduces the input uncertainty problem

and explains the simulation complexity bottleneck in the existing bootstrap schemes. Section 2.3 presents our subsampling idea, procedures and the main statistical results. Section 2.4 discusses the key steps in our theoretical developments. Section 2.5 reports our numerical experiments. All proofs are relegated to Appendix A.

2.2 Problem Motivation

This section describes the problem and our motivation. Section 2.2.2 first describes the input uncertainty problem, Section 2.2.3 presents the existing bootstrap approach, and Section 2.2.4 discusses its computational barrier, thus motivating our subsampling investigation. We aim to provide intuitive explanations in this section, and defer mathematical details to later sections.

2.2.1 Notation

We use the following notations. For any sequences a_n and b_n, both depending on n, we say that $a_n = O(b_n)$ if $|a_n/b_n| \leq C$ for some constant $C > 0$ for all sufficiently large n, and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. Alternately, we say $a_n = \Omega(b_n)$ if $|a_n/b_n| \geq C$ for some constant $C > 0$ for all sufficiently large n, and $a_n = \omega(b_n)$ if $|a_n/b_n| \to \infty$ as $n \to \infty$. We say that $a_n = \Theta(b_n)$ if $\underline{C} \leq |a_n/b_n| \leq \overline{C}$ as $n \to \infty$ for some constants $\underline{C}, \overline{C} > 0$. We use $A_n = O_p(b_n)$ to represent a sequence of random variables A_n that has stochastic order at least b_n, i.e., for any $\epsilon > 0$, there exists $M, N > 0$ such that $P(|A_n/b_n| \leq M) > 1 - \epsilon$ for $n > N$. We use $A_n = o_p(b_n)$ to represent a sequence of random variables A_n that has stochastic order less than b_n, i.e., $A_n/b_n \xrightarrow{p} 0$. We use $A_n = \Theta_p(b_n)$ to represent a sequence A_n that has stochastic order exactly at b_n, i.e., A_n satisfies $A_n = O_p(b_n)$ but not $A_n = o_p(b_n)$.

2.2.2 The Input Uncertainty Problem

Suppose there are m independent input processes driven by input distributions $F_1, F_2, \ldots, F_m$. We consider a generic performance measure $\psi(F_1, \ldots, F_m)$ that is simulable, i.e., given the input distributions, independent unbiased replications of ψ can be generated in a computer. As a primary

example, think of F_1 and F_2 as the interarrival and service time distributions in a queue, and ψ is some output measure such as the mean queue length averaged over a time horizon.

The input uncertainty problem arises in situations where the input distributions $F_1, \ldots, F_m$ are unknown but real-world data are available. One then has to use their estimates $\widehat{F}_1, \ldots, \widehat{F}_m$ to drive the simulation. Denote a point estimate of $\psi(F_1, \ldots, F_m)$ as $\bar{\psi}(\widehat{F}_1, \ldots, \widehat{F}_m)$, where typically we take

$$\bar{\psi}(\widehat{F}_1, \ldots, \widehat{F}_m) = \frac{1}{q} \sum_{r=1}^{q} \hat{\psi}_r(\widehat{F}_1, \ldots, \widehat{F}_m)$$

with $\hat{\psi}_r(\widehat{F}_1, \ldots, \widehat{F}_m)$ being a conditionally unbiased simulation replication driven by $\widehat{F}_1, \ldots, \widehat{F}_m$. This point estimate is affected by both the input statistical noises and the simulation noises. By conditioning on the estimated input distributions (or viewing the point estimate as a random effect model with uncorrelated input and simulation noises), the variance of $\bar{\psi}(\widehat{F}_1, \ldots, \widehat{F}_m)$ can be expressed as

$$\mathrm{Var}[\bar{\psi}(\widehat{F}_1, \ldots, \widehat{F}_m)] = \sigma_I^2 + \sigma_S^2$$

where

$$\sigma_I^2 = \mathrm{Var}[\psi(\widehat{F}_1, \ldots, \widehat{F}_m)] \tag{2.1}$$

is the input variance, and

$$\sigma_S^2 = \frac{\mathbb{E}[\mathrm{Var}[\hat{\psi}_r(\widehat{F}_1, \ldots, \widehat{F}_m) | \widehat{F}_1, \ldots, \widehat{F}_m]]}{q}$$

is the variance contributed from the simulation noises. Assuming that the estimates $\widehat{F}_i$'s are consistent in estimating F_i's, then, as n_i grows, σ_S^2 is approximately $\mathrm{Var}[\hat{\psi}_r(F_1, \ldots, F_m)]/q$ and can be estimated by taking the sample variance of all simulation replications (see, e.g., Cheng and Holland (1997)). On the other hand, σ_I^2 signifies the output variance contributed solely from the input data noises, assuming a fully accurate evaluation of the performance measure ψ. Estimating σ_I^2 is the key and the challenge in quantifying input uncertainty, which is the focus of this chapter.

Before going into details, we discuss two conceptual properties on σ_I^2 that would be relevant in motivating and pinpointing our study. First, suppose further that for each input model i, we have

n_i i.i.d. data $\{X_{i,1}, \ldots, X_{i,n_i}\}$ generated from the distribution F_i. When n_i's are large, typically the overall input variance σ_I^2 is decomposable into

$$\sigma_I^2 \approx \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} \tag{2.2}$$

where σ_i^2/n_i is the variance contributed from the data noise for model i, with σ_i^2 being a constant. In the parametric case where $\widehat{F}_i$ comes from a parametric family containing the estimated parameters, this decomposition is well known from the delta method (Asmussen and Glynn (2007), Chapter 3). Here, σ_i^2/n_i is typically $\nabla_i \psi' \Sigma_i \nabla_i \psi$, where $\nabla_i \psi$ is the collection of sensitivity coefficients, i.e., the gradient, with respect to the parameters in model i, and Σ_i is the asymptotic estimation variance of the point estimates of these parameters (scaled reciprocally with n_i). In the nonparametric case where the empirical distribution $\widehat{F}_i(x) := \sum_{j=1}^{n_i} \delta_{X_{i,j}}(x)/n_i$ is used (where $\delta_{X_{i,j}}$ denotes the delta measure at $X_{i,j}$), (2.2) still holds under mild conditions (e.g., Propositions 2.4.1 and 2.4.6 in the sequel). In this setting the quantity σ_i^2 is equal to $\mathrm{Var}_{F_i}[g_i(X_i)]$, where $g_i(\cdot)$ is the influence function (Hampel (1974)) of ψ with respect to the distribution F_i, whose domain is the value space of the input variate X_i, and $\mathrm{Var}_{F_i}[\cdot]$ denotes the variance under F_i. The influence function can be viewed as a functional derivative taken with respect to the probability distributions F_i's (see Serfling (2009), Chapter 6), and dictates the first-order asymptotic behavior of the plug-in estimate of ψ. Although the mathematical form of σ_i^2's is known, it relies on gradient information that needs to be estimated via simulation itself. Moreover, in the nonparametric case, the gradient dimension in a sense grows with the data size. Thus directly using the delta method in this case could be challenging. In our subsequent developments, we focus on the nonparametric case, both because this is more challenging, and also that this can be viewed as a generalization of the parametric case by viewing the "parameter" simply as a function of F_i's.

Second, under further regularity conditions, a Gaussian approximation holds for $\bar{\psi}(\widehat{F}_1, \ldots, \widehat{F}_m)$ so that

$$\bar{\psi}(\widehat{F}_1, \ldots, \widehat{F}_m) \pm z_{1-\alpha/2} \sqrt{\sigma_I^2 + \sigma_S^2} \tag{2.3}$$

is an asymptotically tight $(1-\alpha)$-level CI for $\psi(F_1, \ldots, F_m)$, where $z_{1-\alpha/2}$ is the standard normal

$1-\alpha/2$ quantile. This CI, which provides a bound-based alternative to quantify input uncertainty, again requires a statistically valid estimate of σ_I^2 or $\sum_{i=1}^m \sigma_i^2/n_i$ (and σ_S^2). In this chapter we primarily focus on the estimation of σ_I^2 and how our proposed approach substantially improves upon previous methods in this regard. Naturally, the improved estimate of σ_I^2 also translates into a better CI when using (2.3). We caution, however, that an optimal procedural configuration to estimate σ_I^2 does not necessarily correspond to an optimal configuration in constructing the CI, as the performance of the latter is measured by different criteria such as coverage or half-width (such a difference in optimally estimating variance versus CI has also been observed in other contexts such as time series (Sun et al. (2008))). Nonetheless, we will show that a direct plug-in of our new estimator of σ_I^2 into (2.3) is already enough to significantly outperform conventional bootstrap-based CIs suggested in the literature, both theoretically and also supported by consistent empirical evidence.

Next we will discuss bootstrap resampling, the commonest estimation technique that forms the basis of our comparison.

2.2.3 Bootstrap Resampling

Let $\widehat{F}_i^*$ represent the empirical distribution constructed using a bootstrap resample from the original data $\{X_{i,1},\ldots,X_{i,n_i}\}$ for input F_i, i.e., n_i points drawn by uniformly sampling with replacement from $\{X_{i,1},\ldots,X_{i,n_i}\}$. The bootstrap variance estimator is $\mathrm{Var}_*[\psi(\widehat{F}_1^*,\ldots,\widehat{F}_m^*)]$, where $\mathrm{Var}_*[\cdot]$ denotes the variance over the bootstrap resamples from the data, conditional on $\widehat{F}_1,\ldots,\widehat{F}_m$.

The principle of bootstrap entails that $\mathrm{Var}_*[\psi(\widehat{F}_1^*,\ldots,\widehat{F}_m^*)] \approx \mathrm{Var}[\psi(\widehat{F}_1,\ldots,\widehat{F}_m)] = \sigma_I^2$. Here $\mathrm{Var}_*[\psi(\widehat{F}_1^*,\ldots,\widehat{F}_m^*)]$ is obtained from a (hypothetical) infinite number of bootstrap resamples and simulation runs per resample. In practice, however, one would need to use a finite bootstrap size and a finite simulation size. This comprises B conditionally independent bootstrap resamples of $\{\widehat{F}_1^*,\ldots,\widehat{F}_m^*\}$, and R simulation replications driven by each realization of the resampled input distributions. This generally incurs two layers of Monte Carlo errors.

Denote $\hat{\psi}_r(\widehat{F}_1^b,\ldots,\widehat{F}_m^b)$ as the r-th simulation run driven by the b-th bootstrap resample. Denote $\bar{\psi}^b$ as the average of the R simulation runs driven by the b-th resample, and $\bar{\bar{\psi}}$ as the grand sample

average from all the BR runs. An unbiased estimator for $\mathrm{Var}_*[\psi(\widehat{F}_1^*, \ldots, \widehat{F}_m^*)]$ is given by

$$\frac{1}{B-1}\sum_{b=1}^{B}(\bar{\psi}^b - \bar{\bar{\psi}})^2 - \frac{V}{R} \tag{2.4}$$

where

$$V = \frac{1}{B(R-1)}\sum_{b=1}^{B}\sum_{r=1}^{R}(\hat{\psi}_r(\widehat{F}_1^b, \ldots, \widehat{F}_m^b) - \bar{\psi}^b)^2.$$

To explain, the first term in (2.4) is an unbiased estimate of the variance of $\bar{\psi}^b$, which can be expressed as $\mathrm{Var}_*[\psi(\widehat{F}_1^*, \ldots, \widehat{F}_m^*)] + (1/R)\mathbb{E}_*[\mathrm{Var}[\hat{\psi}_r(\widehat{F}_1^*, \ldots, \widehat{F}_m^*)|\widehat{F}_1^*, \ldots, \widehat{F}_m^*]]$ (where $\mathbb{E}_*[\cdot]$ denotes the expectation on $\widehat{F}_i^*$'s conditional on $\widehat{F}_i$'s), since $\bar{\psi}^b$ incurs both the bootstrap noise and the simulation noise. In other words, the variance of $\bar{\psi}^b$ is upward biased for $\mathrm{Var}_*[\psi(\widehat{F}_1^*, \ldots, \widehat{F}_m^*)]$. The second term in (2.4), namely V/R, removes this bias. This bias adjustment can be derived by viewing $\mathrm{Var}_*[\psi(\widehat{F}_1^*, \ldots, \widehat{F}_m^*)]$ as the variance of a conditional expectation. Alternately, $\hat{\psi}_r(\widehat{F}_1^*, \ldots, \widehat{F}_m^*)$ can be viewed as a random effect model where each "group" corresponds to each realization of $\widehat{F}_1^*, \ldots, \widehat{F}_m^*$, and (2.4) estimates the "between-group" variance in an analysis-of-variance (ANOVA). Formula (2.4) has appeared in the input uncertainty literature, e.g., Cheng and Holland (1997), Song and Nelson (2015), Lin et al. (2015), and also in Zouaoui and Wilson (2004) in the Bayesian context. Algorithm 1 summarizes the procedure.

More generally, to estimate the variance contribution from the data noise of model i only, namely σ_i^2/n_i, one can bootstrap only from $\{X_{i,1}, \ldots, X_{i,n_i}\}$ and keep other input distributions $\widehat{F}_j, j \neq i$ fixed. Then $\widehat{F}_i^*$ and $\widehat{F}_j, j \neq i$ are used to drive the simulation runs. With this modification, the same formula (2.4) or Algorithm 1 is an unbiased estimate for $\mathrm{Var}_*[\psi(\widehat{F}_1, \ldots, \widehat{F}_{i-1}, \widehat{F}_i^*, \widehat{F}_{i+1}, \ldots, \widehat{F}_m)]$, which is approximately $\mathrm{Var}[\psi(F_1, \ldots, F_{i-1}, \widehat{F}_i, F_{i+1}, \ldots, F_m)]$ by the bootstrap principle, in turn asymptotically equal to σ_i^2/n_i introduced in (2.2). This observation appeared in, e.g., Song et al. (2014); in Section 2.4 we give further justifications.

2.2.4 A Complexity Barrier

We explain intuitively the total number of simulation runs needed to ensure that the variance bootstrap depicted above can meaningfully estimate the input variance. For convenience, we call

Algorithm 1 ANOVA-based Variance Bootstrap

Given: $B \geq 2, R \geq 2$; data $= \{X_{i,j} : i = 1, \ldots, m, j = 1, \ldots, n_i\}$

for $b = 1$ **to** B **do**

For each i, draw a sample $\{X^b_{i,1}, \ldots, X^b_{i,n_i}\}$ uniformly with replacement from the data to obtain a resampled empirical distribution $\widehat{F}^b_i$

for $r = 1$ **to** R **do**

Simulate $\hat{\psi}_r(\widehat{F}^b_1, \ldots, \widehat{F}^b_m)$

end for

Compute $\bar{\psi}^b_{BV} = \frac{1}{R}\sum_{r=1}^{R} \hat{\psi}_r(\widehat{F}^b_1, \ldots, \widehat{F}^b_m)$

end for

Compute $V = \frac{1}{B(R-1)}\sum_{b=1}^{B}\sum_{r=1}^{R}(\hat{\psi}_r(\widehat{F}^b_1, \ldots, \widehat{F}^b_m) - \bar{\psi}^b_{BV})^2$ and $\bar{\bar{\psi}}_{BV} = \frac{1}{B}\sum_{b=1}^{B}\bar{\psi}^b_{BV}$

Output $\hat{\sigma}^2_{BV} = \frac{1}{B-1}\sum_{b=1}^{B}(\bar{\psi}^b_{BV} - \bar{\bar{\psi}}_{BV})^2 - \frac{V}{R}$

this number the *simulation complexity*. This quantity turns out to be of order bigger than the data size. On a high level, it is because the input variance scales reciprocally with the data size (recall (2.2)). Thus, when the data size increases, the input variance becomes smaller and increasingly difficult to estimate with controlled relative error. This in turn necessitates the use of more simulation runs.

To explain more concretely, denote n as a scaling of the data size, i.e., we assume n_i all grow linearly with n, which in particular implies that σ_I^2 is of order $1/n$. We analyze the error of $\hat{\sigma}^2_{BV}$ from Algorithm 1 in estimating σ_I^2. Since $\hat{\sigma}^2_{BV}$ is unbiased for $\mathrm{Var}_*[\psi(\widehat{F}^*_1, \ldots, \widehat{F}^*_m)]$ which is in turn close to σ_I^2, roughly speaking it suffices to focus on the variance of $\hat{\sigma}^2_{BV}$. To analyze this later quantity, we denote a generic simulation run in our procedure, $\hat{\psi}_r(\widehat{F}^*_1, \ldots, \widehat{F}^*_m)$, as

$$\hat{\psi}_r(\widehat{F}^*_1, \ldots, \widehat{F}^*_m) = \psi(\widehat{F}_1, \ldots, \widehat{F}_m) + \delta + \xi$$

where

$$\delta := \psi(\widehat{F}^*_1, \ldots, \widehat{F}^*_m) - \psi(\widehat{F}_1, \ldots, \widehat{F}_m),\ \xi := \hat{\psi}_r(\widehat{F}^*_1, \ldots, \widehat{F}^*_m) - \psi(\widehat{F}^*_1, \ldots, \widehat{F}^*_m).$$

are the errors arising from the bootstrap of the input distributions and the simulation respectively. If ψ is sufficiently smooth, δ elicits a central limit theorem and is of order $\Theta_p(1/\sqrt{n})$. On the other hand, the simulation noise ξ is of order $\Theta_p(1)$.

Via an ANOVA-type analysis as in Sun et al. (2011), we have

$$\begin{aligned}\mathrm{Var}_*[\hat{\sigma}^2_{BV}] =& \frac{1}{B}(\mathbb{E}_*[\delta^4] - (\mathbb{E}_*[\delta^2])^2) + \frac{2}{B(B-1)}(\mathbb{E}_*[\delta^2])^2 + \frac{2}{B^2R^2(B-1)}(\mathbb{E}_*[\xi^2])^2 + \frac{2}{B^2R^3}\mathbb{E}_*[\xi^4] \\ &+ \frac{2(B+1)}{B^2R(B-1)}\mathbb{E}_*[\delta^2]\mathbb{E}_*[\xi^2] + \frac{2(BR^2+R^2-4R+3)}{B^2R^3(R-1)}\mathbb{E}_*[(\mathbb{E}[\xi^2|\widehat{F}^*_1,\ldots,\widehat{F}^*_m])^2] \\ &+ \frac{4B+2}{B^2R}\mathbb{E}_*[\delta^2\xi^2] + \frac{4}{B^2R^2}\mathbb{E}_*[\delta\xi^3]. \end{aligned} \tag{2.5}$$

Now, putting $\delta = \Theta_p(1/\sqrt{n})$ and $\xi = \Theta_p(1)$ formally into (2.5), and ignoring constant factors, results in

$$\mathrm{Var}_*[\hat{\sigma}^2_{BV}] = O_p\left(\frac{1}{Bn^2} + \frac{1}{B^2n^2} + \frac{1}{B^3R^2} + \frac{1}{B^2Rn} + \frac{1}{B^2R^3} + \frac{1}{BR^2} + \frac{1}{BRn} + \frac{1}{B^2R^2\sqrt{n}}\right)$$

or simply

$$O_p\left(\frac{1}{Bn^2} + \frac{1}{BR^2}\right) \tag{2.6}$$

The two terms in (2.6) correspond to the variances coming from the bootstrap resampling and the simulation runs respectively.

Since σ^2_I is of order $1/n$, meaningful estimation of σ^2_I needs measured by the relative error. In other words, we want to achieve $\hat{\sigma}^2_{BV}/\sigma^2_I \xrightarrow{p} 1$ as the simulation budget grows. This property, which we call relative consistency, requires $\hat{\sigma}^2_{BV}$ to have a variance of order $o(1/n^2)$ (i.e., a standard error of $o(1/n)$) in order to compensate for the decreasing order of σ^2_I.

We argue that this implies unfortunately that the total number of simulation runs, BR, must be $\omega(n)$, i.e., of order higher than the data size. To explain, note that the first term in (2.6) forces one to use $B = \omega(1)$, i.e., the bootstrap size needs to grow with n, an implication that is quite natural. The second term in (2.6), on the other hand, dictates also that $BR^2 = \omega(n^2)$, which is satisfied if we use $R = \Theta(n)$ provided that B is already $\omega(1)$. Note that this gives rise to a

total simulation effort $BR = \omega(1) \cdot \Theta(n) = \omega(n)$, which can not be reduced further because the requirement $BR^2 = \omega(n^2)$ already entails that $(BR)^2 \geq BR^2 = \omega(n^2)$ must hold.

We summarize the above with the following result. Let N be the total simulation effort, and recall n as the scaling of the data size. We have:

Theorem 2.2.1 (Simulation complexity of variance bootstrap) *Under Assumptions 2.4.1-2.4.7 to be stated in Section 2.4.1, the required simulation budget to achieve relative consistency in estimating σ_I^2 by Algorithm 1, i.e., $\hat{\sigma}_{BV}^2/\sigma_I^2 \xrightarrow{p} 1$, is $N = \omega(n)$.*

Though out of the scope of this work, there are indications that such a computational barrier occurs in other types of bootstrap. For instance, the percentile bootstrap studied in Barton and Schruben (1993, 2001) appears to also require an inner replication size large enough compared to the data size in order to obtain valid quantile estimates (the authors actually used one inner replication, but Barton (2012) commented that more is needed). Yi and Xie (2017) provides an interesting approach based on ranking and selection to reduce the simulation effort, though they do not investigate the order of the needed effort relative to the data size. The empirical likelihood framework studied in Lam and Qian (2017) requires a similarly higher order of simulation runs to estimate the influence function. Nonetheless, in this work we focus only on how to reduce computation load in variance estimation.

2.3 Procedures and Guarantees in the Subsampling Framework

This section presents our methodologies and results on subsampling. Section 2.3.1 first explains the rationale and the subsampling procedure. Section 2.3.2 then presents our main theoretical guarantees, deferring some elaborate developments to Section 2.4.

2.3.1 Proportionate Subsampled Variance Bootstrap

As explained before, a huge simulation effort is required for the $\hat{\sigma}_{BV}^2$ in Algorithm 1 to achieve relative consistency, because the input variance shrinks at the rate $1/n$ as the input data size grows. In general, in order to estimate a quantity that is of order $1/n$, one must use a sample size

more than n so that the estimation error relatively vanishes. This requirement manifests in the inner replication size $R = \Theta(n)$ needed in constructing $\hat{\sigma}^2_{BV}$.

To reduce the inner replication size, we leverage the relation between the form of the input variance and the estimation variance depicted in (2.6) as follows. The approximate input variance contributed from model i, with data size n_i, has the form σ_i^2/n_i. If we use the variance bootstrap directly as in Algorithm 1, then we need an order more than n total simulation runs due to (2.6). Now, pretend that we have fewer than n_i but still sufficiently many data, say s_i, then the input variance will be approximately σ_i^2/s_i, and the required simulation runs is now only of order higher than s_i due to a reduced inner replication size $R = \Theta(s_i)$. An estimate of σ_i^2/s_i, however, already gives us enough information in estimating σ_i^2/n_i, because we can rescale our estimate of σ_i^2/s_i by s_i/n_i to get an estimate of σ_i^2/n_i. Estimating σ_i^2/s_i can be done by subsampling the input distribution with size s_i. With this, we can both use fewer simulation runs and also retain correct estimation via multiplying by a s_i/n_i factor.

To make the above argument more transparent, the bootstrap principle and the asymptotic approximation of the input variance imply that

$$\mathrm{Var}_*[\psi(\widehat{F}_1^*, \ldots, \widehat{F}_m^*)] = \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}(1 + o_p(1))$$

as the input data size n grows while $F_1, \ldots, F_m$ and ψ are fixed. As a side note, we comment that the $o_p(1)$ error term is usually independent of the dimensions of the inputs F_i's because the variance depends on the inputs only through the scalar quantity ψ (see the proof of Theorem 2.4.7 for a related analysis). The subsampling approach builds on the observation that a similar relation holds for

$$\mathrm{Var}_*[\psi(\widehat{F}_{s_1,1}^*, \ldots, \widehat{F}_{s_m,m}^*)] = \sum_{i=1}^{m} \frac{\sigma_i^2}{s_i}(1 + o_p(1))$$

where $\widehat{F}_{s_i,i}^*$ denotes a bootstrapped input distribution of size s_i (i.e., an empirical distribution of size s_i that is uniformly sampled with replacement from $\{X_{i,1}, \ldots, X_{i,n_i}\}$). If we let $s_i = \lfloor \theta n_i \rfloor$ for some $\theta > 0$ so that $s_i \to \infty$ (where $\lfloor \cdot \rfloor$ is the floor function, i.e. the largest integer less than or

equal to $\cdot$), then we have

$$\mathrm{Var}_*[\psi(\widehat{F}^*_{\lfloor\theta n_1\rfloor,1},\ldots,\widehat{F}^*_{\lfloor\theta n_m\rfloor,m})] = \sum_{i=1}^{m}\frac{\sigma_i^2}{\theta n_i}(1+o_p(1)).$$

Multiplying both sides with θ, we get

$$\theta\mathrm{Var}_*[\psi(\widehat{F}^*_{\lfloor\theta n_1\rfloor,1},\ldots,\widehat{F}^*_{\lfloor\theta n_m\rfloor,m})] = \sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}(1+o_p(1)).$$

Note that the right hand side above is the original input variance of interest. This leads to our *proportionate subsampled variance bootstrap*: We repeatedly subsample collections of input distributions from the data, with size $\lfloor\theta n_i\rfloor$ for model i, and use them to drive simulation replications. We then apply the ANOVA-based estimator in (2.4) on these replications, and multiply it by a factor of θ to obtain our final estimate. We summarize this procedure in Algorithm 2. The term "proportionate" refers to the fact that we scale the subsample size for all models with a single factor θ. For convenience, we call θ the *subsample ratio*.

Algorithm 2 Proportionate Subsampled Variance Bootstrap

Parameters: $B \geq 2, R \geq 2, 0 < \theta \leq 1$; data $= \{X_{i,j} : i = 1,\ldots,m, j = 1,\ldots,n_i\}$

Compute $s_i = \lfloor\theta n_i\rfloor$ for all i

for $b = 1$ **to** B **do**

For each i, draw a subsample $\{X^b_{i,1},\ldots,X^b_{i,s_i}\}$ uniformly with replacement from the data, which forms the empirical distribution $\widehat{F}^b_{s_i,i}$

for $r = 1$ **to** R **do**

Simulate $\hat{\psi}_r(\widehat{F}^b_{s_1,1},\ldots,\widehat{F}^b_{s_m,m})$

end for

Compute $\bar{\psi}^b = \frac{1}{R}\sum_{r=1}^{R}\hat{\psi}_r(\widehat{F}^b_{s_1,1},\ldots,\widehat{F}^b_{s_m,m})$

end for

Compute $V = \frac{1}{B(R-1)}\sum_{b=1}^{B}\sum_{r=1}^{R}(\hat{\psi}_r(\widehat{F}^b_{s_1,1},\ldots,\widehat{F}^b_{s_m,m}) - \bar{\psi}^b)^2$ and $\bar{\bar{\psi}} = \frac{1}{B}\sum_{b=1}^{B}\bar{\psi}^b$

Output $\hat{\sigma}^2_{SVB} = \theta(\frac{1}{B-1}\sum_{b=1}^{B}(\bar{\psi}^b - \bar{\bar{\psi}})^2 - \frac{V}{R})$

Similar ideas apply to estimating the individual variance contribution from each input model, namely σ_i^2/n_i. Instead of subsampling all input distributions, we only subsample the distribution, say $\widehat{F}^*_{s_i,i}$ whose uncertainty is of interest, while fixing all the other distributions as the original empirical distributions, i.e., $\widehat{F}_j, j \neq i$. All the remaining steps in Algorithm 2 remain the same (thus the "proportionate" part can be dropped). This procedure is depicted in Algorithm 3.

Algorithm 3 Subsampled Variance Bootstrap for Variance Contribution from the i-th Input Model

Parameters: $B \geq 2, R \geq 2, 0 < \theta \leq 1$; data $= \{X_{i,j} : i = 1, \ldots, m, j = 1, \ldots, n_i\}$

Compute $s_i = \lfloor \theta n_i \rfloor$

for $b = 1$ **to** B **do**

Draw a subsample $\{X^b_{i,1}, \ldots, X^b_{i,s_i}\}$ uniformly with replacement from the i-th input data set, which forms the empirical distribution $\widehat{F}^b_{s_i,i}$

for $r = 1$ **to** R **do**

Simulate $\hat{\psi}_r(\widehat{F}_1, \ldots, \widehat{F}_{i-1}, \widehat{F}^b_{s_i,i}, \widehat{F}_{i+1}, \ldots, \widehat{F}_m)$

end for

Compute $\bar{\psi}^b = \frac{1}{R}\sum_{r=1}^R \hat{\psi}_r(\widehat{F}_1, \ldots, \widehat{F}_{i-1}, \widehat{F}^b_{s_i,i}, \widehat{F}_{i+1}, \ldots, \widehat{F}_m)$

end for

Compute $V = \frac{1}{B(R-1)}\sum_{b=1}^B \sum_{r=1}^R (\hat{\psi}_r(\widehat{F}_1, \ldots, \widehat{F}_{i-1}, \widehat{F}^b_{s_i,i}, \widehat{F}_{i+1}, \ldots, \widehat{F}_m) - \bar{\psi}^b)^2$ and $\bar{\bar{\psi}} = \frac{1}{B}\sum_{b=1}^B \bar{\psi}^b$

Output $\hat{\sigma}^2_{SVB,i} = \theta(\frac{1}{B-1}\sum_{b=1}^B (\bar{\psi}^b - \bar{\bar{\psi}})^2 - \frac{V}{R})$

2.3.2 Statistical Guarantees

Algorithm 2 provides the following guarantees. Recall that $N = BR$ is the total simulation effort, and n is the scaling of the data size. We have the following result:

Theorem 2.3.1 *Under Assumptions 2.4.1-2.4.7 to be stated in Section 2.4.1, if the parameters B, R, θ of Algorithm 2 are chosen such that*

$$B \to \infty, \ \frac{BR^2}{(\theta n)^2} \to \infty, \ \theta n \to \infty \quad \text{as } n \to \infty \tag{2.7}$$

then the variance estimate $\hat{\sigma}^2_{SVB}$ *is relatively consistent, i.e.* $\hat{\sigma}^2_{SVB}/\sigma_I^2 \xrightarrow{p} 1$.

Theorem 2.3.1 tells us what orders of the bootstrap size B, inner replication size R and subsample ratio θ would guarantee a meaningful estimation of σ_I^2. Note that $\theta \approx s_i/n_i$ for each i, so that $\theta n = \omega(1)$ is equivalent to setting the subsample size $s_i = \omega(1)$. In other words, we need the natural requirement that the subsample size grows with the data size, albeit can have an arbitrary rate.

Given a subsample ratio θ specified according to (2.7), the configurations of B and R under (2.7) that achieve the minimum overall simulation budget is $B = \omega(1)$ and $R = \Omega(\theta n)$. This is because to minimize $N = BR$ while satisfying the second requirement in (2.7), it is more economical to allocate as much budget to R instead of B. This is stated precisely as:

Corollary 2.3.2 *Under the conditions of Theorem 2.3.1, given* θ *such that* $\theta n \to \infty$*, the values of* B *and* R *to achieve* (2.7) *and hence relative consistency that requires the least order of effort are* $B \to \infty$ *and* $R \geq C\theta n$ *for some constant* $C > 0$*, leading to a total simulation budget* N *such that* $\frac{N}{\theta n} \to \infty$.

Note that θn is the order of the subsample size. Thus Corollary 2.3.2 implies that the required simulation budget must grow linearly in the subsample size. However, since the subsample size can be chosen to grow at an arbitrarily small rate, this implies that the total budget can also grow arbitrarily slow relative to the input data size. Therefore, we have:

Corollary 2.3.3 (Simulation complexity) *Under the same conditions of Theorem 2.3.1, the minimum required simulation budget to achieve relative consistency in estimating* σ_I^2 *by Algorithm 2, i.e.,* $\hat{\sigma}^2_{SVB}/\sigma_I^2 \xrightarrow{p} 1$*, is* $N \to \infty$ *as* $n \to \infty$ *by using a* θ *such that* $\theta n \to \infty$.

Compared to Theorem 2.2.1, Corollary 2.3.3 stipulates that our subsampling approach reduces the required simulation effort from a higher order than n to an arbitrary order, i.e., independent of the data size. This is achieved by using a subsample size that grows with n at an arbitrary order, or equivalently a subsample ratio θ that grows faster than $1/n$.

The following result describes the configurations of our scheme when a certain total simulation effort is given. In particular, it shows, for a given total simulation effort, the range of subsample

ratio for which Algorithm 2 can possibly generate valid variance estimates by appropriately choosing B and R:

Theorem 2.3.4 (Valid subsample ratio given total budget) *Assume the same conditions of Theorem 2.3.1. Given a total simulation budget N such that $N \to \infty$, if the subsample ratio satisfies $\theta n \to \infty$ and $\frac{\theta n}{N} \to 0$, then the bootstrap size B and the inner replication size R can be appropriately chosen according to criterion* (2.7) *to achieve relative consistency, i.e., $\hat{\sigma}^2_{SVB}/\sigma^2_I \xrightarrow{p} 1$.*

The next result is on the optimal configurations of our scheme in minimizing the Monte Carlo error. To proceed, define

$$\sigma^2_{SVB} = \theta \mathrm{Var}_*[\psi(\widehat{F}^*_{\lfloor \theta n_1 \rfloor,1}, \ldots, \widehat{F}^*_{\lfloor \theta n_m \rfloor,m})] \tag{2.8}$$

as the perfect form of our proportionate subsampled variance bootstrap introduced in Section 2.3.1, namely without any Monte Carlo noises, and $0 < \theta \leq 1$ is the subsample ratio. We have:

Theorem 2.3.5 (Optimal budget allocation) *Assume the same conditions of Theorem 2.3.1. Given a simulation budget N and a subsample ratio θ such that $\frac{N}{\theta n} \to \infty$ and $\theta n \to \infty$, the optimal outer and inner sizes that minimize the order of the conditional mean squared error $\mathbb{E}_*[(\hat{\sigma}^2_{SVB} - \sigma^2_{SVB})^2]$ are*

$$B^* = \frac{N}{R^*},\ R^* = \Theta(\theta n)$$

giving a conditional mean squared error $\mathbb{E}_*[(\hat{\sigma}^2_{SVB} - \sigma^2_{SVB})^2] = \Theta(\theta/(Nn))(1 + o_p(1))$.

Note that the mean squared error, i.e. $\mathbb{E}_*[(\hat{\sigma}^2_{SVB} - \sigma^2_{SVB})^2]$, of the Monte Carlo estimate $\hat{\sigma}^2_{SVB}$ is random because the underlying resampling is conditioned on the input data, therefore the bound at the end of Theorem 2.3.5 contains a stochastically vanishing term $o_p(1)$.

We next present the optimal tuning of the subsample ratio. This requires a balance of the trade-off between the input statistical error and the Monte Carlo simulation error. To explain, the overall error of $\hat{\sigma}^2_{SVB}$ by Algorithm 2 can be decomposed as

$$\hat{\sigma}^2_{SVB} - \sigma^2_I = (\hat{\sigma}^2_{SVB} - \sigma^2_{SVB}) + (\sigma^2_{SVB} - \sigma^2_I). \tag{2.9}$$

The first term is the Monte Carlo error for which the optimal outer size B, inner size R and the resulting mean squared error are governed by Theorem 2.3.5. In particular, the mean squared error there shows that under a fixed simulation budget N and the optimal allocation $R = \Theta(\theta n)$, the Monte Carlo error gets larger as θ increases. The second term is the statistical errors due to the finiteness of input data and θ. Since θ measures the amount of data contained in the resamples, we expect this second error to become smaller as θ increases. The optimal tuning of θ relies on balancing such a trade-off between the two sources of errors.

We have the following optimal configurations of B, R and θ altogether given a budget N:

Theorem 2.3.6 (Optimal subsample size) *Suppose Assumptions 2.4.1, 2.4.3-2.4.7 in Section 2.4.1 and Assumptions 2.4.10-2.4.12 in Section 2.4.3 hold. For a given simulation budget N such that $N \to \infty$ as $n \to \infty$, if the subsample ratio θ and outer and inner sizes B, R for Algorithm 2 are set to*

$$\begin{cases} \theta^* = \Theta\big(N^{1/3}n^{-1}\big) & \text{if } 1 \ll N \le n^{3/2} \\ \Theta(n^{-1/2}) \le \theta^* \le \Theta\big(Nn^{-2} \wedge 1\big) & \text{if } N > n^{3/2} \end{cases} \tag{2.10}$$

$$R^* = \Theta(\theta^* n),\ B^* = \frac{N}{R^*} \tag{2.11}$$

then the gross error $\hat{\sigma}^2_{SVB} - \sigma^2_I = \mathcal{E} + o_p(N^{-1/3}n^{-1} + n^{-3/2})$, where the leading term has a mean squared error

$$\mathbb{E}[\mathcal{E}^2] = O\big(\frac{1}{N^{2/3}n^2} + \frac{1}{n^3}\big). \tag{2.12}$$

Moreover, if $\mathcal{R} = \Theta((ns)^{-1})$ and at least one of the Σ_i's are positive definite, where $\mathcal{R}$ and Σ_i are as defined in Lemma 2.4.8, then (2.12) *holds with an exact order (i.e., $O(\cdot)$ becomes $\Theta(\cdot)$) and the configuration* (2.10), (2.11) *is optimal in the sense that no configuration gives rise to a gross error $\hat{\sigma}^2_{SVB} - \sigma^2_I = o_p\big(N^{-1/3}n^{-1} + n^{-3/2}\big)$.*

Note from (2.12) that, if the budget $N = \omega(1)$, our optimal configurations guarantee the estimation mean squared error decays faster than $1/n^2$. Recall that the input variance is of order $1/n$, and thus an estimation error of order higher than $1/n^2$ ensures that the estimator is relatively

consistent in the sense $\hat{\sigma}^2_{SVB}/\sigma^2_I \xrightarrow{p} 1$. This recovers the result in Corollary 2.3.3. We also comment that the algorithmic configuration given in Theorem 2.3.6 is chosen to optimize the mean squared error of the input variance estimate, but does not necessarily generates the most accurate CI. There exists evidence (e.g., Sun et al. (2008)) that the optimal choice to minimize the mean squared error of the variance estimate can be different from the one that is optimal for statistical inference, although in our experiments they seem to match closely with each other.

We comment that all the results in this section hold if one estimates the individual variance contribution from each input model i, namely by using Algorithm 3. In this case we are interested in estimating the variance σ_i^2/n_i, and relative consistency means $\hat{\sigma}^2_{SVB,i}/(\sigma_i^2/n_i) \xrightarrow{p} 1$. The data size scaling parameter n can be replaced by n_i in all our results.

Finally, we also comment that the complexity barrier described in Section 2.2.4 and our framework presented in this section applies in principle to the parametric regime, i.e., when the input distributions are known to lie in parametric families with unknown parameters. The assumptions and mathematical details would need to be catered to that situation, which could be done naturally by viewing the "parameter" as a function of F_i's.

2.4 Developments of Theoretical Results

We present our main developments leading to the algorithms and results in Section 2.3. Section 2.4.1 first states in detail our assumptions on the performance measure. Section 2.4.2 presents the theories leading to estimation accuracy, simulation complexity and optimal budget allocation in the proportionate subsampled variance bootstrap. Section 2.4.3 investigates optimal subsample sizes that lead to overall best configurations.

2.4.1 Regularity Assumptions

We first assume that the data sets for all input models are of comparable size.

Assumption 2.4.1 (Balanced data) $\limsup_{all\ n_i\to\infty} \frac{\max_i n_i}{\min_i n_i} < \infty$ *as all* $n_i \to \infty$.

Recall in Sections 2.2 and 2.3 that we have denoted n as a scaling of the data size. More concretely, we take $n = (1/m)\sum_{i=1}^{m} n_i$ as the average input data size under Assumption 2.4.1.

We next state a series of general assumptions on the performance measure ψ. These assumptions hold for common finite-horizon measures, as we will present. For each i let Ξ_i be the support of the i-th true input model F_i, and the collection of distributions $\mathcal{P}_i$ be the convex hull spanned by F_i and all Dirac measures on Ξ_i, i.e.

$$\mathcal{P}_i = \big\{\nu_1 F_i + \sum_{k=2}^{l} \nu_k \mathbf{1}_{x_k} : \sum_{k=1}^{l} \nu_k = 1, \nu_k \geq 0, l < \infty, x_k \in \Xi_i \text{ for all } k\big\}.$$

We assume the following differentiability of the performance measure.

Assumption 2.4.2 (First order differentiability) *For any distributions $P_i, Q_i \in \mathcal{P}_i$, denote $P_i^{\nu_i} = (1-\nu_i)P_i + \nu_i Q_i$ for $\nu_i \in [0,1]$. Assume there exist functions $g_i(P_1, \ldots, P_m; \cdot) : \Xi_i \to \mathbb{R}$ such that $\mathbb{E}_{P_i}[g_i(P_1, \ldots, P_m; X_i)] = 0$ for $i = 1, \ldots, m$ and as all ν_i's approach zero*

$$\psi(P_1^{\nu_1}, \ldots, P_m^{\nu_m}) - \psi(P_1, \ldots, P_m) = \sum_{i=1}^{m} \nu_i \int g_i(P_1, \ldots, P_m; x) d(Q_i - P_i)(x) + o\Big(\sqrt{\sum_{i=1}^{m} \nu_i^2}\Big). \quad (2.13)$$

The differentiability described above is defined with respect to a particular direction, namely $Q_i - P_i$, in the space of probability measures, and is known as Gateaux differentiability or directional differentiability (e.g., Serfling (2009), Van der Vaart (2000)). Assumption 2.4.2 therefore requires the performance measure ψ to be Gateaux differentiable when restricted to the convex set $\mathcal{P}_1 \times \cdots \times \mathcal{P}_m$. The functions g_i's are also called the influence functions (e.g., Hampel (1974)) that play analogous roles as standard gradients in the Euclidean space. The condition of g_i's having vanishing means is without loss of generality since such a condition can always be achieved by centering, i.e., subtracting the mean. Note that doing this does not make any difference to the first term of expansion (2.13) because both Q_i and P_i are probability measures. Taking each $\nu_i = 1$ in (2.13), one informally obtains the Taylor expansion of ψ around P_i's

$$\psi(Q_1, \ldots, Q_m) - \psi(P_1, \ldots, P_m) \approx \sum_{i=1}^{m} \int g_i(P_1, \ldots, P_m; x) d(Q_i - P_i)(x).$$

When each P_i is set to be the true input model F_i and Q_i to be the empirical input model $\widehat{F}_i$, the above linear expansion is expected to be a reasonably good approximation as the data size grows. The next assumption imposes a moment bound on the error of this approximation:

Assumption 2.4.3 (Smoothness at true input models) *Denote by* $g_i(\cdot) := g_i(F_1, \ldots, F_m; \cdot)$ *the influence functions at the true input distributions* $F_i, i = 1, \ldots, m$. *Assume that the remainder in the Taylor expansion of the performance measure*

$$\psi(\widehat{F}_1, \ldots, \widehat{F}_m) = \psi(F_1, \ldots, F_m) + \sum_{i=1}^{m} \int g_i(x) d(\widehat{F}_i - F_i)(x) + \epsilon \tag{2.14}$$

satisfies $\mathbb{E}[\epsilon^2] = o(n^{-1})$ *as* $n \to \infty$, *and the influence functions* g_i*'s are non-degenerate, i.e.* $\sigma_i^2 := \mathrm{Var}_{F_i}[g_i(X_i)] > 0$, *and have finite fourth moments, i.e.* $\mathbb{E}_{F_i}[g_i^4(X_i)] < \infty$.

Assumption 2.4.3 entails that the error of the linear approximation formed by influence functions is negligible in the asymptotic sense. Indeed, the linear term in (2.14) is asymptotically of order $\Theta_p(n^{-1/2})$ by the central limit theorem, whereas the error ϵ is implied by Assumption 2.4.3 to be $o_p(n^{-1/2})$. Hence the variance of the linear term contributes dominantly to the overall input variance as n_i's are large. Note that, like the $o_p(1)$ error in approximating the input variance using the bootstrap principle, the $o(n^{-1})$ order for $\mathbb{E}[\epsilon^2]$ is typically independent of the dimensions of the input distributions, as can be seen from the proof of Theorem 2.4.7. Then, thanks to the independence among the input models, the input variance can be expressed in the additive form described in (2.2) together with a negligible error.

Proposition 2.4.1 *Under Assumptions 2.4.1-2.4.3, the input variance* σ_I^2 *defined in* (2.1) *takes the form*

$$\sigma_I^2 = \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} + o(\frac{1}{n}) \quad \text{as } n \to \infty$$

where each $\sigma_i^2 = \mathrm{Var}_{F_i}[g_i(X_i)]$ *is the variance of the i-th influence function.*

As mentioned before, consistent estimation of input variance σ_I^2 relies on the bootstrap principle, for which we make the following additional assumptions. The assumption states that the error of the linear approximation (2.14) remains small when the underlying distributions F_i are replaced by

the empirical input distributions $\widehat{F}_i$, hence can be viewed as a bootstrapped version of Assumption 2.4.3.

Assumption 2.4.4 (Smoothness at empirical distributions) *Let $\hat{g}_i(\cdot) := g_i(\widehat{F}_1, \ldots, \widehat{F}_m; \cdot)$ be the influence functions at the empirical input distributions $\widehat{F}_i, i = 1, \ldots, m$. Assume the empirical influence function converges to the truth in the sense that $\mathbb{E}[(\hat{g}_i - g_i)^4(X_{i,1})] \to 0$. For each i let $\overline{F}_i$ be either the i-th empirical input model $\widehat{F}_i$ or the resampled model $\widehat{F}^*_{s_i,i}$. For every $(\overline{F}_1, \ldots, \overline{F}_m) \in \prod_{i=1}^m \{\widehat{F}_i, \widehat{F}^*_{s_i,i}\}$, assume the remainder in the Taylor expansion*

$$\psi(\overline{F}_1, \ldots, \overline{F}_m) = \psi(\widehat{F}_1, \ldots, \widehat{F}_m) + \sum_{i=1}^m \int \hat{g}_i(x) d(\overline{F}_i - \widehat{F}_i)(x) + \epsilon^* \tag{2.15}$$

satisfies $\mathbb{E}_*[(\epsilon^*)^4] = o_p\big(s^{-2}\big)$ as both $n, s \to \infty$, where $s := \frac{1}{m}\sum_{i=1}^m s_i$ is the average subsample size.

As the data sizes n_i's grow, the empirical input distributions $\widehat{F}_i$ converge to the true ones F_i. Hence the empirical influence functions $\hat{g}_i$'s are expected to approach the influence functions g_i's associated with the true input distributions, which explains the convergence condition in Assumption 2.4.4. The fourth moment condition on the remainder ϵ^* is needed for controlling the variance of our variance estimator. Since the fourth moment is with respect to the resampling measure and thus depends on the underlying input data, the condition is described in terms of stochastic order. Note that we require (2.15) to hold not just when $\overline{F}_i = \widehat{F}^*_{s_i,i}$ for all i but also when some $\overline{F}_i = \widehat{F}_i$. This allows us to estimate the variance contributed from an arbitrary group of input models and in particular an individual input model.

Assumptions 2.4.2-2.4.4 are on the performance measure ψ itself. Next we impose assumptions on the simulation noise, i.e. the stochastic error $\hat{\psi}_r - \psi$ where $\hat{\psi}_r$ is an unbiased simulation replication for ψ. We denote by $\tau^2(P_1, \ldots, P_m)$ the variance of $\hat{\psi}_r$ when simulation is driven by arbitrary input models $P_1, \ldots, P_m$, i.e.

$$\tau^2(P_1, \ldots, P_m) = \mathbb{E}_{P_1,\ldots,P_m}[(\hat{\psi}_r - \psi(P_1, \ldots, P_m))^2].$$

Similarly we denote by $\mu_4(P_1, \ldots, P_m)$ the fourth central moment of $\hat{\psi}_r$ under the input models

$P_1, \ldots, P_m$

$$\mu_4(P_1, \ldots, P_m) = \mathbb{E}_{P_1,\ldots,P_m}[(\hat{\psi}_r - \psi(P_1, \ldots, P_m))^4].$$

In particular, for convenience we write $\tau^2 = \tau^2(F_1, \ldots, F_m)$ for the variance of $\hat{\psi}$ under the true input models, and $\hat{\tau}^2 = \tau^2(\widehat{F}_1, \ldots, \widehat{F}_m)$ for that under the empirical input models.

The assumptions on the simulation noise are:

Assumption 2.4.5 (Convergence of empirical variance) *$\hat{\tau}^2 \xrightarrow{p} \tau^2$ as $n \to \infty$.*

Assumption 2.4.6 (Convergence of bootstrapped variance) *For every $(\overline{F}_1, \ldots, \overline{F}_m)$ where each $\overline{F}_i = \widehat{F}_i$ or $\widehat{F}^*_{s_i,i}$, it holds that $\mathbb{E}_*[(\tau^2(\overline{F}_1, \ldots, \overline{F}_m) - \hat{\tau}^2)^2] = o_p(1)$ as both $n, s \to \infty$.*

Assumption 2.4.7 (Boundedness of the fourth moment) *For every $(\overline{F}_1, \ldots, \overline{F}_m)$ such that each $\overline{F}_i = \widehat{F}_i$ or $\widehat{F}^*_{s_i,i}$, it holds that $\mathbb{E}_*[\mu_4(\overline{F}_1, \ldots, \overline{F}_m)] = O_p(1)$ as both $n, s \to \infty$.*

Assumptions 2.4.5 and 2.4.6 stipulate that the variance of the simulation replication $\hat{\psi}_r$ as a functional of the underlying input models is smooth enough in the inputs. Conceptually Assumption 2.4.5 is in line with Assumption 2.4.3 in the sense that both concern smoothness of a functional around the true input models, whereas Assumption 2.4.6 is similar to Assumption 2.4.4 since both are about smoothness property around the empirical input models. Assumption 2.4.7 is a fourth moment condition like in Assumption 2.4.4 used to control the variance of the variance estimator. Similar to Assumption 2.4.4, we impose Assumptions 2.4.6 and 2.4.7 for each $\overline{F}_i = \widehat{F}_i$ or $\widehat{F}^*_{s_i,i}$ so that the same guarantees remain valid when estimating input variances from individual input models, i.e., Algorithm 3.

Although the above assumptions may look complicated, they can be verified, under minimal conditions, for generic finite-horizon performance measures in the form

$$\psi(F_1, \ldots, F_m) = \mathbb{E}_{F_1,\ldots,F_m}[h(\mathbf{X}_1, \ldots, \mathbf{X}_m)] \tag{2.16}$$

where $\mathbf{X}_i = (X_i(1), \ldots, X_i(T_i))$ represents the i-th input process consisting of T_i i.i.d. variables distributed under F_i, each T_i being a deterministic time, and h is a performance function. An unbiased simulation replication $\hat{\psi}_r$ of the performance measure is $h(\mathbf{X}_1, \ldots, \mathbf{X}_m)$.

Suppose we have the following conditions for the performance function h:

Assumption 2.4.8 *For each i, $0 < \mathrm{Var}_{F_i}[\sum_{t=1}^{T_i} \mathbb{E}_{F_1,\ldots,F_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t) = X_i]] < \infty$.*

Assumption 2.4.9 (Parameter k) *For each i let $I_i = (I_i(1),\ldots,I_i(T_i))$ be a sequence of indices such that $1 \le I_i(t) \le t$, and $\mathbf{X}_{i,I_i} = (X_i(I_i(1)),\ldots,X_i(I_i(T_i)))$. Assume*

$$\max_{I_1,\ldots,I_m} \mathbb{E}_{F_1,\ldots,F_m}[|h(\mathbf{X}_{1,I_1},\ldots,\mathbf{X}_{m,I_m})|^k] < \infty.$$

The conditional expectation in Assumption 2.4.8 is in fact the influence function of the performance measure (2.16) under the true input models. So Assumption 2.4.8 is precisely the non-degenerate variance condition in Assumption 2.4.3. All other parts of Assumptions 2.4.2-2.4.7 are consequences of the moment condition in Assumption 2.4.9:

Theorem 2.4.2 *Under Assumptions 2.4.1, 2.4.8 and Assumption 2.4.9 with $k = 4$, we have Assumptions 2.4.2-2.4.7 hold for the finite-horizon performance measure ψ given by* (2.16)*.*

2.4.2 Simulation Complexity and Allocation

This section presents theoretical developments on our proportionate subsampled variance bootstrap. We first establish relative consistency assuming infinite computation resources. Recall (2.8) as the proportionate subsampled variance bootstrap estimator without any Monte Carlo errors. The following theorem gives a formal statement on the performance of this estimator discussed in Section 2.3.1.

Theorem 2.4.3 *Under Assumptions 2.4.1-2.4.4, if the subsample ratio θ satisfies $\theta n \to \infty$, then the proportionate subsampled variance bootstrap without Monte Carlo error, namely* (2.8)*, is relatively consistent as $n \to \infty$, i.e.*

$$\sigma^2_{SVB}/\sigma^2_I \xrightarrow{p} 1.$$

The requirement $\theta = \omega(n^{-1})$ implies that $s_i \to \infty$, which is natural as one needs minimally an increasing subsample size to ensure the consistency of our estimator. It turns out that this minimal requirement is enough to ensure consistency even relative to the magnitude of σ^2_I.

Now we turn to the discussion of the Monte Carlo estimate of the bootstrap variance generated from Algorithm 2. The following lemma characterizes the amount of Monte Carlo noise in terms of mean squared error.

Lemma 2.4.4 *The output $\hat{\sigma}^2_{SVB}$ of Algorithm 2 is unbiased for the proportionate subsampled variance bootstrap without Monte Carlo errors, namely σ^2_{SVB}. Furthermore, under Assumptions 2.4.1-2.4.7, if*

$$B \to \infty,\ \theta n \to \infty \quad \text{as } n \to \infty \tag{2.17}$$

and R is arbitrary, then the conditional mean squared error

$$\mathbb{E}_*[(\hat{\sigma}^2_{SVB} - \sigma^2_{SVB})^2] = \frac{2}{B}\Big(\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} + \frac{\tau^2\theta}{R}\Big)^2 (1 + o_p(1)). \tag{2.18}$$

In addition to the condition $\theta = \omega(n^{-1})$ which has appeared in Theorem 2.4.3, we also require $B = \omega(1)$ in Lemma 2.4.4. As the proof reveals, with such a choice of B, we can extract the leading term of the conditional mean squared error shown in (2.18), which takes a neat form and is easy to analyze.

Note that σ_I^2 here is of order n^{-1} by Proposition 2.4.1. Hence the Monte Carlo noise of the variance estimate output by our algorithm has to vanish faster than n^{-1} in order to achieve relative consistency. Combining Theorem 2.4.3 and Lemma 2.4.4, we obtain the simulation complexity of $\hat{\sigma}^2_{SVB}$ in Theorem 2.3.1. To establish the theoretical optimal allocation on the outer and inner sizes B, R, for given data sizes n_i, subsample ratio θ, and total simulation budget N, we minimize the conditional mean square error (2.18) subject to the budget constraint $BR = N$. This gives rise to the following result that gives a more precise (theoretical) statement than Theorem 2.3.5.

Theorem 2.4.5 *Suppose Assumptions 2.4.1-2.4.7 hold. Given a simulation budget N and a subsample ratio θ such that $\frac{N}{\theta n} \to \infty$ and $\theta n \to \infty$, the optimal outer and inner sizes that minimize the conditional mean squared error $\mathbb{E}_*[(\hat{\sigma}^2_{SVB} - \sigma^2_{SVB})^2]$ are*

$$B^* = \frac{N}{R^*},\ R^* = \frac{\theta\tau^2}{\sum_{i=1}^{m} \sigma_i^2/n_i}$$

which gives a conditional mean squared error

$$\mathbb{E}_*[(\hat{\sigma}^2_{SVB} - \sigma^2_{SVB})^2] = \frac{8\theta\tau^2}{N}\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}(1 + o_p(1)). \tag{2.19}$$

Theorem 2.4.5 gives the exact choices of B and R that minimize the Monte Carlo error. However, this is more of theoretical interest because the optimal R^* involves the desired input variance $\sum_{i=1}^m \sigma_i^2/n_i$. Having said that, we can conclude from the theorem that the optimal inner size R is of order $\Theta(\theta n)$, the same as the subsample size, because the input variance is of order $\Theta(1/n)$ by Proposition 2.4.1 and τ^2 is a constant. This results in Theorem 2.3.5 in Section 2.3.2.

2.4.3 Optimal Subsample Ratio

In this section we further establish the optimal subsample ratio θ or equivalently subsample sizes s_i that balance the two sources of errors in (2.9). For this, we need more regularity conditions on the performance measure. The first assumption we need is third order Gateaux differentiability in the convex set $\mathcal{P}_1 \times \cdots \times \mathcal{P}_m$:

Assumption 2.4.10 (Third order differentiability) *Using the same notations $P_i, Q_i, P_i^{\nu_i}$ as in Assumption 2.4.2, assume that there exist second order influence functions $g_{i_1 i_2}(P_1, \ldots, P_m; \cdot) : \Xi_{i_1} \times \Xi_{i_2} \to \mathbb{R}$ and third order influence functions $g_{i_1 i_2 i_3}(P_1, \ldots, P_m; \cdot) : \Xi_{i_1} \times \Xi_{i_2} \times \Xi_{i_3} \to \mathbb{R}$ for $i_1, i_2, i_3 = 1, \ldots, m$ which are symmetric under permutations, namely*

$$g_{i_1 i_2}(P_1, \ldots, P_m; x_1, x_2) = g_{i_2 i_1}(P_1, \ldots, P_m; x_2, x_1)$$
$$g_{i_1 i_2 i_3}(P_1, \ldots, P_m; x_1, x_2, x_3) = g_{i_2 i_1 i_3}(P_1, \ldots, P_m; x_2, x_1, x_3) = g_{i_1 i_3 i_2}(P_1, \ldots, P_m; x_1, x_3, x_2).$$

and for all x, y satisfy

$$\mathbb{E}_{P_{i_2}}[g_{i_1 i_2}(P_1, \ldots, P_m; x, X_{i_2})] = 0,\ \mathbb{E}_{P_{i_3}}[g_{i_1 i_2 i_3}(P_1, \ldots, P_m; x, y, X_{i_3})] = 0.$$

Moreover, as all ν_i's approach zero the following Taylor expansion holds

$$\begin{aligned}
&\psi(P_1^{\nu_1},\ldots,P_m^{\nu_m})-\psi(P_1,\ldots,P_m)\\
=\;&\sum_{i=1}^{m}\nu_i\int g_i(P_1,\ldots,P_m;x)d(Q_i-P_i)(x)\\
&+\frac{1}{2}\sum_{i_1,i_2=1}^{m}\nu_{i_1}\nu_{i_2}\int g_{i_1i_2}(P_1,\ldots,P_m;x_1,x_2)\prod_{k=1}^{2}d(Q_{i_k}-P_{i_k})(x_k)\\
&+\frac{1}{6}\sum_{i_1,i_2,i_3=1}^{m}\nu_{i_1}\nu_{i_2}\nu_{i_3}\int g_{i_1i_2i_3}(P_1,\ldots,P_m;x_1,x_2,x_3)\prod_{k=1}^{3}d(Q_{i_k}-P_{i_k})(x_k)+o\Big((\sum_{i=1}^{m}\nu_i^2)^{\frac{3}{2}}\Big).
\end{aligned}$$

Assumption 2.4.10 complements and strengthens Assumption 2.4.2 in that it imposes stronger differentiability property. Similarly, the following two assumptions strengthen Assumptions 2.4.3 and 2.4.4 respectively by considering cubic expansions.

Assumption 2.4.11 (Third order smoothness at true input models) *Denote by $g_{i_1i_2}(\cdot):=g_{i_1i_2}(F_1,\ldots,F_m;\cdot)$ and $g_{i_1i_2i_3}(\cdot):=g_{i_1i_2i_3}(F_1,\ldots,F_m;\cdot)$ the second and third order influence functions under the true input models. Assume the remainder in the Taylor expansion of the plug-in estimator $\psi(\widehat{F}_1,\ldots,\widehat{F}_m)$*

$$\begin{aligned}
\psi(\widehat{F}_1,\ldots,\widehat{F}_m)\;=\;&\psi(F_1,\ldots,F_m)+\sum_{i=1}^{m}\int g_i(x)d(\widehat{F}_i-F_i)(x)\\
&+\frac{1}{2}\sum_{i_1,i_2=1}^{m}\int g_{i_1i_2}(x_1,x_2)\prod_{k=1}^{2}d(\widehat{F}_{i_k}-F_{i_k})(x_k)\\
&+\frac{1}{6}\sum_{i_1,i_2,i_3=1}^{m}\int g_{i_1i_2i_3}(x_1,x_2,x_3)\prod_{k=1}^{3}d(\widehat{F}_{i_k}-F_{i_k})(x_k)+\epsilon_3
\end{aligned}$$

satisfies $\mathbb{E}[\epsilon_3^2]=o(n^{-3})$ as $n\to\infty$, and the high order influence functions satisfy the moment conditions

$$\mathbb{E}[g_{i_1i_2}^4(X_{i_1,1},X_{i_2,j_2})]<\infty,\ \mathbb{E}[g_{i_1i_2i_3}^2(X_{i_1,1},X_{i_2,j_2},X_{i_3,j_3})]<\infty$$

for all i_1,i_2,i_3 and $j_2\le 2,j_3\le 3$, where $X_{i,j}$ is the j-th data point from the i-th input model.

Similar to the remainder ϵ in Assumption 2.4.3, the moment condition on ϵ_3 here is used to control the error of the cubic approximation of ψ formed by up to third order influence functions. With these additional assumptions, the error term in Proposition 2.4.1 can be refined as follows:

Proposition 2.4.6 *Under Assumptions 2.4.1, 2.4.3 and 2.4.10-2.4.11, the overall input variance, as defined in* (2.1)*, can be expressed as*

$$\sigma_I^2 = \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} + O(\frac{1}{n^2}) \quad \text{as } n \to \infty.$$

We also need third order differentiability around the empirical input models:

Assumption 2.4.12 (Third order smoothness at empirical input models) *Let* $\hat{g}_{i_1 i_2}(\cdot) := g_{i_1 i_2}(\widehat{F}_1, \ldots, \widehat{F}_m; \cdot)$ *and* $\hat{g}_{i_1 i_2 i_3}(\cdot) := g_{i_1 i_2 i_3}(\widehat{F}_1, \ldots, \widehat{F}_m; \cdot)$ *be the second and third order influence functions under the empirical input models. Assume that the remainder in the Taylor expansion of the bootstrapped performance measure* $\psi(\widehat{F}^*_{s_1,1}, \ldots, \widehat{F}^*_{s_m,m})$

$$\begin{aligned}
\psi(\widehat{F}^*_{s_1,1}, \ldots, \widehat{F}^*_{s_m,m}) &= \psi(\widehat{F}_1, \ldots, \widehat{F}_m) + \int \hat{g}_i(x) d(\widehat{F}^*_{s_i,i} - \widehat{F}_i)(x) \\
&\quad + \frac{1}{2} \sum_{i_1,i_2=1}^{m} \int \hat{g}_{i_1 i_2}(x_1, x_2) \prod_{k=1}^{2} d(\widehat{F}^*_{s_{i_k},i_k} - \widehat{F}_{i_k})(x_k) \\
&\quad + \frac{1}{6} \sum_{i_1,i_2,i_3=1}^{m} \int \hat{g}_{i_1 i_2 i_3}(x_1, x_2, x_3) \prod_{k=1}^{3} d(\widehat{F}^*_{s_{i_k},i_k} - \widehat{F}_{i_k})(x_k) + \epsilon_3^*
\end{aligned}$$

satisfies $\mathbb{E}_*[(\epsilon_3^*)^2] = o_p(s^{-3})$ *as both* $n, s \to \infty$*. In addition, assume the high order empirical influence functions* $\hat{g}_{i_1 i_2}$ *and* $\hat{g}_{i_1 i_2 i_3}$ *converge in mean square error, i.e.*

$$\mathbb{E}[(\hat{g}_{i_1 i_2} - g_{i_1 i_2})^2(X_{i_1,1}, X_{i_2,j_2})] \to 0, \ \mathbb{E}[(\hat{g}_{i_1 i_2 i_3} - g_{i_1 i_2 i_3})^2(X_{i_1,1}, X_{i_2,j_2}, X_{i_3,j_3})] \to 0$$

for all i_1, i_2, i_3 *and* $j_2 \leq 2, j_3 \leq 3$*, where* $X_{i,j}$ *is the* j*-th data point from the* i*-th input model. For the first order influence function* $\hat{g}_i$*, assume the remainder in the Taylor expansion*

$$\hat{g}_i(X_{i,1}) = g_i(X_{i,1}) + \sum_{i'=1}^{m} \int g_{ii'}(X_{i,1}, x) d(\widehat{F}_{i'} - F_{i'})(x) - \int g_i(x) d(\widehat{F}_i - F_i)(x) + \epsilon_g$$

satisfies $\mathbb{E}[\epsilon_g^2] = o(n^{-1})$ *as* $n \to \infty$.

As for Assumptions 2.4.3 and 2.4.4, finite-horizon performance measures under mild conditions satisfy the above two assumptions:

Theorem 2.4.7 *Under Assumptions 2.4.1, 2.4.8 and Assumption 2.4.9 with* $k = 4$*, we have Assumptions 2.4.10-2.4.12 hold for the finite-horizon performance measure* ψ *given by* (2.16).

With Assumptions 2.4.11 and 2.4.12, we can identify the statistical error of our variance estimator assuming infinite computation resources, which we summarize in the following lemma.

Lemma 2.4.8 *Under Assumptions 2.4.1, 2.4.3-2.4.4 and 2.4.10-2.4.12, the statistical error of the proportionate subsampled bootstrap variance is characterized by*

$$\sigma_{SVB}^2 - \sigma_I^2 = \mathcal{Z} + \mathcal{R} + o_p(\frac{1}{n^{3/2}} + \frac{1}{ns}) \tag{2.20}$$

where $\mathcal{Z}$ *is a random variable such that*

$$\mathbb{E}[\mathcal{Z}] = 0,\ \mathrm{Var}[\mathcal{Z}] = \sum_{i=1}^{m} \frac{\lambda_i^T \Sigma_i \lambda_i}{n_i}$$

with $\lambda_i = (1/n_i, 2/n_1, \ldots, 2/n_m)^T$ *and*

$$\Sigma_i = \textit{covariance matrix of } (g_i^2(X_i), \mathbb{E}_{X_1'}[g_1(X_1')g_{1i}(X_1', X_i)], \ldots, \mathbb{E}_{X_m'}[g_m(X_m')g_{mi}(X_m', X_i)]).$$

$\mathcal{R}$ *is defined as*

$$\begin{aligned}\mathcal{R} \;=\; & \sum_{i=1}^{m} \frac{1}{n_i s_i} \mathrm{Cov}(g_i(X_i), g_{ii}(X_i, X_i)) + \sum_{i,i'=1}^{m} \frac{1}{n_i s_{i'}} \mathrm{Cov}(g_i(X_i), \mathbb{E}_{X_{i'}'}[g_{ii'i'}(X_i, X_{i'}', X_{i'}')]) \\ & + \sum_{i=1}^{m} \frac{\mathrm{frac}(\theta n_i)\sigma_i^2}{n_i s_i} + \sum_{i,i'=1}^{m} \frac{\mathrm{Var}[g_{ii'}(X_i, X_{i'}')]}{4 n_i s_{i'}}\end{aligned}$$

where $\mathrm{frac}(x) := x - \lfloor x \rfloor$ *denotes the fraction part of* $x \in \mathbb{R}$*, and for each* i*,* X_i, X_i' *are independent copies of the random variable distributed under* F_i.

Combining the statistical error (2.20), and the minimum Monte Carlo error (2.19) under the optimal budget allocation into the trade-off (2.9), we obtain the overall error of the output $\hat{\sigma}^2_{SVB}$ of Algorithm 2:

Theorem 2.4.9 (Overall error of the variance estimate) *Suppose Assumptions 2.4.1, 2.4.3-2.4.7 and 2.4.10-2.4.12 hold. Given a simulation budget N and a subsample ratio θ such that $\frac{N}{\theta n} \to \infty$ and $\theta n \to \infty$, if outer and inner sizes B, R for Algorithm 2 are chosen to be $R = \Theta(\theta n), B = N/R$, then the gross error of our Monte Carlo estimate $\hat{\sigma}^2_{SVB} - \sigma^2_I = \mathcal{E} + o_p(\theta^{1/2}(Nn)^{-1/2} + \theta^{-1}n^{-2} + n^{-3/2})$, where the leading term has a mean squared error*

$$\mathbb{E}[\mathcal{E}^2] = \Theta\big(\frac{\theta}{Nn} + \mathcal{R}^2 + \sum_{i=1}^{m} \frac{\lambda_i^T \Sigma_i \lambda_i}{n_i}\big) \tag{2.21}$$

where $\mathcal{R}$, λ_i's and Σ_i's are defined in Lemma 2.4.8.

It is clear from their definitions in Lemma 2.4.8 that $\mathcal{R} = O(\theta^{-1}n^{-2})$ and each $(\lambda_i^T \Sigma_i \lambda_i)/n_i = O(n^{-3})$, hence the mean squared error (2.21) is in general of order $O(\theta(Nn)^{-1} + \theta^{-2}n^{-4} + n^{-3})$. When $\mathcal{R}$ and at least one of the $\lambda_i^T \Sigma_i \lambda_i$'s satisfy the non-degeneracy condition in Theorem 2.3.6, this bound becomes tight in order, and the optimal subsample ratio can be established by minimizing the order of the leading overall error $\mathcal{E}$.

2.5 Numerical Experiments

This section reports our experimental findings. We consider two examples with different scales and complexities:

M/M/1 queue: The first example we consider is an M/M/1 queue that has true arrival rate 0.5 and service rate 1. Suppose the system is empty at time zero. The performance measure of interest is the probability that the waiting time of the 20-th arrival exceeds 2 units of time, whose true value is approximately 0.182. Specifically, the system has two input distributions, i.e., the inter-arrival time distribution $F_1 = \text{Exp}(0.5)$ and the service time distribution $F_2 = \text{Exp}(1)$, for which we have n_1 and n_2 i.i.d. data available respectively. If A_t is the inter-arrival time between the

t-th and $(t+1)$-th arrivals, and S_t is the service time for the t-th arrival, then the system output

$$\psi(F_1, F_2) = \mathbb{E}_{F_1,F_2}[\mathbf{1}\{W_{20} > 2\}]$$

where the waiting time W_{20} is calculated by the Lindley recursion $W_{t+1} = \max\{W_t + S_t - A_t, 0\}$ for $t = 1, \ldots, 19$ and $W_1 = 0$. To test the proposed approach under different levels of utilization, we also consider true arrival rate 0.9 and service rate 1, for which case the target performance measure is taken to be the probability that the waiting time of the 20-th arrival exceeds 6 units of time (true value 0.190). The data sizes n_1, n_2 are chosen so that $n_1 = 2n_2$ in the experiments, so only the minimum $\min_i n_i$ is reported for convenience.

Computer network: We also consider a computer communication network borrowed from Cheng and Holland (1997) and Lin et al. (2015). The structure of the system is characterized by the undirected graph in Figure 2.1: Four message-processing units, which correspond to the nodes, are connected by four transport channels that are represented by the edges. For every pair i, j

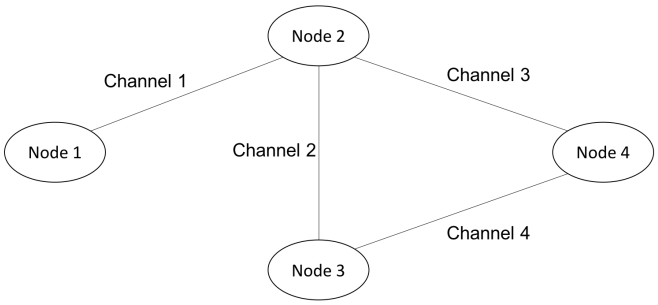


Figure 2.1: A computer network with four nodes and four channels.

of processing units with $i \neq j$, there are external messages that enter into unit i and are to be transmitted to unit j through a fixed path, and their arrival follows a Poisson process with rate $\lambda_{i,j}$. The specific values for $\lambda_{i,j}$'s are summarized in Table 2.1. Each unit takes a constant time of 0.001 seconds to process a message, and has unlimited storage capacity. The messages have lengths that are independent and follow an exponential distribution with mean 300 bits, and each channel has a capacity of 275000 bits, therefore there are queuing and transmission delays. The messages

node i \ node j	1	2	3	4
1	n.a.	40	30	35
2	50	n.a.	45	15
3	60	15	n.a.	20
4	25	30	40	n.a.

Table 2.1: True arrival rates $\lambda_{i,j}$ of messages to be transmitted from node i to node j.

travel through the channels with a velocity of 150000 miles per second, and the i-th channel has a length of $100 \cdot i$ miles for $i = 1, 2, 3, 4$, leading to a propagation delay of $\frac{100 \cdot i}{150000}$ seconds along the i-th channel. The total time that a message of length l bits occupies the i-th channel is therefore $\frac{l}{275000} + \frac{100 \cdot i}{150000}$ seconds. Suppose the system is empty at time zero. The performance measure of interest is the average delay of the first 30 messages that arrive to the system, or mathematically, $\mathbb{E}[\frac{1}{30} \sum_{k=1}^{30} D_k]$, where D_k is the time for the k-th message to be transmitted from its entering node to destination node. The true value of the performance measure is approximately 6.91×10^{-3} seconds. In the experiment, we assume that the arrival rates of the different types of messages, as well as the distribution of the message length, are unknown, therefore there are 13 input models in total. Like in the example of M/M/1 queue, the data sizes across different input models are kept proportional to each other and only the minimum size is reported.

In the experiments we investigate the simulation efforts needed for our subsampling procedure to generate accurate estimates of the input variance, the impacts of the procedural parameters θ, B, R on the estimation accuracy, and practical guidelines on optimal choices of these parameters. Regarding performance metrics of the method, we primarily focus on the mean squared error of the obtained input variance estimate. In addition, note that our estimated input variance can also be used to construct CIs by plugging into formula (2.3). We also examine the quality of these CIs, measured by coverage accuracy and width, as impacted by the estimation accuracy of the input variance.

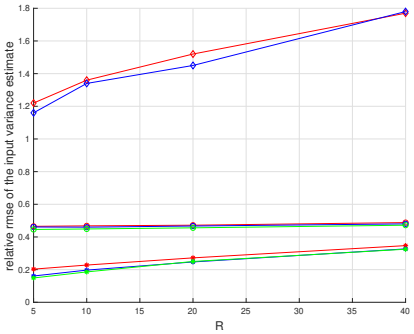
We compare our subsampling approach with the variance bootstrap depicted in Algorithm 1 and the percentile bootstrap suggested by Barton and Schruben (1993, 2001). The percentile bootstrap adopts the same nested simulation structure as in variance bootstrap, but does not estimate the

input variance and instead directly outputs order statistics of the resampled performance measures to construct CIs. Specifically, after obtaining B bootstrapped performance measure estimates $\bar{\psi}^b := \frac{1}{R}\sum_{r=1}^{R}\hat{\psi}_r(\widehat{F}_1^b,\ldots,\widehat{F}_m^b)$, each averaged over R i.i.d. replications, the percentile bootstrap outputs the $\frac{\alpha}{2}(B+1)$-th and $(1-\frac{\alpha}{2})(B+1)$-th order statistics of $\{\bar{\psi}^b : b=1,\ldots,B\}$ as a $(1-\alpha)$-level CI.

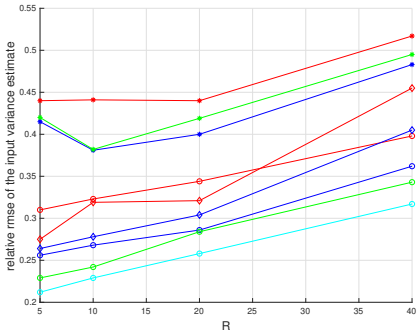
In converting our subsampled input variance estimate to CI, we also investigate the use of a "splitting" versus a "non-splitting" approach. In most part of this section, we use the splitting approach that divides the budget into two portions with one used to estimate the input variance and the other to compute the point estimator. To describe it in detail, suppose we have a total budget of N simulation runs. We allocate R_v simulation runs to estimate σ_I^2 using either Algorithm 1 or 2, and the remaining $R_e = N - R_v$ simulation runs driven by the empirical input distributions to compute the point estimator $\bar{\psi}(\widehat{F}_1,\ldots,\widehat{F}_m)$. When constructing the CI in (2.3), the simulation variance σ_S^2 is calculated as $\frac{\tilde{\tau}^2}{R_e}$, where $\tilde{\tau}^2$ is the sample variance computed from the R_e simulation replications. The second, "non-splitting", approach invests all the N simulation runs in estimating σ_I^2, and constructs the point estimator by averaging all the replications, i.e., $\bar{\psi} = \frac{1}{B}\sum_{b=1}^{B}\bar{\psi}^b$, where $\bar{\psi}^b$ is the performance measure estimate for the b-th resample from Algorithm 2. The simulation variance σ_S^2 in this case is taken to be the sample variance of all the $\bar{\psi}^b$'s divided by the bootstrap size B. The rationale for this approach is that, when the subsample size θn is large, $\mathbb{E}_*[\bar{\psi}]$ should accurately approximate the plug-in estimator $\psi(\widehat{F}_1,\ldots,\widehat{F}_m)$ with an error that is negligible relative to the input variability. Using the former as a surrogate for the latter avoids splitting the budget; however, we will see later that this may introduce too much bias to maintain the desired coverage level when the subsample size is relatively small.

The rest of this section is organized as follows. Section 2.5.1 investigates practical guidelines for choosing the algorithmic parameters in our procedure. Using these guidelines, in Section 2.5.2 we compare the proposed procedure with the variance bootstrap and the percentile bootstrap. Section 2.5.3 studies further the conversion of input variance estimate into CI, and compares the associated splitting and nonsplitting approaches.

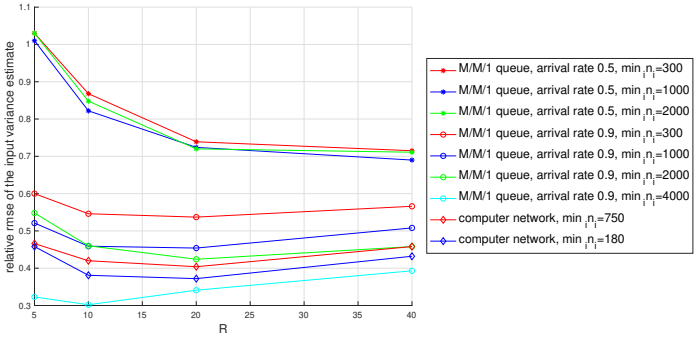
2.5.1 Guidelines for Algorithmic Configuration



(a) $\theta \min_i n_i = 5$.



(b) $\theta \min_i n_i = 30$.



(c) $\theta \min_i n_i = 120$.

Figure 2.2: Input variance estimation accuracy under different configurations of B, R such that $BR = 1000$.

We examine the performances using a wide range of parameter choices for θ, B, R. For each of the two considered examples, and input data sizes from 30 to 2000, we test our subsampling approach at various combinations of θ, B, R where the subsample size $\theta \min_i n_i \in \{5, 15, 30, 60, 120\}$ and the budget allocation parameters $(B, R) \in \{(25, 40), (50, 20), (100, 10), (200, 5)\}$ (a total of 1000 simulation runs). To calculate the mean square error of the input variance estimate, we perform 1000 independent runs of the procedure, each on an independently generated input data set, and then take the average of the squared errors. The reported error metric is the relative root mean squared error (rmse) which can be expressed as $\frac{\sqrt{\mathbb{E}[(\hat{\sigma}_I^2 - \sigma_I^2)^2]}}{\sigma_I^2}$ where $\hat{\sigma}_I^2$ and σ_I^2 are the estimated and true input variances respectively.

We first study and establish guidelines for the outer size B and inner size R for a given subsample

size. Figure 2.2 shows how the estimation error changes as the inner replication size R grows from 5 to 40 (correspondingly the outer size B drops from 200 to 25) and the subsample size $\theta \min_i n_i$ is fixed at a certain value. Each curve represents the results for one of the considered examples under a particular input data size. Although the precise optimal choice for B, R varies from one example to another even when the subsample size is chosen the same, the estimation error appears robust to the parameter choices, with a range of values that only slightly underperform the optimal. In particular, compared to the unknown optimal choice, an R between $\frac{1}{6}\theta \min_i n_i$ and $\frac{1}{3}\theta \min_i n_i$ seems to achieve a comparable accuracy level in the variance estimation, hence is recommended as a general choice.

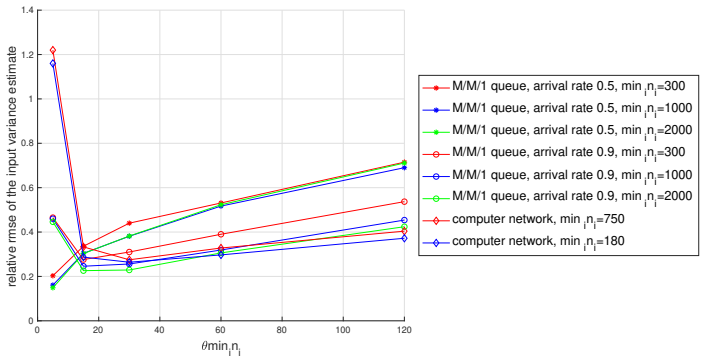


Figure 2.3: Input variance estimation accuracy under different subsample sizes with B, R optimally tuned.

Now we turn to optimal choices for the subsample size. Provided that B, R is properly chosen as above, we examine the behavior of the variance estimation error as the subsample size varies. As we have discussed in Section 2.3.1, subsampling is preferred when the input data size is relatively large, and thus we consider input data sizes ≥ 500 for our M/M/1 queue and computer network, and for each considered data size we plot the variance estimation error versus the subsample size in Figure 2.3. We see that a too large size such as 120 always leads to a larger estimation error than moderate sizes like 30, whereas a too small size around 5 can lift the error by even more in some cases, which is consistent with the theoretical insight from the bound (2.21). Therefore, in general we recommend the use of a subsample size $\theta \min_i n_i$ between 20 and 40 to optimize the estimation accuracy. Figure 2.3 shows that, under the suggested subsample size, the relative rmse is as low as

0.2-0.5 across all the cases.

2.5.2 Comparisons with the Variance Bootstrap and the Percentile Bootstrap

We compare our subsampling method with the standard variance bootstrap and the percentile bootstrap, under the same total budget of 1500 simulation runs. In addition to the relative rmse of the input variance estimate, we also report the actual coverage probability and width of the CI constructed by plugging in the input variance estimate. To estimate all these performance metrics, we construct 1000 95%-level CIs for the target performance measures, each from an independently generated input data set. The "splitting" approach that splits the total budget into $R_v = 1000, R_e = 500$ is adopted for the subsampling approach and the variance bootstrap, whereas for the percentile bootstrap all the 1500 simulation runs are used for the resamples. As suggested in Section 2.5.1, we use the parameter values $\theta = \frac{30}{\min_i n_i}, B = 100, R = 10$ in our method in all the cases, whereas for the other two methods we vary the parameter configurations over a reasonable range constrained by the simulation budget and then report the best results generated by these considered configurations. In particular, the parameters for the variance bootstrap are chosen to minimize the mean square error of the input variance estimate from four combinations, "$B = 25, R = 40$", "$B = 50, R = 20$", "$B = 100, R = 10$", "$B = 200, R = 5$", and those for the percentile bootstrap are chosen to achieve the best the coverage accuracy from four combinations, "$B = 50, R = 30$", "$B = 100, R = 15$", "$B = 300, R = 5$", "$B = 1500, R = 1$". Note that these give an upper hand to our competing alternatives in the comparisons.

Tables 2.2 and 2.3 summarize the experimental results for the M/M/1 queue when the true arrival rate is 0.5 and 0.9 respectively, and Table 2.4 shows those for the computer network. The shorthand "PSVB" stands for proportionate subsampled variance bootstrap, i.e., our subsampling approach. For each method, the "coverage estimate" column displays estimates of the actual coverage probability based on 1000 independent CIs, and the "CI width" column shows their average width. The second column of each table shows the ratio between the input standard error σ_I and the simulation standard error σ_S for different input data sizes in our "splitting" approach. A ratio close to or greater than 1 means that the input noise is a major source of uncertainty relative

to the simulation noise, thus indicating the need to be taken into account in output analysis.

$\min_i n_i$	$\frac{\sigma_I}{\sigma_S}$	PSVB			variance bootstrap			percentile bootstrap	
		relative rmse	coverage estimate	CI width	relative rmse	coverage estimate	CI width	coverage estimate	CI width
30	7.74	0.73	84.3%	0.422	0.73	84.3%	0.422	91.9%	0.467
100	3.77	0.55	92.5%	0.251	0.80	88.6%	0.248	98.8%	0.356
300	2.13	0.44	94.8%	0.156	1.04	85.6%	0.148	99.9%	0.307
1000	1.15	0.38	95.0%	0.103	2.48	89.4%	0.111	100%	0.285
2000	0.79	0.38	95.9%	0.087	5.43	92.8%	0.107	100%	0.280

Table 2.2: Results for the M/M/1 queue with arrival rate 0.5 and service rate 1.

$\min_i n_i$	$\frac{\sigma_I}{\sigma_S}$	PSVB			variance bootstrap			percentile bootstrap	
		relative rmse	coverage estimate	CI width	relative rmse	coverage estimate	CI width	coverage estimate	CI width
30	11.12	0.59	81.4%	0.609	0.59	81.4%	0.609	94.6%	0.639
100	6.22	0.42	89.9%	0.372	0.63	88.6%	0.386	97.2%	0.446
300	3.46	0.32	92.6%	0.225	0.71	87.0%	0.225	99.3%	0.348
1000	1.86	0.27	93.3%	0.137	1.21	86.3%	0.137	100%	0.307
2000	1.30	0.24	95.0%	0.108	2.19	90.7%	0.119	100%	0.294
4000	0.91	0.23	94.9%	0.089	3.61	91.2%	0.106	100%	0.288

Table 2.3: Results for the M/M/1 queue with arrival rate 0.9 and service rate 1.

$\min_i n_i$	$\frac{\sigma_I}{\sigma_S}$	PSVB			variance bootstrap			percentile bootstrap	
		relative rmse	coverage estimate	CI width ($\times 10^{-4}$)	relative rmse	coverage estimate	CI width ($\times 10^{-4}$)	coverage estimate	CI width ($\times 10^{-4}$)
30	12.60	0.74	92.0%	19.3	0.74	92.0%	19.3	95.2%	22.0
150	5.36	0.41	94.3%	8.85	0.53	91.3%	8.50	98.3%	11.2
750	2.35	0.32	94.2%	4.27	0.94	86.9%	3.88	100%	7.97
1800	1.53	0.28	95.3%	3.03	1.63	87.1%	3.01	100%	7.34

Table 2.4: Results for the computer network.

We compare the approaches based on Tables 2.2-2.4. Firstly, our subsampling approach significantly outperforms the variance bootstrap in terms of estimation accuracy of the input variance. The estimates generated by our approach have a smaller relative error than those by the variance bootstrap in all considered cases, and the gap becomes more significant as the data size grows larger. In particular, as the data size grows from 30 to thousands, the estimation error keeps de-

creasing from 0.7 to 0.25 in our approach, whereas in variance bootstrap it keeps increasing from 0.7 to larger than 1, a level that makes the estimate too crude to be useful. These demonstrate the computational advantage and dictate the use of subsampling especially when the input data size is relatively large. Note that the same budget of 1000 simulation runs are used in input variance estimation for all considered data sizes and that the estimation accuracy seems much better for large data sizes than for small sizes, and one may wonder whether more simulation runs should be used for small data sizes to further improve the estimation accuracy. It turns out that the estimation errors are mostly due to the inadequacy of the input data rather than the simulation budget, hence a budget of 1000 is already large enough and further increasing the budget does not bring much benefit. For instance, in the case of data size 30 in Table 2.2, the relative error of the input variance estimate remains as large as 0.69 even if the simulation budget is increased by 10 times.

Secondly, thanks to the high accuracy in the input variance estimates, our subsampling approach generates accurate CIs whose coverage probabilities quickly approach the nominal level 95% as the input data size grows. In contrast, the CIs using the variance bootstrap exhibit under-coverage, and the percentile bootstrap CIs significantly over-cover the truth. We see that the coverage of the variance bootstrap is below 90% in most considered cases, and in the very few cases where the CIs happen to have relatively good coverages, the intervals are much wider than those by our subsampling approach. For example, in the case of data size 2000 in Table 2.2, the variance bootstrap gives a fairly accurate coverage 92.8%, but on average the interval is 1.23 times as wide as that by our method. This shows that the better estimates of the input variance using subsampling translate to better CIs significantly compared to using the variance bootstrap, in terms of both coverage accuracy and width. The percentile bootstrap CIs show an overly high coverage probability close to 100% and are 2-3 times wider than those by subsampling for all considered input data sizes except 30. The over-coverage issue of the percentile CIs arises because the order statistics capture only the input noise but not the simulation noise in the resampled performance measures, a phenomenon that has been discussed in Barton et al. (2007, 2018). When one can afford a sufficiently large budget of simulation relative to the input data size, the simulation noise can be

made negligible so that the CIs have the correct coverage. However, when simulation resources are relatively limited (e.g., when data size ≥ 100 in Tables 2.2-2.4), the CIs are unnecessarily widened by the extra simulation noise that leads to over-coverage. We also notice that the percentile bootstrap CIs do show more accurate coverage than the other two methods when the input data size is 30, which may suggest that the percentile bootstrap is the preferred approach to constructing CIs in small data cases. However, this outperformance is a result of optimally choosing the parameters B, R in hindsight. In our experiments, this best parameter set varies from one case to another, and the actual coverage under different configurations varies in a range of 8%.

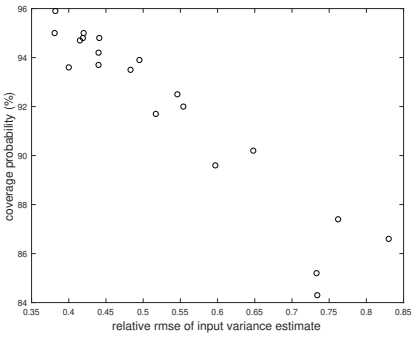
Thirdly, results across different input data sizes show that, the advantages of subsampling in both input variance estimation and CI construction are most significant in situations with relatively large input data size. Note that one may argue in such situations input uncertainty is negligible. However, whether this is indeed the case relates to the error tolerance of the decision-maker and the magnitude of the target performance measure itself. For the large data sizes we consider, the input noise appears still relatively substantial. For instance, when the input data size is 2000 in Table 2.3, the average width of the CIs as a measure of the input uncertainty and simulation uncertainty combined amounts to as much as 57% of the target tail probability, and that the input uncertainty serves as a major component of the total uncertainty (a ratio of 1.3 relative to the simulation uncertainty).

Lastly, in situations with small input data size like 30 the CI coverage clearly falls below 95% in Tables 2.2 and 2.3. This under-coverage phenomenon may appear to stem from the nonlinear effect of the performance measure that is inadequately captured by the Gaussian-approximation-based CI given in (2.3). The real reason, as our experiments suggest, turns out to be the insufficient accuracy of the input variance estimates. In fact, if the true input variance σ_I^2 (which can be accurately estimated by repeatedly generating independent input data sets) is plugged into (2.3) to construct CIs, the coverage probability under the data size 30 rises to 94%-95% for both the M/M/1 queue and the computer network. This indicates a positive impact of an accurate input variance estimate on the CI quality, a point that we will discuss further momentarily.

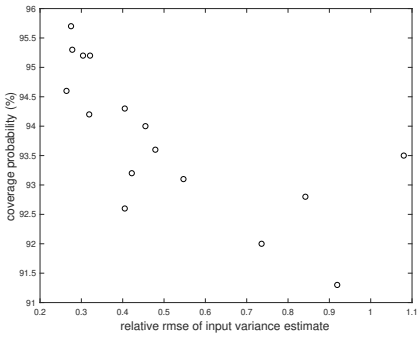
2.5.3 Constructing CI via Input Variance and Comparisons of the Splitting and Nonsplitting Approaches

We study in more depth the relation between the input variance estimation accuracy and CI quality, and compare the splitting approach for CI construction that has been used in previous subsections, with the alternate nonsplitting approach described at the beginning of this section. Finally, we provide practical budget allocation strategies for the splitting approach.

First, to see how the estimation accuracy of the input variance affects the coverage accuracy of the CIs, we use the splitting approach to compute 95%-level CIs, with 1000 simulation runs assigned to input variance estimation and another 500 runs to point estimator evaluation. Figure 2.4a plots the coverage probability versus the relative rmse when the subsample size $\theta \min_i n_i$ is chosen 30 in the M/M/1 queue example, where each point corresponds to a particular combination of the data size $\min_i n_i$, the outer replication size B, and the inner replication size R. Figure 2.4b plots the same for the computer network example with subsample size 30. Both figures clearly show that, the more accurately the input variance is estimated, the closer to the nominal level 95% the coverage probability will be. Accurate estimation of the input variance thus appears to play a crucial role in the construction of accurate CIs.



(a) M/M/1 queue with arrival rate 0.5, $\theta \min_i n_i = 30$.



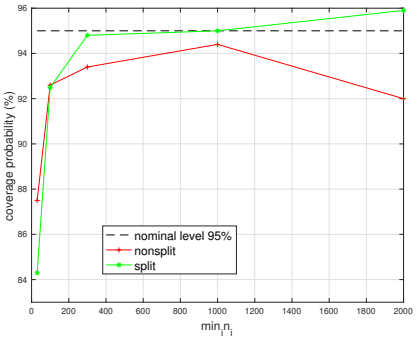
(b) Computer network, $\theta \min_i n_i = 30$.

Figure 2.4: Monotonicity between coverage accuracy and input variance estimation accuracy.

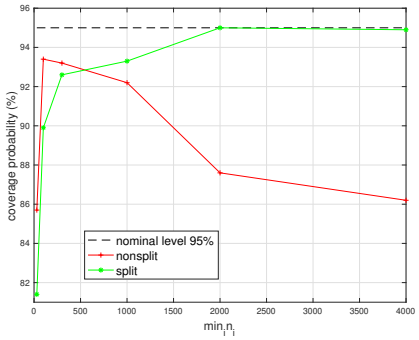
Next we compare the splitting and nonsplitting approaches under the same total budget of 1500 simulation runs. Like in the splitting approach, we use a subsample size $\theta \min_i n_i = 30$ for our nonsplitting approach, but use $B = 75, B = 20$ to consume all the 1500 simulation runs. We

find that the CIs generated from the two approaches have similar lengths, but the nonsplitting approach underperforms in terms of coverage accuracy. Each plot in Figure 2.5 shows the coverage probabilities of the nonsplitting CIs versus the splitting ones for each of the considered example systems, as the input data size grows from 30 to thousands. We see that when the data size is relatively small (e.g., below 500), the two approaches generate CIs with similar coverage accuracy. When the data size grows larger, however, the coverage probability of the nonsplitting CIs keeps dropping in all the three examples, especially in the M/M/1 queue with arrival rate 0.9 where a drop towards 86% is observed, whereas the splitting CIs exhibit almost exact 95% coverage. A possible cause of the undercoverage is the overly small subsample size compared to the input data size, which leads to a high bias in the point estimator. With a subsample size s, the bias of the nonsplitting point estimator $\mathbb{E}_*[\bar{\psi}]$ with respect to the truth $\psi(F_1, \ldots, F_m)$ can be as large as $O(1/s)$. Given that the input standard error is $\Theta(1/\sqrt{n})$, $\mathbb{E}_*[\bar{\psi}]$ has a negligible bias only when the subsample size is large enough, namely when $s = \omega(\sqrt{n})$, indicating that a small subsample size relative to the data size can corrupt the CI. In our experiment, we find that the (supposedly unobservable) bias can be as large as 25% of the CI width when the input data size is 2000 in the M/M/1 queue with arrival rate 0.9, and that artificially removing the bias from the point estimator can improve the coverage to a similar level achieved by the splitting approach. Because of the bias and the consequent under-coverage issue, we caution the use of the non-splitting approach, that it should only be used when a relatively large subsample size is adopted.

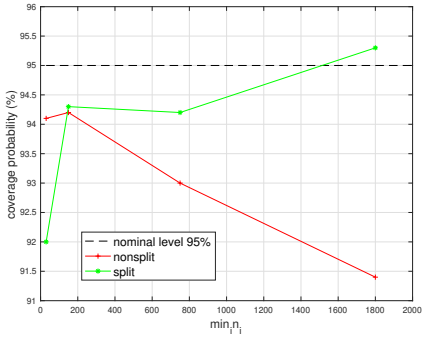
Since the splitting approach is recommended, next we explore strategies of splitting a given budget. Our goal is to generate shortest possible CIs that have a sufficiently accurate coverage probability. As in the beginning of the section, denote by R_v the number of simulation runs used to estimate the input variance, and by R_e to construct the point estimator. Under a fixed total budget $R_v + R_e = 1500$, we try four different splits $R_v = 100, 250, 500, 1000$ (accordingly $R_e = 1500 - R_v$), and for each split the subsample size is fixed at $\theta \min_i n_i = 30$ and several choices of B, R are tested among which the one with the best coverage probability is reported. Figure 2.6 plots the coverage probability versus the CI width for the four considered splits, where the M/M/1 queue with arrival rate 0.9 is considered and input data size is 2000. We notice that the split controls a tradeoff



(a) M/M/1 queue with arrival rate 0.5.



(b) M/M/1 queue with arrival rate 0.9.



(c) Computer network.

Figure 2.5: Coverage comparison under the splitting and nonsplitting approaches.

between the coverage accuracy and the CI width. The more simulation runs one allocates to input variance estimation, the more accurate but wider CIs one would obtain, because the input variance is more accurately estimated while the point estimator becomes more noisy. The plot suggests that allocating 500-1000 replications to variance estimation achieves a good balance of accuracy and width, in the sense that the intervals from the split "500+1000" or "1000 + 500" are only slightly wider than those by other splits and that allocating less (say 250) to variance estimation results in a considerable drop in coverage probability from the nominal level 95%. The results from Tables 2.2-2.4, where the split "1000 + 500" is used, also validates the effectiveness of such a strategy. Therefore, for a given simulation budget, we recommend that the user allocate 500-1000 replications to input variance estimation with our subsampling approach and all the remaining budget to the construction of the point estimator.

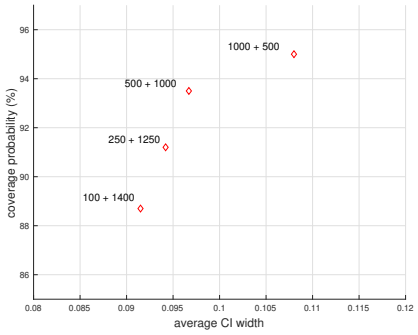


Figure 2.6: Coverage probability versus CI width, under different budget splits in the form of "$R_v + R_e$".

2.6 Conclusion

We have explained how estimating input variances in stochastic simulation can require large computation effort when using conventional bootstrapping. This arises as the bootstrap involves a two-layer sampling, which adds up to a total effort of larger order than the data size in order to achieve relative consistency. To alleviate this issue, we have proposed a subsampling method that leverages the relation between the structure of input variance and the estimation error from the two-layer sampling, so that the resulting total effort can be reduced to being independent of the data size. We have presented the theoretical results in this effort reduction, and the optimal choices of the subsample ratio and simulation budget allocation in terms of the data size and the budget. We have also demonstrated numerical results to support our theoretical findings, and provided guidelines in using our proposed methods to estimate input variances and also construct output CIs.

Chapter 3

Optimization-Based Quantification of Simulation Input Uncertainty via Empirical Likelihood

3.1 Introduction

Following Chapter 2, we continue investigating input uncertainty quantification in this chapter, but switch focus from input variance estimation to the construction of CIs for simulation outputs that account for input uncertainty in addition to the noises in generating the random variates in the simulation process. Again we focus particularly on the nonparametric regime. A common approach is the bootstrap (e.g., Barton and Schruben 1993, 2001), which repeatedly generates resampled distributions to drive simulation runs and uses the quantiles of the simulated outputs to construct the CIs. Another approach is the delta method (e.g., Asmussen and Glynn 2007, Chapter III) that estimates the asymptotic variance in the central limit theorem (CLT) directly. The latter has been considered mostly in the parametric setting (e.g., Cheng and Holland 1997, 1998, 2004) but bears a straightforward analog in our considered nonparametric scenario (as has been discussed in Section 2.2.2). Estimating this variance can also be conducted by subsampling (i.e.,Chapter 2) or bootstrapping (e.g., Cheng and Holland 1997, Song and Nelson 2015).

Our focus in this chapter is a new approach to construct input-induced CIs by using optimization as an underpinning tool. Our approach looks for a set of "maximal" and a set of "minimal" probability weights on the input data, obtained by solving a pair of convex optimization problems with constraints involving a suitably averaged statistical divergence. These weights can be viewed as "worst-case" representations of the input distributions which are then used to generate the input variates to drive the simulation, giving rise to upper and lower bounds that together form a CI on the performance measure of interest.

We will illustrate how this optimization-based approach offers benefits relative to the bootstrap and the delta method. The bootstrap typically involves nested simulation due to the resampling step before simulation runs, which leads to a multiplicative computational requirement. Its performance can be sensitive to the simulation budget size and allocation rule in the nested procedure. A key element of our approach is to use convex optimization to replace the resampling step, which bypasses the multiplicative budget allocation problem and gives more robust performances, especially under small input data sizes. On the other hand, the overhead in setting up our optimization is more substantial than the bootstrap when the input data size is large. It should be noted that the bootstrap possesses more flexibility as the resampled simulation replications can be used to approximate many statistics and to construct CIs at different confidence levels, without re-running the bootstrap procedure again. On the contrary, our approach needs re-optimization and a re-evaluation step for each new confidence level or statistic of interest. Nonetheless, the monotonicity structure of our decision space in the target confidence level allows us to speed up the re-optimization by starting the optimization for a high confidence level at the solution obtained for a low confidence level, while the re-evaluation step only requires a sample size for standard output analysis that is free of input uncertainty.

Our method is closer to the delta method than the bootstrap in that, like the former, we need to estimate gradient information. While our approach and the delta method have similar asymptotic behaviors, we will demonstrate how our approach tends to outperform in finite sample. Roughly speaking, this outperformance arises since the delta method relies heavily on a linear approximation in constructing CIs. When the standard error, which arises from this linear approximation, is noisily

estimated, the resulting CI tends to exhibit under-coverage issues. In contrast, using the weighted distributions to drive simulation runs in our approach is less sensitive to the quality of the standard error estimation, and moreover can introduce nonlinearity that naturally follows the boundaries of a given problem, which in turn alleviates the under-coverage issue experienced in the delta method.

As our main technical contributions, we design and analyze procedures to achieve tight statistical coverage guarantees for the resulting optimization-based CIs. Our approach aligns with the recent surge of robust optimization (Ben-Tal and Nemirovski 2002, Bertsimas et al. 2011) in handling decision-making under uncertainty, where decisions are chosen to perform well under the worst-case scenario among a so-called uncertainty or ambiguity set of possibilities. Our approach particularly resembles distributionally robust optimization (DRO) (e.g., Ben-Tal et al. 2013, Delage and Ye 2010, Goh and Sim 2010, Wiesemann et al. 2014) where the uncertainty of the considered problem lies in the probability distributions, as our involved optimization formulation contains decision variables that are probability weights of the input distributions. However, contrary to the DRO rationale that postulates the uncertainty sets to contain the truth (including those studied recently in the simulation literature; Hu et al. 2012, Glasserman and Xu 2014, Lam 2016b, 2018, Ghosh and Lam 2019), we will explain our procedures by viewing the constraints as log-likelihoods on the input data, and develop the resulting statistical guarantees from a multi-sample generalization of the empirical likelihood (EL) method (Owen 2001), a nonparametric analog of the celebrated maximum likelihood method in parametric statistics. Consequently, the form of our proposed constraint (i.e., the averaged statistical divergence constraint) differs drastically from previous DRO suggestions, and the guarantee is provably tight asymptotically. We mention that, though EL has appeared in statistics for a long time, its use in operations research has appeared only recently and is limited to optimization problems (e.g., Lam and Zhou 2017, Duchi et al. 2016, Lam 2019, Blanchet and Kang 2020, Blanchet et al. 2019). We therefore contribute by showing that a judicious use of this idea can offer new benefits in the equally important area of simulation analysis.

The rest of this chapter is as follows. Section 3.2 reviews some related literature. Section 3.3 presents our procedure and main results on statistical guarantees. Section 3.4 explains the underlying theory giving rise to our approach and statistical results. Section 3.5 shows some

numerical results and compares with previous approaches. All technical proofs are left to Appendix B.

3.2 Related Literature

A review on the problem of input uncertainty has been given in Chapter 2, therefore here we briefly survey two methodologically related areas. Our methodologies are related to several tools in statistics. First is the EL method. Initially proposed by Owen (1988) as a nonparametric counterpart of the maximum likelihood theory, the EL method has been widely studied in statistical problems like regression and hypothesis testing etc. (e.g., Qin and Lawless 1994, Owen 2001, Hjort et al. 2009). Its use in operations research is relatively recent and is limited to optimization. Lam and Zhou (2017) investigates the use of EL in quantifying uncertainty in sample average approximation. Lam (2019) uses EL to derive uncertainty sets for DRO that guarantees feasibility for stochastic constraints. Duchi et al. (2016) generalizes the EL method to Hadamard differentiable functions and obtains tight optimality bounds for stochastic optimization problems. Blanchet and Kang (2020), Blanchet et al. (2019) generalize the EL method to inference using the Wasserstein distance. In addition, our work also utilizes the influence function, which captures nonparametric sensitivity information of a statistic, and is first proposed by Hampel (1974) in the context of robust statistics (Huber and Ronchetti 2009, Hampel et al. 2011) as a heuristic tool to measure the effect of data contamination. Influence function is also used in deriving asymptotic results for von Mises differentiable functionals which have profound applications in U-statistics (Serfling 2009).

Our approach resembles DRO, which utilizes worst-case perspectives in stochastic decision-making problems under ambiguous probability distributions. In particular, our optimization posited over the space of input probability distributions has a similar spirit as the search for the worst-case distribution in the inner optimization in DRO. The DRO framework has been applied in various disciplines such as economics (Hansen and Sargent 2008), finance (Glasserman and Xu 2013, 2014), stochastic control (Petersen et al. 2000, Iyengar 2005, Nilim and El Ghaoui 2005, Xu and Mannor 2012), queueing (Jain et al. 2010) and dynamic pricing (Lim and Shanthikumar 2007). Among

them, constraints in terms of ϕ-divergences, which include the Burg-entropy divergence appearing in our approach, have been considered in, e.g. Ben-Tal et al. (2013), Bayraksan and Love (2015), Jiang and Guan (2016), so are other types of statistical distances such as Renyi divergence (e.g., Atar et al. 2015, Dey and Juneja 2012, Blanchet et al. 2020) and the Wasserstein distance (e.g., Esfahani and Kuhn 2018, Blanchet and Murthy 2019, Gao and Kleywegt 2016), and other constraint types including moments and support (e.g., Delage and Ye 2010, Goh and Sim 2010, Hu et al. 2012, Wiesemann et al. 2014). In simulation, the DRO idea has appeared in Glasserman and Xu (2014), Lam (2016b, 2018), Ghosh and Lam (2019) in quantifying model risks. Nonetheless, although our involved optimization looks similar to DRO, the underpinning statistical guarantees of our approach stem from the EL method. As we will explain, our constraints possess properties that are dramatically different from those studied in DRO, and their precise forms also deviate from any known DRO suggestions.

3.3 Optimization-Based Confidence Intervals

This section presents our main procedure and statistical guarantees. We start with our problem setting and some notations.

3.3.1 Problem Setting

We consider a performance measure in the form

$$Z^* = Z(P_1, \ldots, P_m) = \mathbb{E}_{P_1,\ldots,P_m}\left[h(\mathbf{X}_1, \ldots, \mathbf{X}_m)\right], \tag{3.1}$$

where $P_1, \ldots, P_m$ are m independent input models, $\mathbf{X}_i = (X_i(1), \ldots, X_i(T_i))$ is a sequence of T_i i.i.d. input variates each distributed under P_i, and T_i is a deterministic run length. The distribution P_i has (possibly multivariate) domain $\mathfrak{X}_i$. The function h mapping from $\mathfrak{X}_1^{T_1} \times \cdots \times \mathfrak{X}_m^{T_m}$ to $\mathbb{R}$ is assumed computable given the inputs $\mathbf{X}_i$. In other words, given the sequence $\mathbf{X}_1, \ldots, \mathbf{X}_m$, the value of $h(\mathbf{X}_1, \ldots, \mathbf{X}_m)$ can be evaluated by the computer. The notation $\mathbb{E}_{P_1,\ldots,P_m}[\cdot]$ is a shorthand for $\mathbb{E}_{P_1^{T_1}\times\cdots\times P_m^{T_m}}[\cdot]$, the expectation taken over all the independent i.i.d. sequences $\mathbf{X}_1, \ldots, \mathbf{X}_m$, i.e.,

under the product measure $P_1^{T_1} \times \cdots \times P_m^{T_m}$. We use X_i to denote a generic random variable/vector distributed under P_i.

As a simple example, $\mathbf{X}_1$ and $\mathbf{X}_2$ can represent respectively the sequences of inter-arrival times and service times in a queueing system. P_1 and P_2 represent the corresponding input distributions. h denotes the indicator function of the exceedance of some waiting time above a threshold. Then $Z(P_1, P_2)$ becomes the waiting time tail probability.

Our premise is that there exists a true P_i that is unknown for each i, but a sample of n_i i.i.d. observations $\{X_{i,1}, \ldots, X_{i,n_i}\}$ is available from each P_i. The true value of (3.1) is therefore unknown even under abundant simulation runs. Our goal is to find an asymptotically accurate $(1-\alpha)$-level CI for the true performance measure Z^*. To be more precise, we call a CI asymptotically exact if it consists of two numbers $\mathscr{L}, \mathscr{U}$, derived from the data and the simulation, such that

$$\lim_{\text{each } n_i \text{ and } R\to\infty} P(\mathscr{L} \le Z^* \le \mathscr{U}) = 1 - \alpha$$

where R is the total number of simulation replications involved in generating the CI, and the probability P is taken with respect to the joint randomness in the data and the simulation. The asymptotic above is qualified by certain growth rates of n_i and R that we will detail.

Along our development will also arise cases in which a coverage guarantee is provided as a lower bound, i.e.,

$$\liminf_{\text{each } n_i \text{ and } R\to\infty} P(\mathscr{L} \le Z^* \le \mathscr{U}) \ge 1 - \alpha$$

We call $[\mathscr{L}, \mathscr{U}]$ an asymptotically valid $(1-\alpha)$-level CI. The CIs constructed from our procedures will be either asymptotically exact or, asymptotically valid and accompanied with an associated upper bound that quantities the tightness of the coverage. Lastly, our developments fix the number of independent input models m and the run lengths T_i's, i.e., we focus primarily on transient performance measures with a moderate number of input models relative to the data and simulation sizes.

3.3.2 Main Procedure

Algorithm 4 gives a step-by-step description of our basic procedure for computing $\mathscr{L}$ and $\mathscr{U}$. The quantity $\hat{\tilde{G}}_i(X_{i,j})$ for each $i = 1, \ldots, m, j = 1, \ldots, n_i$ introduced in Step 1 is the sample estimate of the so-called influence function of Z, which can be viewed as the gradient of Z taken with respect to the input distributions (see Assumption 3.3.2 and the subsequent discussion). This sample estimate of the influence function is obtained from R_1 simulation runs. The same simulation replications are also used to compute an estimate $\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)$ of the performance measure under the empirical input models, that is, each $\hat{P}_i(\cdot) = \frac{1}{n_i}\sum_{j=1}^{n_i} \delta_{X_{i,j}}(\cdot)$ where $\delta_{X_{i,j}}(\cdot)$ denotes the delta measure on $X_{i,j}$.

Step 2 in Algorithm 4 outputs a minimizer and a maximizer of the optimization (3.3) in which "min / max" denotes a pair of minimization and maximization, and the calibrating constant $\mathcal{X}^2_{1,1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with degree of freedom one. Optimization (3.3) can be viewed as a sample average approximation (SAA) (Shapiro et al. 2014) on the influence function (expressible as an expectation), with decision variables being the probability weights $w_{i,j}, i = 1, \ldots, m, j = 1, \ldots, n_i$ on the influence function evaluated at each observation $X_{i,j}$ of input model i. For convenience, we denote $\mathbf{w}_i = (w_{i,j})_{j=1,\ldots,n_i}$ as the weight vector associated with input model i, and $\mathbf{w} = (\mathbf{w}_i)_{i=1,\ldots,m}$ be the aggregate weight vector.

Optimization (3.3) can be interpreted as two worst-case optimization problems over m independent input distributions, each on support $\{X_{i,1}, \ldots, X_{i,n_i}\}$, subject to a weighted average of individual statistical divergences (Pardo 2005). To explain, the quantity $D_{n_i}(\mathbf{w}_i) = -\frac{1}{n_i}\sum_{j=1}^{n_i} \log(n_i w_{i,j})$ is the Burg-entropy divergence (Ben-Tal et al. 2013) (or the Kullback-Leibler (KL) divergence) between the probability weights $\mathbf{w}_i$ and the uniform weights. Thus, letting $N = \sum_{i=1}^{m} n_i$ be the total number of observations from all input models, we have

$$-\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log(n_i w_{i,j}) = \sum_{i=1}^{m} \frac{n_i}{N}\left(-\frac{1}{n_i}\sum_{j=1}^{n_i} \log(n_i w_{i,j})\right) = \sum_{i=1}^{m} \frac{n_i}{N} D_{n_i}(\mathbf{w}_i)$$

which is an average of the Burg-entropy divergences imposed on different input models, each weighted by the proportion of the respective observations, n_i/N. The first constraint in (3.3)

Algorithm 4 Basic Empirical-Likelihood-Based Procedure (BEL)

Input: Data $\{X_{i,1},\ldots,X_{i,n_i}\}$ for each input model $i=1,\ldots,m$. A target confidence level $1-\alpha$, and numbers of simulation replications, $R_1, 2R_2$, to be used in Step 1 and Step 3 respectively.

Procedure:

1. Influence Function Estimation: For each $i=1,\ldots,m, j=1,\ldots,n_i$ compute estimate of the influence function evaluated at $X_{i,j}$

$$\hat{\hat{G}}_i(X_{i,j}) = \frac{1}{R_1}\sum_{r=1}^{R_1}\left[(h(\mathbf{X}_1^r,\ldots,\mathbf{X}_m^r) - \hat{Z}(\hat{P}_1,\ldots,\hat{P}_m))(n_i\sum_{t=1}^{T_i}\mathbf{1}\{X_i^r(t)=X_{i,j}\}-T_i)\right] \tag{3.2}$$

where for each $r=1,\ldots,R_1$, $\mathbf{X}_i^r=(X_i^r(1),\ldots,X_i^r(T_i))$ are i.i.d. variates drawn independently from the uniform distribution on $\{X_{i,1},\ldots,X_{i,n_i}\}$ for each i, $\mathbf{1}\{\cdot\}$ is the indicator function, and $\hat{Z}(\hat{P}_1,\ldots,\hat{P}_m)=\sum_{r=1}^{R_1}h(\mathbf{X}_1^r,\ldots,\mathbf{X}_m^r)/R_1$ is the sample mean of the outputs.

2. Optimization: Compute respective optimal solutions $(\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min})$ and $(\mathbf{w}_1^{\max},\ldots,\mathbf{w}_m^{\max})$ of the following pair of programs

$$\begin{aligned}
\min\,/\,\max \quad & \sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{\hat{G}}_i(X_{i,j})w_{i,j} \\
\text{subject to} \quad & -2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}) \le \mathcal{X}^2_{1,1-\alpha} \\
& \sum_{j=1}^{n_i} w_{i,j}=1, \text{ for all } i=1,\ldots,m \\
& w_{i,j}\ge 0, \text{ for all } i=1,\ldots,m, j=1,\ldots,n_i.
\end{aligned} \tag{3.3}$$

3. Evaluation: Compute

$$\mathscr{L}^{BEL} = \frac{1}{R_2}\sum_{r=1}^{R_2}h(\mathbf{X}_1^{r,\min},\ldots,\mathbf{X}_m^{r,\min}),\ \mathscr{U}^{BEL} = \frac{1}{R_2}\sum_{r=1}^{R_2}h(\mathbf{X}_1^{r,\max},\ldots,\mathbf{X}_m^{r,\max})$$

where for each $r=1,\ldots,R_2$, $\mathbf{X}_i^{r,\min}=(X_i^{r,\min}(1),\ldots,X_i^{r,\min}(T_i))$ and $\mathbf{X}_i^{r,\max}=(X_i^{r,\max}(1),\ldots,X_i^{r,\max}(T_i))$ are i.i.d. variates drawn independently from a weighted distribution on $\{X_{i,1},\ldots,X_{i,n_i}\}$, according to weights $\mathbf{w}_i^{\min}$ and $\mathbf{w}_i^{\max}$, respectively for each i.

Output: The CI $[\mathscr{L}^{BEL},\mathscr{U}^{BEL}]$.

can thus be written as

$$\sum_{i=1}^{m} \frac{n_i}{N} D_{n_i}(\mathbf{w}_i) \leq \frac{\mathcal{X}^2_{1,1-\alpha}}{2N}$$

which constitutes a neighborhood ball of size $\mathcal{X}^2_{1,1-\alpha}/(2N)$ measured by the averaged Burg-entropy divergence.

Finally, Step 3 in Algorithm 4 uses the obtained optimal probability weights $\mathbf{w}_i^{\min}$ and $\mathbf{w}_i^{\max}$ to form two weighted empirical distributions on $\{X_{i,j}\}_{j=1,\ldots,n_i}$ for input model i, which are used to drive two independent sets of simulation runs, each of size R_2, in order to output the lower and upper confidence bounds respectively.

An efficient method to solve optimization (3.3) is discussed in the following proposition:

Proposition 3.3.1 *For each i and every $\beta > 0$ define $\lambda_i(\beta)$ to be the unique solution of the equation*

$$\sum_{j=1}^{n_i} \frac{2\beta}{\hat{\hat{G}}_i(X_{i,j}) + \lambda_i} = 1 \tag{3.4}$$

on the interval $(-\min_j \hat{\hat{G}}_i(X_{i,j}), \infty)$. Let $\beta^* > 0$ solve the equation

$$2\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log \frac{2n_i\beta}{\hat{\hat{G}}_i(X_{i,j}) + \lambda_i(\beta)} + \mathcal{X}^2_{1,1-\alpha} = 0. \tag{3.5}$$

If there exist some $i_0 \in \{1,\ldots,m\}$ and $j_1, j_2 \in \{1,\ldots,n_{i_0}\}$ such that $\hat{\hat{G}}_{i_0}(X_{i_0,j_1}) \neq \hat{\hat{G}}_{i_0}(X_{i_0,j_2})$, then $\beta^* \in \big(0, D/\big(2\big(1 - e^{-\frac{\mathcal{X}^2_{1,1-\alpha}}{2N}}\big)\min_i n_i\big)\big)$ and is unique, where $D = \max\{\max_j \hat{\hat{G}}_i(X_{i,j}) - \min_j \hat{\hat{G}}_i(X_{i,j}) | i = 1,\ldots,m\}, N = \sum_{i=1}^m n_i$, and the minimizer $(\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min})$ of (3.3) can be obtained by

$$w_{i,j}^{\min} = \frac{2\beta^*}{\hat{\hat{G}}_i(X_{i,j}) + \lambda_i(\beta^*)}.$$

The maximizer $(\mathbf{w}_1^{\max},\ldots,\mathbf{w}_m^{\max})$ can be computed in the same way except that each $\hat{\hat{G}}_i(X_{i,j})$ is replaced by $-\hat{\hat{G}}_i(X_{i,j})$.

Otherwise, if for each $i = 1,\ldots,m$ the coefficient $\hat{\hat{G}}_i(X_{i,j})$ takes the same value across all $j = 1,\ldots,n_i$, then (3.3) has a constant objective hence becomes trivial. In this case, one can output $(\mathbf{w}_1^{min},\ldots,\mathbf{w}_m^{min})$ and $(\mathbf{w}_1^{max},\ldots,\mathbf{w}_m^{max})$ as any solution in the feasible set (e.g., the uniform

weights, i.e., $w_{i,j}^{\max} = w_{i,j}^{\min} = 1/n_i$ *for all* i, j*).*

The proof of Proposition 3.3.1 uses the Karush-Kuhn-Tucker (KKT) conditions of (3.3), and can be found in Section B.7 of the Appendix. To implement what Proposition 3.3.1 suggests, we first check whether $\hat{\hat{G}}_i(X_{i,j_1}) \neq \hat{\hat{G}}_i(X_{i,j_2})$ for some $i = 1, \ldots, m$ and some $1 \leq j_1 < j_2 \leq n_i$, in which case the maximizer $(\mathbf{w}_1^{\max}, \ldots, \mathbf{w}_m^{\max})$ and minimizer $(\mathbf{w}_1^{\min}, \ldots, \mathbf{w}_m^{\min})$ of (3.3) are unique. Here, given a value of β we can efficiently evaluate each $\lambda_i(\beta)$ by solving (3.4) with Newton's method. Then, β^* is obtained by running a bisection on (3.5) over the interval $(0, D/\big(2\big(1-e^{-\frac{\chi^2_{1,1-\alpha}}{2N}}\big)\min_i n_i\big))$ whose convergence is guaranteed by the uniqueness of β^* and the continuity of (3.5) in β, and finally each $w_{i,j}^{\min}$ or $w_{i,j}^{\max}$ is computed from β^*, $\lambda_i(\beta^*)$'s and $\hat{\hat{G}}_i(X_{i,j})$'s. Note that for any $\beta > 0$ the left hand side of (3.4) is monotonically decreasing and convex in λ_i, hence Newton's method is guaranteed to converge to $\lambda_i(\beta)$ as long as it starts within $(-\min_j \hat{\hat{G}}_i(X_{i,j}), \lambda_i(\beta))$, say at $2\beta - \min_j \hat{\hat{G}}_i(X_{i,j})$. The advantage of this approach over directly solving the convex optimization (3.3) is that we reduce the dimension of the decision space, from linear in the sample sizes to only solving univariate equations in (3.4) and (3.5), which is much more favorable when the sample sizes are large. Finally, note that if $\hat{\hat{G}}_i(X_{i,j_1}) = \hat{\hat{G}}_i(X_{i,j_2})$ for all $i = 1, \ldots, m$ and all $1 \leq j_1 < j_2 \leq n_i$, then, as the objective function of (3.3) becomes a constant, we can take any solution in the feasible set; for concreteness we can take $(\mathbf{w}_1^{\min}, \ldots, \mathbf{w}_m^{\min})$ and $(\mathbf{w}_1^{\max}, \ldots, \mathbf{w}_m^{\max})$ to be the uniform weights.

Next we provide two variants of Algorithm 4, depicted as Algorithms 5 and 6, which differ only by the last step. The motivation (with more details in Section 3.4.6) is that Algorithm 4 tends to under-cover the true performance value because its last step only outputs the sample mean of the simulation replications and does not take full account of the stochastic uncertainty. Algorithm 5 takes care of this uncertainty by outputting the standard normal lower and upper confidence bounds in the last step. However, this simple adjustment does not account for the joint variances from the input data and the stochasticity in a tight manner, and tends to generate conservative CIs that over-cover the truth. This motivates the refined adjustment in Algorithm 6 that is designed to match the CI inflation from combined input and stochastic uncertainties, by taking into account the asymptotic form of the joint variance, and subsequently leads to accurate coverage performances. To explain intuitively, with no simulation noise, an asymptotically exact CI under input uncertainty

would output a point estimate of the target performance measure plus or minus a standard error of magnitude $z_{1-\alpha/2}\sigma_I$, where σ_I^2 denotes the input-induced variance. As we will develop in Section 3.4, using the weights $(\mathbf{w}_1^{\min}, \ldots, \mathbf{w}_m^{\min})$ and $(\mathbf{w}_1^{\max}, \ldots, \mathbf{w}_m^{\max})$ on $\{X_{i,1}, \ldots, X_{i,n_i}\}, i = 1, \ldots, m$, obtained in Step 2 of the algorithms, to evaluate the performance measure essentially matches this CI as long as no simulation noise is present in the evaluation step. When R_2 is finite, however, the standard error should be inflated to $z_{1-\alpha/2}\sqrt{\sigma_I^2 + \sigma^2/R_2}$ where σ^2 denotes the variance of one simulation run. But as $z_{1-\alpha/2}\sigma_I$ is already implicitly elicited by the optimal weights, it needs to be removed from this standard error to avoid over-coverage. This leads to the adjustment in Step 3 of Algorithm 6, where $\hat{\sigma}_I^2$ is an estimate of σ_I^2 and $\hat{\sigma}_{\max}^2$ and $\hat{\sigma}_{\min}^2$ are estimates of σ^2. In the expression of $\hat{\sigma}_I^2$, the sample variance $\sum_{j=1}^{n_i} \big(\hat{\bar{G}}_i(X_{i,j})\big)^2/n_i$ for input model i is upward biased due to the simulation noise in each $\hat{\bar{G}}_i(X_{i,j})$, which is removed by introducing the term $n_i T_i \hat{\sigma}^2/R_1$. This correction term is chosen to match the leading variance of the influence function estimate $\hat{\bar{G}}_i(X_{i,j})$. To explain, the performance function $h - \hat{Z}$ and the multiplier $n_i \sum_{t=1}^{T_i} \mathbf{1}\{X_i^r(t) = X_{i,j}\} - T_i$ that form the product term in (3.2) are nearly independent (see Proposition B.6.2 in Section B.6), hence the variance of each product can be well approximated by the product of the variances of the performance function and the multiplier which give rise to $\hat{\sigma}^2$ and $n_i T_i$ respectively, and the $1/R_1$ factor simply results from the averaging over R_1 replications. The positive-part operation is to handle small R_1 situations where such a variance estimate could yield negative values due to the bias correction, in which case we reset it to zero. Note that the latter occurs increasingly rarely as R_1 grows and would not affect asymptotic properties of the estimate.

3.3.3 Statistical Guarantees

We present statistical guarantees of Algorithms 4, 5 and 6. We assume the following:

Assumption 3.3.1 $\liminf \frac{\min_{i=1,\ldots,m} n_i}{\max_{i=1,\ldots,m} n_i} > 0$ *as all* $n_i \to \infty$.

Assumption 3.3.1 postulates that data sizes across different input models grow at the same rate. For convenience, we shall use the averaged size $n := \frac{1}{m}\sum_{i=1}^m n_i$ to represent the overall scale of the data size throughout the chapter.

Algorithm 5 Evaluation-Adjusted Empirical Likelihood (EEL)

Follow Algorithm 4 until Step 3. Replace Step 3 by

$$\mathscr{L}^{EEL} = \hat{Z}^{\min} - z_{1-\alpha/2}\frac{\hat{\sigma}_{\min}}{\sqrt{R_2}},\ \mathscr{U}^{EEL} = \hat{Z}^{\max} + z_{1-\alpha/2}\frac{\hat{\sigma}_{\max}}{\sqrt{R_2}}$$

where

$$\hat{Z}^{\min} = \frac{1}{R_2}\sum_{r=1}^{R_2} h(\mathbf{X}_1^{r,\min},\ldots,\mathbf{X}_m^{r,\min}),\ \hat{\sigma}^2_{\min} = \frac{1}{R_2-1}\sum_{r=1}^{R_2}(h(\mathbf{X}_1^{r,\min},\ldots,\mathbf{X}_m^{r,\min}) - \hat{Z}^{\min})^2$$

are the sample mean and variance of the R_2 simulation runs driven by distributions on $\{X_{i,1},\ldots,X_{i,n_i}\}$ with weights $\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min}$, and $\hat{Z}^{\max}, \hat{\sigma}^2_{\max}$ are defined accordingly. $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal.

Output: The CI $[\mathscr{L}^{EEL}, \mathscr{U}^{EEL}]$.

Algorithm 6 Fully Adjusted Empirical Likelihood (FEL)

Follow Algorithm 4 until Step 3. Replace Step 3 by

$$\mathscr{L}^{FEL} = \hat{Z}^{\min} - z_{1-\alpha/2}\Big(\sqrt{\hat{\sigma}_I^2 + \frac{\hat{\sigma}^2_{\min}}{R_2}} - \hat{\sigma}_I\Big),\ \mathscr{U}^{FEL} = \hat{Z}^{\max} + z_{1-\alpha/2}\Big(\sqrt{\hat{\sigma}_I^2 + \frac{\hat{\sigma}^2_{\max}}{R_2}} - \hat{\sigma}_I\Big)$$

where $z_{1-\alpha/2}, \hat{Z}^{\min}, \hat{\sigma}^2_{\min}, \hat{Z}^{\max}, \hat{\sigma}^2_{\max}$ are the same as in Algorithm 5, and

$$\hat{\sigma}_I^2 = \max\Big\{\sum_{i=1}^{m}\frac{1}{n_i}\Big[\sum_{j=1}^{n_i}\frac{\big(\hat{\bar{G}}_i(X_{i,j})\big)^2}{n_i} - \frac{n_i T_i \hat{\sigma}^2}{R_1}\Big], 0\Big\},\ \text{with } \hat{\sigma}^2 = \frac{1}{R_1-1}\sum_{r=1}^{R_1}(h(\mathbf{X}_1^r,\ldots,\mathbf{X}_m^r) - \hat{Z})^2 \tag{3.6}$$

is computed from the R_1 replications generated in Step 1.

Output: The CI $[\mathscr{L}^{FEL}, \mathscr{U}^{FEL}]$.

Assumption 3.3.2 *At least one of* $\mathrm{Var}(G_i(X_i)), i = 1, \ldots, m$ *is non-zero, where*

$$G_i(x) = \sum_{t=1}^{T_i} \mathbb{E}_{P_1,\ldots,P_m}[h(\mathbf{X}_1, \ldots, \mathbf{X}_m)|X_i(t) = x] - T_i Z(P_1, \ldots, P_m).$$

Assumption 3.3.3 *For each* i *let* $I_i = (I_i(1), \ldots, I_i(T_i))$ *be a sequence of indices such that* $1 \leq I_i(t) \leq T_i$, *and* $\mathbf{X}_{i,I_i} = (X_i(I_i(1)), \ldots, X_i(I_i(T_i)))$. *Assume* $\mathbb{E}_{P_1,\ldots,P_m}[|h(\mathbf{X}_{1,I_1}, \ldots, \mathbf{X}_{m,I_m})|^8]$ *is finite for all such* I_i*'s.*

The function $G_i(x)$ in Assumption 3.3.2 is the influence function (Hampel 1974, Hampel et al. 2011) of the performance measure $Z(P_1, \ldots, P_m)$ with respect to the input distribution P_i, which measures the infinitesimal effect caused by perturbing P_i and represents the Gateaux derivative of Z in the sense

$$\frac{d}{d\epsilon} Z(P_1, \ldots, P_{i-1}, (1-\epsilon)P_i + \epsilon Q_i, P_{i+1}, \ldots, P_m)\Big|_{\epsilon=0^+} = \int G_i(x) dQ_i(x) \tag{3.7}$$

for any distribution Q_i on $\mathfrak{X}_i$. A rigorous treatment of the validity of (3.7) can be found in Proposition 3.4.1 in Section 3.4.2. Assumption 3.3.2 entails that at least one of the influence functions is non-degenerate at the true input distributions P_i's, or in other words, at least one of these distributions would exert a first-order effect on the performance measure. This assumption is essential in ensuring a normality asymptotic for the output performance measure. In lack of this assumption, the output performance measure will satisfy a χ^2 or even higher-order asymptotic behavior as the input data size grows, which has never been observed in the simulation literature to our best knowledge (the parametric analog of this would be to say that the first-order sensitivities to all input parameters are zero).

Note that the $\hat{\hat{G}}_i(X_{i,j})$ in Step 1 of Algorithm 4 is a sample version of $G_i(X_{i,j})$. Assumption 3.3.3 is a moment condition that, as we will see, controls the magnitude of the linearization error in Step 2 and the simulation error in Steps 1 and 3 of our algorithms. It holds if, for instance, h is bounded.

We have the following statistical guarantees in using the three proposed algorithms to construct

input-induced CIs:

Theorem 3.3.2 *Suppose Assumptions 3.3.1, 3.3.2 and 3.3.3 hold. If the simulation sizes R_1, R_2 are chosen such that $\frac{R_1}{n} \to \infty, \frac{R_2}{n} \to \infty$, then the outputs $\mathscr{L}^{BEL}, \mathscr{U}^{BEL}$ of Algorithm 4 constitute an asymptotically exact $(1-\alpha)$-level CI, i.e.,*

$$\lim_{n,R_1,R_2\to\infty:\ \frac{R_1}{n}\to\infty,\frac{R_2}{n}\to\infty} P\left(\mathscr{L}^{BEL} \le Z^* \le \mathscr{U}^{BEL}\right) = 1-\alpha. \tag{3.8}$$

Theorem 3.3.3 *Suppose Assumptions 3.3.1, 3.3.2 and 3.3.3 hold. If the simulation sizes R_1, R_2 are chosen such that $\frac{R_1}{n} \to \infty, \frac{R_2}{n} \le M$ for some constant $M > 0$, then the outputs $\mathscr{L}^{EEL}, \mathscr{U}^{EEL}$ of Algorithm 5 constitute an asymptotically valid $(1-\alpha)$-level CI, i.e.,*

$$\liminf_{n,R_1,R_2\to\infty:\ \frac{R_1}{n}\to\infty,\frac{R_2}{n}\ bounded} P\left(\mathscr{L}^{EEL} \le Z^* \le \mathscr{U}^{EEL}\right) \ge 1-\alpha$$

$$\limsup_{n,R_1,R_2\to\infty:\ \frac{R_1}{n}\to\infty,\frac{R_2}{n}\ bounded} P\left(\mathscr{L}^{EEL} \le Z^* \le \mathscr{U}^{EEL}\right) \le 1-\tilde{\alpha}+\frac{\tilde{\alpha}^2}{4}$$

where $1-\frac{\tilde{\alpha}}{2} = \Phi(\sqrt{2}z_{1-\alpha/2})$ with Φ being the distribution function of the standard normal. Moreover, if $\frac{R_2}{n} \to \infty$ like in Theorem 3.3.2, then the CI is asymptotically exact, i.e., (3.8) *holds for $\mathscr{L}^{EEL}, \mathscr{U}^{EEL}$.*

Theorem 3.3.4 *Suppose Assumptions 3.3.1, 3.3.2 and 3.3.3 hold. If the simulation sizes R_1, R_2 are chosen such that $\frac{R_1}{n} \to \infty, \frac{R_2}{n} \le M$ for some constant $M > 0$, then the outputs $\mathscr{L}^{FEL}, \mathscr{U}^{FEL}$ of Algorithm 6 constitute an asymptotically valid $(1-\alpha)$-level CI, i.e.,*

$$\liminf_{n,R_1,R_2\to\infty:\ \frac{R_1}{n}\to\infty,\frac{R_2}{n}\ bounded} P\left(\mathscr{L}^{FEL} \le Z^* \le \mathscr{U}^{FEL}\right) \ge 1-\alpha$$

$$\limsup_{n,R_1,R_2\to\infty:\ \frac{R_1}{n}\to\infty,\frac{R_2}{n}\ bounded} P\left(\mathscr{L}^{FEL} \le Z^* \le \mathscr{U}^{FEL}\right) \le 1-\alpha+\frac{\alpha^2}{4}.$$

Moreover, if $\frac{R_2}{n} \to \infty$ like in Theorem 3.3.2, then the CI is asymptotically exact, i.e., (3.8) *holds for $\mathscr{L}^{FEL}, \mathscr{U}^{FEL}$.*

Theorem 3.3.2 states that Algorithm 4 generates an asymptotically exact CI for the true per-

formance measure, when the simulation budgets available to both Step 1 and Step 3 dominate the data size. Theorems 3.3.3 and 3.3.4 show that in Algorithms 5 and 6 the simulation effort for Step 3 can be reduced to grow independent of the data size. This is thanks to the adjustment in the evaluation of the confidence bounds that accounts for the stochastic uncertainty in Step 3. The CI from Algorithm 5 tends to be conservative and can over-cover the truth with a level of $1-\tilde{\alpha}+\tilde{\alpha}^2/2$. To get a sense of this conservativeness, when the desired coverage level $1-\alpha = 90\%$, the guaranteed level can be as high as $1-\tilde{\alpha}+\tilde{\alpha}^2/2 \approx 98\%$. On the other hand, the further refinement in Algorithm 6 is able to recover the exact coverage up to an error of $\alpha^2/4$, which is negligible for most purposes (e.g., when $\alpha = 5\%$, $\alpha^2/4 = 0.0625\%$).

3.4 Theory on Statistical Guarantees

This section further elaborates on Algorithms 4, 5 and 6, and explains the underlying theories leading to Theorems 3.3.2, 3.3.3 and 3.3.4. Section 3.4.1 starts with an initial interpretation of our approach from a distributionally robust optimization (DRO) perspective. The subsequent subsections then discuss the guarantees in several steps. Section 3.4.2 first presents a linear approximation on the performance measures to bypass some statistical and computational bottlenecks. Sections 3.4.3 and 3.4.4 develop the EL method for the linearized problem and CI construction. Section 3.4.5 incorporates the simulation errors. Lastly, Section 3.4.6 discusses the last evaluation steps in our procedures and links them to the conclusions of Theorems 3.3.2, 3.3.3 and 3.3.4.

3.4.1 An Initial Interpretation from DRO

On a high level, our algorithms in Section 3.3.2 can be interpreted as attempting to solve the following problem. Given the observations $\{X_{i,1}, \ldots, X_{i,n_i}\}$ for each input model i, we consider the weighted empirical distribution $\sum_{j=1}^{n_i} w_{i,j}\delta_{X_{i,j}}(\cdot)$. To ease the notation, these weighted distributions will be superseded by their corresponding weight vector $\mathbf{w}_i$ in expressing the performance measure $Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ (and its linear approximations that will be introduced in later sections) evaluated

at these distributions. We consider

$$\begin{aligned} \mathscr{L}/\mathscr{U} := \min / \max \quad & Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) \\ \text{subject to} \quad & \mathbf{w} \in \mathcal{U}_\alpha \end{aligned} \tag{3.9}$$

where

$$\mathcal{U}_\alpha = \left\{ (\mathbf{w}_1, \ldots, \mathbf{w}_m) \in \mathbb{R}^N \middle| \begin{array}{l} -2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log(n_i w_{i,j}) \leq \mathcal{X}^2_{1,1-\alpha} \\ \sum_{j=1}^{n_i} w_{i,j} = 1, \text{ for all } i = 1, \ldots, m \\ w_{i,j} \geq 0, \text{ for all } i, j \end{array} \right\} \tag{3.10}$$

This problem resembles DRO, which is a special class of robust optimization whose uncertainty is on the probability distribution. More specifically, robust optimization considers decision-making under uncertainty or ambiguity of the underlying parameters, and hinges on optimizing the objective under the worst-case scenario, where the worst-case is over all parameters within the so-called uncertainty set or ambiguity set. In DRO, the uncertain quantities are the probability distributions that govern a stochastic optimization, so that the uncertainty set lies in the space of distributions. From this view, optimization (3.9) calculates the worst-case performance measure subject to the uncertainty set $\mathcal{U}_\alpha$. In particular, as discussed in Section 3.3.2, the constraint in (3.10) resembles an averaged Burg-entropy divergence, comprising m terms each being the divergence between the distribution weighted by $\mathbf{w}_i$ and the uniform distribution, on the support generated by the empirical data $\{X_{i,1}, \ldots, X_{i,n_i}\}$.

Despite this Burg-entropy divergence interpretation that ties the optimal weights in (3.9) to "worst-case" distributions, the conceptual reasoning of $\mathcal{U}_\alpha$ that we present below is fundamentally different from DRO. The latter advocates the use of uncertainty sets that contain the true distribution with a certain confidence. To this end, a divergence ball used as an uncertainty set must use a "baseline" distribution that is absolutely continuous to the true distribution, in order to have an overwhelming (or at least non-zero) probability of containing the truth (Jiang and Guan 2016, Esfahani and Kuhn 2018). This condition is violated in formulation (3.9) when the true input distribution is continuous. As the baseline distribution in our divergence (namely the empirical distribution) is supported only on the data, the resulting ball does not contain any continuous

distributions. Moreover, the use of weighted average and its particular weights put on each of these empirically defined divergences is also an unnatural choice from a DRO perspective. Therefore, a key difference between DRO and our approach in terms of coverage guarantees is that, DRO provides simultaneous confidence bounds for all performance measures because of the inclusion of the true distribution in the uncertainty set, whereas our uncertainty set, as we show in later sections, is constructed just rich enough to correctly bound a particular performance measure. In particular, our approach bypasses the inclusion of the truth in the uncertainty set by directly targeting at the uncertainty in the performance measure incurred by the input data.

Thus, instead of arguing the statistical behaviors of (3.9) through the conventional reasoning of DRO, we will explain them using a generalization of the empirical likelihood (EL) method, which is a nonparametric analog of maximum likelihood and endows a tight statistical confidence guarantee in using (3.9) that can be translated to our procedures. Moreover, we also note that, from a computational viewpoint, (3.9) is non-convex and intractable in general. Our procedures as well as statistical developments thus rely on a linearization of the objective function in (3.9). Furthermore, estimating the objective (i.e., the performance measure) and its linearization involves running simulation and incurs the associated errors. The next several subsections detail the linearization, the EL method development, and the sampling error control.

3.4.2 Linearization of Performance Measure

We first state a property related to the influence function in (3.7) that shows up in Assumption 3.3.2:

Proposition 3.4.1 *Let* $(Q_1^1, \ldots, Q_m^1), (Q_1^2, \ldots, Q_m^2)$ *be two sets of distributions such that for any* $s_{i,t} \in \{1, 2\}$ *with* $i = 1, \ldots, m$ *and* $t = 1, \ldots, T_i$

$$\int |h(\mathbf{x}_1, \ldots, \mathbf{x}_m)| \prod_{i=1}^{m} \prod_{t=1}^{T_i} dQ_i^{s_{i,t}}(x_{i,t}) < +\infty,$$

where $\mathbf{x}_i = (x_{i,t})_{t=1,\ldots,T_i}$. *Then*

$$\lim_{\epsilon \to 0+} \frac{1}{\epsilon} \left(Z((1-\epsilon)Q_1^1 + \epsilon Q_1^2, \ldots, (1-\epsilon)Q_m^1 + \epsilon Q_m^2) - Z(Q_1^1, \ldots, Q_m^1) \right) = \sum_{i=1}^{m} \mathbb{E}_{Q_i^2}[G_i^{Q_1^1,\ldots,Q_m^1}(X)], \tag{3.11}$$

where $\mathbb{E}_{Q_i^2}[\cdot]$ *denotes the expectation with respect to* Q_i^2 *that governs* X, *and* $G_i^{Q_1^1,\ldots,Q_m^1}$ *is the influence function of* $Z(Q_1^1, \ldots, Q_m^1)$ *with respect to the distribution* Q_i^1, *given by*

$$G_i^{Q_1^1,\ldots,Q_m^1}(x) = \sum_{t=1}^{T_i} \mathbb{E}_{Q_1^1,\ldots,Q_m^1}[h(\mathbf{X}_1, \ldots, \mathbf{X}_m) | X_i(t) = x] - T_i Z(Q_1^1, \ldots, Q_m^1).$$

Moreover, $\mathbb{E}_{Q_i^1}[G_i^{Q_1^1,\ldots,Q_m^1}(X)] = 0$ *for all* $i = 1, \ldots, m$.

Proposition 3.4.1 can be shown by generalizing the techniques in the asymptotic analysis of von Mises statistical functionals (e.g., Serfling 2009). It suggests the following linear approximation of $Z(Q_1^2, \ldots, Q_m^2)$ around $(Q_1^1, \ldots, Q_m^1)$

$$Z(Q_1^1, \ldots, Q_m^1) + \sum_{i=1}^{m} \mathbb{E}_{Q_i^2}[G_i^{Q_1^1,\ldots,Q_m^1}(X)] \tag{3.12}$$

where the sum consists of expectations of influence functions under Q_i^2 and hence is linear in Q_i^2. In particular, when $Q_i^1 = P_i$, i.e., the true input distribution, and $Q_i^2 = \mathbf{w}_i$ (like at the beginning of Section 3.4.1, we abuse notations slightly here to denote $\mathbf{w}_i$ as the weighted distribution supported on the observations $\{X_{i,j}\}_{j=1,\ldots,n_i}$), (3.12) suggests a linear approximation of $Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ given by

$$Z_L(\mathbf{w}_1, \ldots, \mathbf{w}_m) := Z^* + \sum_{i=1}^{m} \sum_{j=1}^{n_i} G_i(X_{i,j}) w_{i,j} \tag{3.13}$$

where the G_i's are defined in Assumption 3.3.2 and correspond to the influence functions of Z at the true input distributions.

Furthermore, taking $Q_i^1 = \hat{P}_i$, i.e., the empirical input distribution, and $Q_i^2 = \mathbf{w}_i$ in (3.12), we

arrive at the linearization of $Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ around the uniform weights $w_{i,j} = 1/n_i$

$$\widehat{Z_L}(\mathbf{w}_1, \ldots, \mathbf{w}_m) := Z(\hat{P}_1, \ldots, \hat{P}_m) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{G}_i(X_{i,j}) w_{i,j} \tag{3.14}$$

where the $\hat{G}_i$'s are the influence functions of Z at the empirical input distributions, defined by

$$\hat{G}_i(x) = \sum_{t=1}^{T_i} \mathbb{E}_{\hat{P}_1, \ldots, \hat{P}_m}[h(\mathbf{X}_1, \ldots, \mathbf{X}_m) | X_i(t) = x] - T_i Z(\hat{P}_1, \ldots, \hat{P}_m). \tag{3.15}$$

The following result characterizes the quality of the above two linear approximations:

Proposition 3.4.2 *Under Assumptions 3.3.1 and 3.3.3, as the input data size $n \to \infty$ we have*

$$\mathbb{E}\Big[\sup_{(\mathbf{w}_1, \ldots, \mathbf{w}_m) \in \mathcal{U}_\alpha} \big| Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - Z_L(\mathbf{w}_1, \ldots, \mathbf{w}_m) \big|^2 \Big] = O\big(\frac{1}{n^2}\big) \tag{3.16}$$

$$\mathbb{E}\Big[\sup_{(\mathbf{w}_1, \ldots, \mathbf{w}_m) \in \mathcal{U}_\alpha} \big| Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - \widehat{Z_L}(\mathbf{w}_1, \ldots, \mathbf{w}_m) \big|^2 \Big] = O\big(\frac{1}{n^2}\big) \tag{3.17}$$

where $\mathcal{U}_\alpha$ is defined in (3.10)*, and the expectation is with respect to the input data.*

Proposition 3.4.2 suggests that, restricting to $\mathcal{U}_\alpha$, the maximum deviations of the linear approximations from the true performance measure vanish as fast as $1/n$. Next we will build the theories and explain our procedures for a linearized performance measure, and relate them back to the original nonlinear performance measure Z through Proposition 3.4.2.

3.4.3 Empirical Likelihood Theory for Sums of Means

First proposed by Owen (1988), the EL method can be viewed as a nonparametric counterpart of the maximum likelihood theory. Here we will develop this method for the linear approximation Z_L. Note that the second term in (3.13) can be viewed as the sum of the expected influence function values under the weighted empirical distributions. Therefore, to ease notation and emphasize its generality, we will present our EL method as a generic inference tool for estimating sums of means.

Suppose we are given m independent samples of i.i.d. observations $\{Y_{i,1}, \ldots, Y_{i,n_i}\}, i = 1, \ldots, m$, with $Y_{i,j}, j = 1, \ldots, n_i$ distributed according to a common distribution F_i. For the i-th sample, we

define its nonparametric likelihood, in terms of the probability weights $\mathbf{w}_i$ over the support points of the data, to be $\prod_{j=1}^{n_i} w_{i,j}$. The multi-sample likelihood is $\prod_{i=1}^{m}\prod_{j=1}^{n_i} w_{i,j}$. By a simple convexity argument, it can be shown that assigning uniform weights $w_{i,j} = 1/n_i$ for each sample yields the maximum value $\prod_{i=1}^{m}(1/n_i)^{n_i}$. Moreover, uniform weights still maximize even if one allows putting weights outside the support of data, in which case $\sum_{j=1}^{n_i} w_{i,j} < 1$ for some i, making $\prod_{j=1}^{n_i} w_{i,j}$ even smaller. Therefore, the uniform weights $w_{i,j} = 1/n_i$ for all $j = 1, \ldots, n_i$ can be viewed as the nonparametric maximum likelihood estimate for the i-th distribution F_i, and $w_{i,j} = 1/n_i$ for all i, j is the multi-sample counterpart.

To proceed, we need to define a parameter of interest that is determined by the distributions F_i's. In our case, the parameter of interest is the sum of means $\mu_0 := \sum_{i=1}^{m} \mathbb{E}Y_i$ where each Y_i is distributed under F_i.

The key of the EL method is to establish limit theorems analogous to the celebrated Wilks' Theorem (Wilks (1938)) in the maximum likelihood theory, which stipulates that a suitably defined logarithmic likelihood ratio converges to a $\mathcal{X}^2$ random variable. In the EL setting, we use the so-called profile nonparametric likelihood ratio to carry out inference on parameters. To explain this, first, the nonparametric likelihood ratio is defined as the ratio between the nonparametric likelihood of a given set of weights and the uniform weights (i.e., the nonparametric maximum likelihood estimate). The profile nonparametric likelihood ratio is defined as the maximum ratio among all probability weights giving rise to a particular value μ for the sum of means, i.e.,

$$R(\mu) = \max\left\{ \prod_{i=1}^{m}\prod_{j=1}^{n_i} n_i w_{i,j} \,\Big|\, \sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = \mu,\ \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for all } i,\ w_{i,j} \geq 0 \text{ for all } i, j \right\}, \tag{3.18}$$

and is defined to be 0 if the optimization problem in (3.18) is infeasible. Profiling here refers to the categorization of weights that lead to the same value μ.

The quantity $R(\mu)$ satisfies the following asymptotic property:

Theorem 3.4.3 *Let Y_i be a random variable distributed under F_i. Assume $\mathrm{Var}(Y_i) < \infty$ for all $i = 1, \ldots, m$ and at least one of them is non-zero, and that the sample sizes n_i's satisfy Assumption 3.3.1. Then $-2\log R(\mu_0)$, where μ_0 is the sum of the true means, converges in distribution to $\mathcal{X}_1^2$,*

the chi-square distribution with degree of freedom one, as $n \to \infty$.

In other words, the logarithmic profile nonparametric likelihood ratio at the true value asymptotically follows a chi-square distribution with degree of freedom one. This degree of freedom is the effective number of parameters to be estimated which, in this case, is one since there is only a single target parameter μ_0. Note that this is independent of the number of input distributions m.

Theorem 3.4.3 is a sum-of-mean generalization of the well-known empirical likelihood theorem (ELT) for single-sample mean:

Theorem 3.4.4 (Owen (2001) Theorem 2.2) *Consider only the first sample* $\{Y_{1,1}, \ldots, Y_{1,n_1}\}$. *Assume* $0 < \mathrm{Var}(Y_1) < \infty$. *Then* $-2 \log R(\mathbb{E}Y_1)$ *converges in distribution to* $\mathcal{X}_1^2$, *as* $n_1 \to \infty$. *The function* $R(\cdot)$ *here is the same as that in* (3.18) *but with* $m = 1$.

Extensions of this theorem have been studied in the literature (e.g., Owen 1990, 1991, Qin and Lawless 1994, Hjort et al. 2009). The most relevant one is in the context of analysis-of-variance (ANOVA), in which the logarithmic profile nonparametric likelihood ratio at the true means of multiple independent samples are shown to converge to $\mathcal{X}_m^2$, where m is the number of samples (or groups). However, the argument for this result relies on viewing the multiple samples as a collection of heteroscedastic data and applying the triangular array ELT (Owen 1991), which does not apply obviously to our case. Another related extension is the plug-in EL (Hjort et al. 2009) which entails that, under p estimating functions that possibly involve unknown nuisance parameters, the associated logarithmic profile likelihood ratio converges to a weighted sum of p independent $\mathcal{X}_1^2$'s, if "good enough" estimators of the unknown nuisance parameters are used in evaluating the profile likelihood ratio. However, Hjort et al. (2009) focuses on the single-sample case, thus is not directly applicable. There have also been studies on applying EL to hypothesis testing of two-sample mean differences (Liu et al. 2008, Wu and Yan 2012), but it appears that a fully rigorous proof is not available for our general multi-sample sum-of-means setting. In view of these, we provide a detailed proof of Theorem 3.4.3 in Section B.3 of the Appendix.

A sketch of the key idea is as follows. We first introduce the auxiliary variables μ_i that represent the means of individual samples, so that the constraint $\sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = \mu$ in (3.18) is replaced

by $\sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = \mu_i, i = 1, \ldots, m$ and $\sum_{i=1}^{m} \mu_i = \mu$. The KKT conditions then enforce the optimal weights to be

$$w_{i,j}^* = \frac{1}{n_i + \lambda^*(Y_{i,j} - \mu_i^*)}$$

where λ^* is the Lagrange multiplier for the constraint $\sum_{i=1}^{m} \mu_i = \mu$ and μ_i^* is the optimal solution for μ_i. When μ is the true value μ_0, an asymptotic analysis on the KKT conditions approximates λ^* as

$$\lambda^* \approx \frac{\sum_{i=1}^{m} (\bar{Y}_i - \mathbb{E}Y_i)}{\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}}$$

where $\bar{Y}_i = (1/n_i) \sum_{j=1}^{n_i} Y_{i,j}$ is the sample mean and σ_i^2 is the variance of Y_i. Moreover, we have the approximation $\mu_i^* \approx \mathbb{E}Y_i$. By Taylor's expansion, the logarithmic profile nonparametric likelihood ratio can be approximated as

$$\begin{aligned}
-2 \log R(\mu_0) &= 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log \left(1 + \frac{\lambda^*}{n_i} (Y_{i,j} - \mu_i^*) \right) \\
&\approx 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\frac{\lambda^*}{n_i} (Y_{i,j} - \mu_i^*) - \frac{\lambda^{*2}}{2n_i^2} (Y_{i,j} - \mu_i^*)^2 \right) \\
&\approx 2 \sum_{i=1}^{m} \lambda^* (\bar{Y}_i - \mu_i^*) - \sum_{i=1}^{m} \frac{\lambda^{*2} \sigma_i^2}{n_i} \\
&\approx \left(\frac{\sum_{i=1}^{m} (\bar{Y}_i - \mathbb{E}Y_i)}{\sqrt{\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}}} \right)^2 \\
&\Rightarrow \mathcal{X}_1^2
\end{aligned}$$

where "$\Rightarrow$" denotes convergence in distribution. This gives our result in Theorem 3.4.3.

3.4.4 Duality and Optimization-Based Confidence Intervals

From Theorem 3.4.3, a duality-type argument will give rise to a pair of optimization problems whose optimal values will serve as confidence bounds for the sum of the true means. We have the following:

Theorem 3.4.5 *Under the same conditions of Theorem 3.4.3, we have*

$$\lim_{n\to\infty} P\left(\underline{\mu} \le \mu_0 \le \overline{\mu}\right) = 1-\alpha$$

where

$$\underline{\mu}/\overline{\mu} := \min / \max \Big\{ \sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{i,j} w_{i,j} \Big| (\mathbf{w}_1, \ldots, \mathbf{w}_m) \in \mathcal{U}_\alpha \Big\}. \tag{3.19}$$

Theorem 3.4.5 thus translates the asymptotic convergence in Theorem 3.4.3 into an asymptotically exact confidence bound. This is argued by a duality argument that turns the first constraint in (3.18) into objective and vice versa. The concept is similar to Wilks' Theorem for maximum likelihood, but with the profiling that leads to the resulting optimization problems in (3.19).

Moreover, in terms of their positions, the optimization-based confidence bounds $\underline{\mu}$ and $\overline{\mu}$ are equivalent to the standard normality-based confidence bounds up to negligible errors, as described below:

Proposition 3.4.6 *Under the same conditions of Theorem 3.4.3, the confidence bounds $\underline{\mu}, \overline{\mu}$ from Theorem 3.4.5 satisfy*

$$\underline{\mu} = \sum_{i=1}^{m} \bar{Y}_i - z_{1-\alpha/2}\sqrt{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}} + o_p\big(\frac{1}{\sqrt{n}}\big)$$
$$\overline{\mu} = \sum_{i=1}^{m} \bar{Y}_i + z_{1-\alpha/2}\sqrt{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}} + o_p\big(\frac{1}{\sqrt{n}}\big)$$

where $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{i,j}/n_i$ is the sample mean of $\{Y_{i,1}, \ldots, Y_{i,n_i}\}$ and σ_i^2 is the true variance of Y_i, and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal.

The errors between $\underline{\mu}$, $\overline{\mu}$ and the normality-based bounds $\sum_{i=1}^{m} \bar{Y}_i \pm z_{1-\alpha/2}\sqrt{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}}$ are negligible in the sense that they are of smaller order than the width of the resulting CI, which is of order $1/\sqrt{n}$.

Applying the above two results to the linear approximation Z_L, we have the following:

Corollary 3.4.7 *Under Assumptions 3.3.1, 3.3.2 and 3.3.3, we have*

$$\lim_{n\to\infty} P\left(\mathscr{L}_L \le Z^* \le \mathscr{U}_L\right) = 1-\alpha \tag{3.20}$$

where

$$\mathscr{L}_L/\mathscr{U}_L := \min / \max \left\{ Z_L(\mathbf{w}_1,\ldots,\mathbf{w}_m) \middle| (\mathbf{w}_1,\ldots,\mathbf{w}_m) \in \mathcal{U}_\alpha \right\}. \tag{3.21}$$

Moreover

$$\begin{aligned} \mathscr{L}_L &= Z^* + \sum_{i=1}^m \bar{G}_i - z_{1-\alpha/2}\sigma_I + o_p\big(\frac{1}{\sqrt{n}}\big) \\ \mathscr{U}_L &= Z^* + \sum_{i=1}^m \bar{G}_i + z_{1-\alpha/2}\sigma_I + o_p\big(\frac{1}{\sqrt{n}}\big) \end{aligned} \tag{3.22}$$

where each $\bar{G}_i = \sum_{j=1}^{n_i} G_i(X_{i,j})/n_i$ *is the sample mean of the influence function values* $G_i(X_{i,j}), j = 1,\ldots,n_i$, $\sigma_I^2 = \sum_{i=1}^m \mathrm{Var}(G_i(X_i))/n_i$, *and* $z_{1-\alpha/2}$ *is the* $1-\alpha/2$ *quantile of the standard normal.*

Note that the influence functions in (3.13) satisfy $\sum_{i=1}^m \mathbb{E}[G_i(X_i)] = 0$ due to the last claim in Proposition 3.4.1. Thus, letting $Y_{i,j} = G_i(X_{i,j})$ in Theorem 3.4.5 and Proposition 3.4.6, and noting that the Z^* in (3.20) and (3.22) can be cancelled out, we arrive at the conclusion in Corollary 3.4.7.

Next, combining Corollary 3.4.7 and the linearization error (3.16), we can establish similar results for $\mathscr{L}, \mathscr{U}$ that arise in (3.9):

Theorem 3.4.8 *Under Assumptions 3.3.1, 3.3.2 and 3.3.3, the minimum and maximum values* $\mathscr{L}, \mathscr{U}$ *of* (3.9) *satisfy*

$$\lim_{n\to\infty} P\left(\mathscr{L} \le Z^* \le \mathscr{U}\right) = 1-\alpha,$$

and the asymptotic equivalence (3.22) *holds true with* $\mathscr{L}_L$, $\mathscr{U}_L$ *replaced by* $\mathscr{L}$, $\mathscr{U}$.

The proof of Theorem 3.4.8 consists of first approximating the discrepancies between the optimal values, i.e., $\mathscr{L} = \mathscr{L}_L + O_p(1/n)$ and $\mathscr{U} = \mathscr{U}_L + O_p(1/n)$, using (3.16), and then showing that any quantities that equal (3.22), up to a small order of discrepancies, deliver an interval with asymptotically exact coverage probability by a standard application of Slutsky's Theorem.

3.4.5 Estimating Influence Function

Our proposed CIs in Algorithms 4, 5 and 6 use a combination of the intervals suggested in Corollary 3.4.7 and Theorem 3.4.8. Before we explain this concretely, note that directly using the definition of $\mathscr{L}, \mathscr{U}$ in (3.9) will encounter computational difficulties due to the general intractability of the optimization. Thus, we consider using optimization (3.21) or expression (3.22) in obtaining our confidence bounds. In either case, we need to estimate the influence function represented by $G_i(X_{i,j})$'s.

There are two sources of errors in estimating $G_i(X_{i,j})$. First, since we do not know the true distribution P_i, we approximate it by the influence function evaluated at the empirical distribution, namely $\hat{G}_i(X_{i,j})$ defined in (3.15) (which in turn forms the coefficient in $\widehat{Z_L}$). Second, $\hat{G}_i(X_{i,j})$, like $G_i(X_{i,j})$, is a sum of conditional expectations, which needs to be estimated by simulation. Ghosh and Lam (2019, 2015) propose an unbiased estimator for such quantities where the input distributions have arbitrary weights $w_{i,j}$ on their support points. Here we use their scheme for the special case of uniform weights. Similar approaches also arise in the so-called infinitesimal jackknife for bagging estimators (e.g., Efron (2014), Wager et al. (2014)). Proposition 3.4.9 shows the scheme (see Ghosh and Lam 2019 for the proof).

Proposition 3.4.9 *Given input data $\{X_{i,j}\}$, the empirical influence function $\hat{G}_i$ evaluated at data point $X_{i,j}$ satisfies*

$$\hat{G}_i(X_{i,j}) = \mathrm{Cov}_{\hat{P}_1,\ldots,\hat{P}_m}(h(\mathbf{X}_1,\ldots,\mathbf{X}_m), S_{i,j}(\mathbf{X}_i)),$$

where $\mathrm{Cov}_{\hat{P}_1,\ldots,\hat{P}_m}$ denotes the covariance under the empirical input distributions, and

$$S_{i,j}(\mathbf{X}_i) = \sum_{t=1}^{T_i} n_i \mathbf{1}\{X_i(t) = X_{i,j}\} - T_i.$$

Such a covariance interpretation of the influence function leads us to the Monte Carlo estimate (3.2) of $\hat{G}_i(X_{i,j})$ in Step 1, denoted $\hat{\hat{G}}_i(X_{i,j})$, that takes the form of a sample covariance from R_1

simulation runs. Next, we introduce a sampled linear approximation for $Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ given by

$$\widehat{\widehat{Z}}_L(\mathbf{w}_1, \ldots, \mathbf{w}_m) := \hat{Z}(\hat{P}_1, \ldots, \hat{P}_m) + \sum_{i=1}^{m} \sum_{j=1}^{n_i} \hat{\bar{G}}_i(X_{i,j}) w_{i,j}. \tag{3.23}$$

where $\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)$ is the sample mean of the R_1 replications. Optimization (3.3) in Step 2 of the procedures uses $\widehat{\widehat{Z}}_L(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ as the objective function. But since $\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)$ does not depend on the weights $w_{i,j}$'s, it is dropped from the expression.

The quality of the sample linear approximation (3.23) is quantified as:

Proposition 3.4.10 *Under Assumptions 3.3.1 and 3.3.3, as the input data size $n \to \infty$ and simulation effort $R_1 \to \infty$ we have $\mathbb{E}\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|\widehat{Z}_L(\mathbf{w}_1, \ldots, \mathbf{w}_m) - \widehat{\widehat{Z}}_L(\mathbf{w}_1, \ldots, \mathbf{w}_m)\big|^2\big] = O\big(\frac{1}{R_1}\big)$, hence together with (3.17) we have*

$$\mathbb{E}\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - \widehat{\widehat{Z}}_L(\mathbf{w}_1, \ldots, \mathbf{w}_m)\big|^2\big] = O\big(\frac{1}{n^2} + \frac{1}{R_1}\big). \tag{3.24}$$

Here both expectations are with respect to the input data and the simulation runs in Step 1 of Algorithms 4-6.

The uniform error (3.24) of $\widehat{\widehat{Z}}_L$ as an approximation to Z then implies the following guarantee on the difference between the weights $\{\mathbf{w}_i^{\min}\}_{i=1}^m, \{\mathbf{w}_i^{\max}\}_{i=1}^m$ obtained in Step 2 of Algorithm 4, and the optimal weights for the optimization pair (3.9), measured in terms of their evaluations of the performance measure Z:

Theorem 3.4.11 *Let $Z^{\min} := Z(\mathbf{w}_1^{\min}, \ldots, \mathbf{w}_m^{\min})$ and $Z^{\max} := Z(\mathbf{w}_1^{\max}, \ldots, \mathbf{w}_m^{\max})$. Under Assumptions 3.3.1 and 3.3.3, as the input data size $n \to \infty$ and simulation effort $R_1 \to \infty$ we have*

$$\mathbb{E}[(Z^{\min} - \mathscr{L})^2] = O\big(\frac{1}{n^2} + \frac{1}{R_1}\big),\ \mathbb{E}[(Z^{\max} - \mathscr{U})^2] = O\big(\frac{1}{n^2} + \frac{1}{R_1}\big)$$

where $\mathscr{L}, \mathscr{U}$ are defined in (3.9), and the expectation is with respect to the input data and the simulation runs in Step 1 of Algorithms 4-6.

Theorem 3.4.11 justifies using $\{\mathbf{w}_i^{\min}\}_{i=1}^m, \{\mathbf{w}_i^{\max}\}_{i=1}^m$ to evaluate the performance measure, which

give rise to the asymptotically exact confidence bounds $\mathscr{L}, \mathscr{U}$ up to a small-order error. Step 3 of the algorithms utilizes this implication. However, we need to properly control the simulation error in evaluating the performance measure, which is detailed in the next subsection.

As a side note, we can also use the linear approximation $\widehat{\widehat{Z}}_L$ evaluated at the weights $\{\mathbf{w}_i^{\min}\}_{i=1}^m$ and $\{\mathbf{w}_i^{\max}\}_{i=1}^m$ directly as our confidence interval. This forms another asymptotically exact CI (see Theorem B.5.3 in Appendix B.5). Moreover, this approach would require less simulation effort than our procedures (R_1 versus $R_1 + 2R_2$). However, like the delta method, this approach relies heavily on the linear approximation to construct the CI. In contrast, the CIs in our procedures are constructed from simulating the (nonlinear) performance measure, under the carefully chosen empirical weights $\{\mathbf{w}_i^{\min}\}_{i=1}^m, \{\mathbf{w}_i^{\max}\}_{i=1}^m$. As a result, they conform more closely to the boundaries of a given problem and in turn can lead to better coverages. For example, when the performance measure is within a range (e.g., a probability that is between 0 and 1), using only the linear approximation frequently incurs under-coverage as the CIs can lie significantly outside the meaningful range (note that truncating at the boundaries would not solve the issue, which is intrinsic in the linear approximation), whereas our procedures would generate confidence bounds that much more often lie within the range and consequently offer better coverages.

3.4.6 Evaluation of CI Bounds

This section explains and compares Step 3 in Algorithms 4, 5 and 6 to evaluate the final confidence bounds, and relates these to the justification of Theorems 3.3.2, 3.3.3 and 3.3.4.

Algorithm 4 constructs CIs by taking averages of R_2 independent simulation runs driven by the weighted empirical input distributions, with weights being $\{\mathbf{w}_i^{\min}\}_{i=1}^m, \{\mathbf{w}_i^{\max}\}_{i=1}^m$, to evaluate the lower and upper bounds respectively. Note that by Theorem 3.4.11, the performance measures evaluated at the weighted empirical distributions, $Z^{\min}$ and $Z^{\max}$, are close to $\mathscr{L}$ and $\mathscr{U}$, which in turn by Theorem 3.4.8 satisfy exact coverage guarantees. Step 3 of Algorithm 4 adds simulation noises from the R_2 simulation runs in estimating $Z^{\min}$ and $Z^{\max}$. This results in the following discrepancies between the outputs of Algorithm 4 and $\mathscr{L}$, $\mathscr{U}$:

Proposition 3.4.12 *Under Assumptions 3.3.1 and 3.3.3, as the input data size $n \to \infty$ and*

simulation effort $R_1 \to \infty, R_2 \to \infty$, the outputs $\mathscr{L}^{BEL}, \mathscr{U}^{BEL}$ of Algorithm 4 satisfy

$$\mathbb{E}[(\mathscr{L}^{BEL} - \mathscr{L})^2] = O\big(\frac{1}{n^2} + \frac{1}{R_1} + \frac{1}{R_2}\big),\ \mathbb{E}[(\mathscr{U}^{BEL} - \mathscr{U})^2] = O\big(\frac{1}{n^2} + \frac{1}{R_1} + \frac{1}{R_2}\big)$$

where the expectation is with respect to the input data and the simulation runs in Steps 1 and 3 of Algorithm 4.

Proposition 3.4.12 implies that, when the simulation sizes R_1 and R_2 both dominate the input data size n, the root-mean-square discrepancies between the outputs from Algorithm 4, $\mathscr{L}^{BEL}$, $\mathscr{U}^{BEL}$, and the asymptotically exact CIs formed by $\mathscr{L}$, $\mathscr{U}$, become $o(1/\sqrt{n})$, which is of smaller order than the width of the CI that is of order $1/\sqrt{n}$. This then leads to the asymptotic exactness of $[\mathscr{L}^{BEL}, \mathscr{U}^{BEL}]$ in Theorem 3.3.2.

Algorithm 4 requires both R_1 and R_2 to be large relative to n. Algorithms 5 and 6, on the other hand, are designed to work well for smaller R_2. To explain, note that the reason of needing R_2 to be large in Algorithm 4 is to wash away the simulation noises to a smaller magnitude than the CI width in Step 3. Instead of simply washing them away, Algorithms 5 and 6 suitably enlarge the CI to incorporate these errors in Step 3, so that R_2 can now be chosen independent of n. In particular, Algorithm 5 simply takes the upper or lower confidence bound for $Z^{\max}$ or $Z^{\min}$ respectively to account for the simulation noises. Algorithm 6, on the other hand, uses a more delicate adjustment reasoned from the representation (3.22). Consider the upper bound $\mathscr{U}^{FEL}$ for instance. Theorem 3.4.11, the interchangeability between $\mathscr{L}, \mathscr{U}$ and $\mathscr{L}_L, \mathscr{U}_L$ in Theorem 3.4.8, and the expressions in (3.22) together stipulate that, with no simulation error in Step 3 of the algorithm, an asymptotically exact upper bound can be expressed as $Z^{\max} = Z(\mathbf{w}_1^{\max}, \ldots, \mathbf{w}_m^{\max}) \approx Z^* + \sum_{i=1}^m \bar{G}_i + z_{1-\alpha/2}\sigma_I$. This corresponds to a point estimate (using plain empirical distributions as the inputs) that is approximated by $Z^* + \sum_{i=1}^m \bar{G}_i$, and adding a standard error $z_{1-\alpha/2}\sigma_I$ that captures its uncertainty. When R_2 is only moderate, $\hat{Z}^{\max}$ becomes approximately $Z^* + \sum_{i=1}^m \bar{G}_i + (\hat{Z}^{\max} - Z^{\max}) + z_{1-\alpha/2}\sigma_I$, where $Z^* + \sum_{i=1}^m \bar{G}_i + (\hat{Z}^{\max} - Z^{\max})$ can be regarded as the more noisy point estimate that contains the stochastic error $(\hat{Z}^{\max} - Z^{\max})$. In order to capture the joint input and simulation uncertainties, the standard error to this point estimate should be inflated to $z_{1-\alpha/2}\sqrt{\sigma_I^2 + \sigma^2/R_2}$ where σ^2 is the

variance of each simulation run, i.e., $Z^* + \sum_{i=1}^m \bar{G}_i + (\hat{Z}^{\max} - Z^{\max}) + z_{1-\alpha/2}\sqrt{\sigma_I^2 + \sigma^2/R_2}$ is an asymptotically exact upper bound. Comparing this with the approximate expression for $\hat{Z}^{\max}$, we thus conclude a margin adjustment of $z_{1-\alpha/2}\sqrt{\sigma_I^2 + \sigma^2/R_2} - z_{1-\alpha/2}\sigma_I$, which is estimated by the adjustment in Step 3 of Algorithm 6.

To make the above arguments precise, we have the following decompositions:

Proposition 3.4.13 *Let $Z^{\min} := Z(\mathbf{w}_1^{\min}, \ldots, \mathbf{w}_m^{\min})$ and $Z^{\max} := Z(\mathbf{w}_1^{\max}, \ldots, \mathbf{w}_m^{\max})$, and recall $\hat{Z}^{\min}$ and $\hat{Z}^{\max}$ in Step 3 of Algorithms 5 and 6. Under Assumptions 3.3.1, 3.3.2 and 3.3.3, as the input data size $n \to \infty$ and simulation effort $\frac{R_1}{n} \to \infty, R_2 \to \infty$, the outputs $\mathscr{L}^{EEL}, \mathscr{U}^{EEL}$ of Algorithm 5 satisfy*

$$\mathscr{L}^{EEL} = Z^* + \sum_{i=1}^m \bar{G}_i + (\hat{Z}^{\min} - Z^{\min}) - z_{1-\alpha/2}\Big(\sigma_I + \frac{\sigma}{\sqrt{R_2}}\Big) + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big)$$
$$\mathscr{U}^{EEL} = Z^* + \sum_{i=1}^m \bar{G}_i + (\hat{Z}^{\max} - Z^{\max}) + z_{1-\alpha/2}\Big(\sigma_I + \frac{\sigma}{\sqrt{R_2}}\Big) + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big)$$

whereas the outputs $\mathscr{L}^{FEL}, \mathscr{U}^{FEL}$ of Algorithm 6 satisfy

$$\mathscr{L}^{FEL} = Z^* + \sum_{i=1}^m \bar{G}_i + (\hat{Z}^{\min} - Z^{\min}) - z_{1-\alpha/2}\sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}} + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big)$$
$$\mathscr{U}^{FEL} = Z^* + \sum_{i=1}^m \bar{G}_i + (\hat{Z}^{\max} - Z^{\max}) + z_{1-\alpha/2}\sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}} + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big)$$

where $\sigma_I^2 = \sum_{i=1}^m \mathrm{Var}(G_i(X_i))/n_i$ is as defined in Corollary 3.4.7, $\sigma^2 = \mathrm{Var}_{P_1,\ldots,P_m}(h(\mathbf{X}_1, \ldots, \mathbf{X}_m))$ is the output variance, and the o_p is with respect to the input data and the simulation runs in Steps 1 and 3 of Algorithms 5-6.

To see how these decompositions arise, we can write the outputs of Algorithm 5 as (for the lower bound, say) $\hat{Z}^{\min} - z_{1-\alpha/2}\hat{\sigma}_{\min}/\sqrt{R_2} = Z^{\min} + (\hat{Z}^{\min} - Z^{\min}) - z_{1-\alpha/2}\hat{\sigma}_{\min}/\sqrt{R_2}$, where $Z^{\min}$, by Theorem 3.4.11, is close to $\mathscr{L}$ that is in turn representable as $Z^* + \sum_{i=1}^m \bar{G}_i - z_{1-\alpha/2}\sigma_I$ up to a small error by Theorem 3.4.8. Noting that $\hat{\sigma}_{\min}$ approximates σ, these together show the representation for $\mathscr{L}^{EEL}$ in Proposition 3.4.13. The other expressions for $\mathscr{U}^{EEL}$, and $\mathscr{L}^{FEL}$, $\mathscr{U}^{FEL}$, follow

analogously using the adjustments shown in Algorithms 5 and 6.

From Proposition 3.4.13, we can arrive at Theorems 3.3.3 and 3.3.4. For FEL, like discussed above, the two terms $\sum_{i=1}^{m} \bar{G}_i$ and $\hat{Z}^{\min} - Z^{\min}$ in the expression of $\mathscr{L}^{FEL}$ contain the input error and the simulation error in Step 3 respectively. They possess variances that are approximately σ_I^2 and σ^2/R_2, and their asymptotic independence implies a standard error of $z_{1-\alpha/2}\sqrt{\sigma_I^2 + \sigma^2/R_2}$, so that the representation of $\mathscr{L}^{FEL}$ matches the lower bound of an asymptotically exact CI. Similar conclusion holds for the upper bound $\mathscr{U}^{FEL}$. These almost put together the CI, except that $\mathscr{L}^{FEL}$ and $\mathscr{U}^{FEL}$ contain some inter-dependence that slightly corrupts the coverage, which leads to Theorem 3.3.4. The argument for EEL in Theorem 3.3.3 follows similarly, but with the standard error term in $\mathscr{L}^{EEL}$ or $\mathscr{U}^{EEL}$ overestimating the uncertainty by a factor as large as $\sqrt{2}$ (because $1 \le \frac{\sigma_I + \sigma/\sqrt{R_2}}{\sqrt{\sigma_I^2 + \sigma^2/R_2}} \le \sqrt{2}$, where $\sqrt{2}$ is attained when $\sigma_I^2 = \sigma^2/R_2$). In fact, under a coupling between all the simulation runs in Algorithms 5 and 6, $\hat{\sigma}_{\min}/\sqrt{R_2}$ always upper bounds $\sqrt{\hat{\sigma}_I^2 + \hat{\sigma}_{\min}^2/R_2} - \hat{\sigma}_I$ and hence Algorithm 5 always generates wider CIs than Algorithm 6.

3.5 Numerical Experiments

We present some numerical results for Algorithm 4 (BEL), Algorithm 5 (EEL) and Algorithm 6 (FEL). These include coverage probabilities and the statistical indicators, such as mean and standard deviation, of the positions or widths of the resulting CIs. We conduct experiments on two settings, a queueing model in Section 3.5.1 and stochastic activity networks in Section 3.5.2. We consider various levels of simulation budgets, data sizes, and problem dimensions (i.e., number of estimated input models). Throughout this section we set the target confidence level to 95%.

We also compare our procedures with three methods:

1. Percentile bootstrap resampling ("standard BT"): This scheme is suggested in Barton and Schruben (1993, 2001). Given m input data sets $\{X_{1,1}, \ldots, X_{1,n_1}\}, \ldots, \{X_{m,1}, \ldots, X_{m,n_m}\}$, it proceeds as follows. First choose B, the number of bootstrap resamples of the input empirical distributions, and R_b, the number of simulation replications for each bootstrap resample. For each $l = 1, 2, \ldots, B$, draw a simple random sample of size n_i with replacement, denoted by

$\{X^l_{i,1}, \ldots, X^l_{i,n_i}\}$, for each input model i, then generate R_b simulation replications driven by the empirical distributions formed by $\{X^l_{i,1}, \ldots, X^l_{i,n_i}\}, i = 1, \ldots, m$, and take their average to obtain Z^l. Finally output the $0.025(B+1)$-th and $0.975(B+1)$-th order statistics of $\{Z^l\}_{l=1}^B$.

2. Adaptive percentile bootstrap ("adaptive BT"): Proposed by Yi and Xie (2017), this approach adaptively allocates simulation budget in order to obtain percentile bootstrap CIs more efficiently than the standard percentile bootstrap. It aims to allocate more simulation runs to the resamples whose corresponding performance measures are closer to the 0.025 or 0.975 quantiles. The procedure consists of two phases. The first phase uses simulation to sequentially screen out bootstrap resamples that will less likely give the target quantiles. The second phase allocates the remaining simulation budget to the surviving resamples to more accurately estimate their performance measures. For a given simulation budget, the tuning parameters B, n_0, r, M (see Yi and Xie (2017)) are needed. In our subsequent comparisons we offer it some advantages by randomly drawing 10 different combinations of these parameters from a broad enough range of values, and reporting results on the top combinations ranked by the closeness of the coverage level to the nominal level.

3. The nonparametric delta method: This method has not been explicitly suggested in the simulation literature (in the nonparametric regime), and here we provide a heuristic version inspired from our analyses. The CI takes the form

$$\hat{Z} \pm z_{1-\frac{\alpha}{2}} \sqrt{\text{input-induced variance} + \text{stochastic variance}}$$

where $\hat{Z}$ is an estimate of the performance measure under the empirical input distributions. We estimate the stochastic variance using the sample variance of the generated simulation replications, and estimate the input-induced variance using the $\hat{\sigma}_I$ in Algorithm 6. To be specific, we carry out Step 1 of Algorithm 4 with $R_1 = R_d$, and then construct the CI

$$\hat{Z} \pm z_{1-\frac{\alpha}{2}} \sqrt{\sum_{i=1}^{m} \frac{1}{n_i} \Big(\frac{1}{n_i} \sum_{j=1}^{n_i} \big(\hat{\bar{G}}_i(X_{i,j})\big)^2 - \frac{n_i T_i \hat{\sigma}^2}{R_d} \Big) + \frac{\hat{\sigma}^2}{R_d}} \tag{3.25}$$

where $\hat{Z}$ and $\hat{\sigma}^2$ are respectively the sample mean and variance of the R_d simulation replications.

We will detail our comparisons under various problem and algorithmic configurations in the two experimental setups that follow. After that, in Section 3.5.3, we summarize some highlights and provide further comparisons with the bootstrap.

3.5.1 Mean Waiting Time of an M/M/1 Queue

We first consider a canonical M/M/1 queue with arrival rate 0.95 and service rate 1. The system is empty when the first customer comes in. We set our target performance measure as the expected waiting time of the 10-th customer. To put it in the form of (3.1), let A_t be the inter-arrival time between the t-th and $(t+1)$-th customers, S_t be the service time of the t-th customer, and

$$h(A_1, A_2, \ldots, A_9, S_1, S_2, \ldots, S_9) = W_{10},$$

where the waiting time W_{10} is calculated via the Lindley recursion

$$W_1 = 0, W_{t+1} = \max\{W_t + S_t - A_t, 0\}, \text{ for } t = 1, \ldots, 9.$$

Both the inter-arrival time distribution and the service time distribution are assumed unknown. Table 3.1 shows the results of all the methods under a simulation budget 2000 and input data sizes $n_1 = 30, n_2 = 25$. Table 3.2 summarizes results under a budget 8000 and data sizes $n_1 = 120, n_2 = 100$. For each row of the tables, 1000 i.i.d. input data sets are drawn from the true input distributions, and then a CI is constructed from each of them, from which the coverage probability, mean CI length and standard deviation of CI length are estimated. The word "overshoot" means that the CI limits exceed the natural bounds of the performance measure, i.e., the lower bound being negative given that waiting time must be non-negative.

We test the coverage probabilities of the optimization-based CIs. For each of Tables 3.1 and 3.2, we compute a "benchmark" coverage of each method by generating 5000 CIs each of which

consumes 5×10^4 simulation runs, to approximate the simulation-error-free coverage for comparison (the bracketed number underneath the name of each method in the tables). We observe first that the benchmark coverage of our optimization-based CIs are close to the nominal value 95% in both tables (roughly 92% in Table 3.1 and 94% in Table 3.2), which provides a sanity check for the validity of the EL method in our setting. Moreover, consistent with the asymptotic results, the benchmark coverage is closer to 95% when the data size is bigger (Table 3.2). Second, under the simulation budget of the experiments, Tables 3.1 and 3.2 show that in general BEL under-covers compared to the benchmark, EEL over-covers, whereas FEL is accurate. For instance, in Table 3.2 where the benchmark coverage of the EL method is 93.7%, BEL varies from 90% to 92%, EEL ranges from 96% to 99%, whereas FEL stays around 94%. This phenomenon is in line with Theorems 3.3.2, 3.3.3 and 3.3.4 since, as we have discussed in Sections 3.3.3 and 3.4.6, BEL does not take into account the stochastic uncertainty in the final evaluation, EEL captures the stochastic uncertainty but in a conservative manner, while FEL is designed to tightly match the magnitude of the uncertainty. The under-coverage issue of BEL and the over-coverage issue of EEL, especially for the larger-data case (Table 3.2), become more severe when R_2 is chosen small, while FEL delivers accurate coverage for all considered parameter values. Thus FEL seems to be more reliable over the other two procedures when the user has a limited simulation budget.

We compare our methods with the percentile bootstrap procedures in terms of coverage accuracy and algorithmic configuration. The benchmark coverages of our methods and the bootstrap appear to be quite similar in all considered cases (within 1% in both Tables 3.1 and 3.2). Moreover, the bootstrap methods perform competitively in terms of the actual coverages, when the budget allocation or tuning parameters are optimally chosen. Nonetheless, FEL appears to show more robust performance with respect to these tuning needs. In the standard bootstrap, when R_b is chosen large relative to the data size and B is set around 50, the coverages of the CIs are close to the benchmark coverages in all cases. However, as R_b decreases, the coverage probabilities of bootstrap CIs quickly rise towards 100%. This over-coverage issue can be attributed to the higher variability caused by small R_b that is not properly accounted for, as discussed in Barton et al. (2002) and Barton et al. (2007). The adaptive bootstrap appears to mitigate this issue

	methods & parameters	coverage estimate	mean CI length	std. CI length	% of overshoot
BEL (91.8%*)	$R_1 = 1000, R_2 = 500$	89.6%	4.76	2.17	0%
	$R_1 = 1500, R_2 = 250$	90.7%	4.72	1.99	0%
	$R_1 = 1800, R_2 = 100$	88.7%	4.76	2.15	0%
	$R_1 = 1900, R_2 = 50$	89.2%	4.79	2.24	0%
EEL (91.8%*)	$R_1 = 1000, R_2 = 500$	93.1%	5.21	2.19	0%
	$R_1 = 1500, R_2 = 250$	94.1%	5.38	2.21	0%
	$R_1 = 1800, R_2 = 100$	95.1%	5.67	2.42	0%
	$R_1 = 1900, R_2 = 50$	96.0%	6.16	2.64	0.1%
FEL (91.8%*)	$R_1 = 1000, R_2 = 500$	90.5%	4.72	2.06	0%
	$R_1 = 1500, R_2 = 250$	91.9%	4.83	2.07	0%
	$R_1 = 1800, R_2 = 100$	91.9%	4.93	2.08	0%
	$R_1 = 1900, R_2 = 50$	91.5%	5.06	2.20	0%
standard BT (91.0%*)	$B = 50, R_b = 40$	91.2%	4.90	2.23	0%
	$B = 100, R_b = 20$	93.5%	4.98	2.02	0%
	$B = 400, R_b = 5$	96.9%	6.09	2.28	0%
	$B = 1000, R_b = 2$	99.2%	7.74	2.82	0%
adaptive BT (4 best combinations) (91.0%*)	$B = 100, n_0 = 10, r = 1.2, M = 3$	92.7%	5.01	2.18	0%
	$B = 100, n_0 = 10, r = 1.2, M = 1$	92.0%	5.02	2.22	0%
	$B = 100, n_0 = 10, r = 1.4, M = 1$	92.3%	4.93	2.08	0%
	$B = 100, n_0 = 10, r = 1.8, M = 1$	92.5%	5.00	2.24	0%
nonparametric delta method (86.6%*)	$R_d = 2000$	84.9%	4.66	2.08	54%

$*$ denotes the benchmark coverage with negligible simulation noise.

Table 3.1: $M/M/1$ queue. $n_1 = 30, n_2 = 25$. Total simulation budget 2000. Run times (second/CI): three EL methods 1.1×10^{-2}, the bootstrap 1.2×10^{-2}, delta method 1.0×10^{-2}.

by more efficient allocation of the budget. It requires, however, a careful selection of the best parameter configurations (while the tables show the top four configurations, the worst case among our randomly selected 10 choices has a coverage of 80%). In practice these parameters needs to be obtained via discrete simulation optimization (Yi and Xie 2017). In contrast, the coverage probabilities of FEL stay almost unchanged under various budget allocations (including the case that R_2 is as small as 50). FEL thus seems easy to use in terms of algorithmic configuration; in particular, merely setting $R_2 = 50$ appears doing well.

To further illustrate the robustness of the proposed approach in terms of algorithmic configurations, relative to the bootstrap, we show in Table 3.3 the coverages as we increase the simulation

	methods & parameters	coverage estimate	mean CI length	std. CI length	% of overshoot
BEL (93.7%*)	$R_1 = 4000, R_2 = 2000$	92.6%	2.47	0.597	0%
	$R_1 = 7000, R_2 = 500$	92.4%	2.46	0.606	0%
	$R_1 = 7800, R_2 = 100$	91.9%	2.48	0.713	0%
	$R_1 = 7900, R_2 = 50$	89.6%	2.45	0.787	0%
EEL (93.7%*)	$R_1 = 4000, R_2 = 2000$	95.7%	2.66	0.626	0%
	$R_1 = 7000, R_2 = 500$	97.7%	2.90	0.678	0%
	$R_1 = 7800, R_2 = 100$	98.0%	3.50	0.870	0%
	$R_1 = 7900, R_2 = 50$	98.8%	3.94	1.04	0%
FEL (93.7%*)	$R_1 = 4000, R_2 = 2000$	93.6%	2.45	0.591	0%
	$R_1 = 7000, R_2 = 500$	94.3%	2.45	0.594	0%
	$R_1 = 7800, R_2 = 100$	94.1%	2.74	0.705	0%
	$R_1 = 7900, R_2 = 50$	94.3%	2.90	0.865	0%
standard BT (94.2%*)	$B = 50, R_b = 160$	92.7%	2.56	0.675	0%
	$B = 100, R_b = 80$	96.4%	2.64	0.613	0%
	$B = 400, R_b = 20$	98.8%	3.19	0.658	0%
	$B = 1000, R_b = 8$	100%	4.19	0.800	0%
adaptive BT (4 best combinations) (94.2%*)	$B = 200, n_0 = 20, r = 1.6, M = 1$	93.6%	2.64	0.657	0%
	$B = 200, n_0 = 15, r = 2, M = 1$	95.0%	2.68	0.687	0%
	$B = 200, n_0 = 5, r = 1.6, M = 3$	94.5%	2.71	0.688	0%
	$B = 400, n_0 = 10, r = 1.8, M = 1$	94.5%	2.72	0.654	0%
nonparametric delta method (91.5%*)	$R_d = 8000$	92.0%	2.45	0.560	0%

$*$ denotes the benchmark coverage with negligible simulation noise.

Table 3.2: $M/M/1$ queue. $n_1 = 120, n_2 = 100$. Total simulation budget 8000. Run times (second/CI): three EL methods 4.0×10^{-2}, the bootstrap 3.4×10^{-2}, delta method 5.3×10^{-2}.

budget. The first row shows the coverage estimates of the bootstrap and FEL under allocations that satisfy the same overall simulation budget. Both appear to be close to their respective benchmark coverages shown in Table 3.1. However, the coverages of the bootstrap could be illusory in this case since, as the bootstrap size B increases with R_b fixed, the coverage rises from 91% to 95% as shown in the following rows. These deviate from the benchmark coverages, and indicate that neither B nor R_b is large enough for the bootstrap to work properly. In contrast, the coverage of FEL appears quite stable and remains close to the benchmark when R_1 or R_2 increases.

Compared to the nonparametric delta method, our optimization-based CIs possess better coverages, especially in the situation of limited input data size. When the data size is less than 30 for

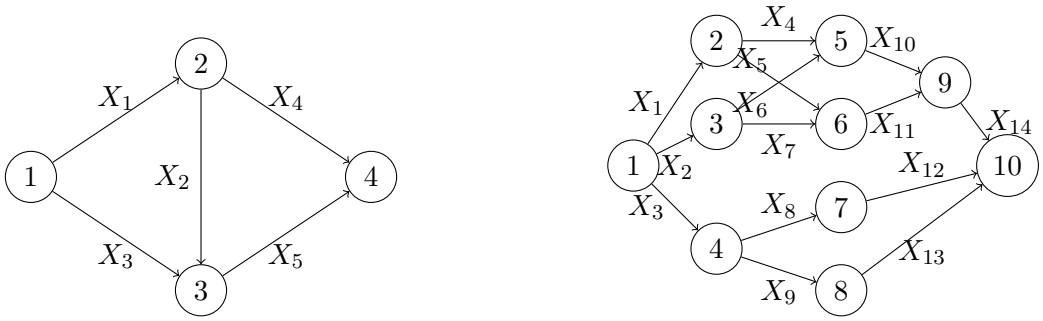
standard BT		FEL	
parameters	coverage estimate	parameters	coverage estimate
$B = 40, R_b = 15$	90.9%	$R_1 = 500, R_2 = 50$	90.3%
$B = 100, R_b = 15$	92.4%	$R_1 = 2000, R_2 = 50$	91.9%
$B = 200, R_b = 15$	93.6%	$R_1 = 500, R_2 = 200$	90.2%
$B = 500, R_b = 15$	94.7%	$R_1 = 2000, R_2 = 200$	90.8%

Table 3.3: $M/M/1$ queue. $n_1 = 30, n_2 = 25$.

each input model (Table 3.1), the coverage probabilities of the delta-method CIs are around 85%, while our methods are around 90% to 96%, depending on the particular variants. The unsatisfactory coverage of the delta-method CI could be attributed to the shrinkage of the interval caused by two factors. The first is the overshoot issue. Table 3.1 shows that frequently the delta-method CI exceeds the natural bounds of the target performance measure, which renders its effective length shorter and hence an inferior coverage. The second factor that may shorten the interval is the noise in estimating the input-induced variance. When only 30 input data are available, the estimate of the input-induced variance used in constructing the CI (3.25) can be highly noisy, which potentially makes the overall standard error estimate downward biased (hence shorter CIs) because of the concavity of the square root operation. This may explain the smaller length of the delta-method CI compared to CIs by other methods in Table 3.1. The coverage gets much better for the delta-method CI when input data size rises above 100 (Table 3.2), which gets close to, but still falls short of, our optimization-based counterparts especially FEL.

3.5.2 Stochastic Activity Networks

We consider a larger-scale problem and larger ranges of data sizes, in the setting of stochastic activity networks shown in Figure 3.1. The first network Figure 3.1a is borrowed from Yi and Xie (2017). Each edge $i = 1, \ldots, 5$ of the network represents a task that can be completed in X_i units of time. Assigning each X_i to edge i as its length, the total time to finish the project is the length of the longest path from node 1 to node 4, i.e. $h(X_1, \ldots, X_5) = \max\{X_1+X_2+X_5, X_1+X_4, X_3+X_5\}$. Assume that the unknown distributions of the X_i's are exponential with rate $10, 5, 12, 11, 5$ for i from



(a) 4 nodes and $m = 5$ tasks.

(b) 10 nodes and $m = 14$ tasks.

Figure 3.1: Stochastic activity networks.

1 to 5, and we are interested in computing the expected time to finish the project $\mathbb{E}[h(X_1, \ldots, X_5)]$.

	methods & parameters	coverage estimate	mean CI length	std. CI length	% of overshoot
BEL	$R_1 = 4000, R_2 = 2000$	92.7%	0.17	0.03	0%
	$R_1 = 7000, R_2 = 500$	91.9%	0.17	0.04	0%
	$R_1 = 7800, R_2 = 100$	84.9%	0.18	0.06	0%
	$R_1 = 7900, R_2 = 50$	81.7%	0.18	0.07	0%
EEL	$R_1 = 4000, R_2 = 2000$	96.1%	0.20	0.03	0%
	$R_1 = 7000, R_2 = 500$	97.7%	0.23	0.04	0%
	$R_1 = 7800, R_2 = 100$	99.0%	0.30	0.07	0%
	$R_1 = 7900, R_2 = 50$	99.4%	0.35	0.09	0%
FEL	$R_1 = 4000, R_2 = 2000$	92.2%	0.17	0.03	0%
	$R_1 = 7000, R_2 = 500$	93.2%	0.18	0.04	0%
	$R_1 = 7800, R_2 = 100$	94.6%	0.22	0.06	0%
	$R_1 = 7900, R_2 = 50$	94.5%	0.25	0.08	0%
standard BT	$B = 50, R_b = 160$	94.0%	0.21	0.04	0%
	$B = 100, R_b = 80$	97.1%	0.22	0.04	0%
	$B = 400, R_b = 20$	99.7%	0.33	0.04	0%
	$B = 1000, R_b = 8$	100%	0.47	0.05	0%
adaptive BT (4 best combinations)	$B = 300, n_0 = 15, r = 1.2, M = 1$	94.9%	0.22	0.04	0%
	$B = 100, n_0 = 20, r = 1.2, M = 1$	93.9%	0.22	0.05	0%
	$B = 400, n_0 = 10, r = 1.2, M = 3$	95.6%	0.24	0.04	0%
	$B = 100, n_0 = 5, r = 1.2, M = 3$	96.2%	0.22	0.04	0%
nonparametric delta method	$R_d = 8000$	94.9%	0.18	0.03	0%

Table 3.4: Stochastic activity network in Figure 3.1a. $n_1 = n_2 = 200, n_3 = n_4 = n_5 = 30$. Total simulation budget 8000. Run times (second/CI): three EL methods 3.3×10^{-2}, the bootstrap 1.7×10^{-2}, delta method 3.2×10^{-2}.

We test our method in cases where the data sizes for different input models vary significantly.

Specifically we consider the case where $n_1 = n_2 = 200$ and $n_3 = n_4 = n_5 = 30$, which produce a ratio of roughly 7 between the maximum and minimum data sizes. Table 3.4 shows the results under a simulation budget of 8000. All the methods seem to exhibit performances similar to the cases with more balanced observations in Tables 3.1 and 3.2. For example, FEL and the adaptive bootstrap generate CIs with similar coverage probabilities (around the nominal level 95%), EEL and the standard bootstrap tend to over-cover, and BEL tends to under-cover especially for small values of R_2. In contrast to the last example, the nonparametric delta method in this case seems to have a good performance that is similar to our FEL. This could be because the performance function h here is piecewise linear with only three pieces, hence can be well approximated by a single linear function and in turn leads to the better finite-sample performance of the delta method that relies crucially on linearization.

Next we consider a bigger stochastic activity network, shown in Figure 3.1b, that is borrowed from Chu et al. (2014) that consists of 14 tasks. The time to completion X_i of each task follows exponential distribution with rate $10, 5, 12, 11, 5, 8, 4, 9, 13, 7, 6, 9, 10, 6$ for i from 1 to 14. In addition to computing the expected time to complete the project (Table 3.5), which is represented by the length of the longest path from node 1 to 10, we also test our methods in estimating the tail probability that the time to finish the project exceeds 1.5 units of time (Tables 3.6 and 3.7). The true value of the probability is 0.0747 (estimated from abundunt simulation).

Table 3.5 shows that our FEL and the adaptive bootstrap consistently exhibit satisfactory coverage levels when the number of input models is fairly big compared with the input data size (per input model). Here we use a simulation budget of 4000, and a data size of 30 for the first 7 input models, and 25 for the other 7 inputs. The coverage probabilities and their trends in each method are similar to our observations before (e.g., in Tables 3.2 and 3.4). For example, the coverage of FEL stays around 94%, the standard bootstrap over-covers for small R_b, and BEL under-covers for small R_2.

Table 3.6 shows the tail probability estimation results, with a data size around 100 per input model. Table 3.7 considers a bigger data size of 400-500. The simulation budgets are 16000 and 60000 respectively. FEL and the delta method seem to have accurate coverage probabilities (93%

	methods & parameters	coverage estimate	mean CI length	std. CI length	% of overshoot
BEL	$R_1 = 3000, R_2 = 500$	91.6%	0.24	0.04	0%
	$R_1 = 3500, R_2 = 250$	90.4%	0.24	0.05	0%
	$R_1 = 3800, R_2 = 100$	89.1%	0.24	0.06	0%
	$R_1 = 3900, R_2 = 50$	85.0%	0.24	0.09	0%
EEL	$R_1 = 3000, R_2 = 500$	97.3%	0.31	0.05	0%
	$R_1 = 3500, R_2 = 250$	96.9%	0.33	0.06	0%
	$R_1 = 3800, R_2 = 100$	98.3%	0.39	0.08	0%
	$R_1 = 3900, R_2 = 50$	98.9%	0.45	0.11	0%
FEL	$R_1 = 3000, R_2 = 500$	93.3%	0.25	0.04	0%
	$R_1 = 3500, R_2 = 250$	93.2%	0.26	0.05	0%
	$R_1 = 3800, R_2 = 100$	93.3%	0.29	0.07	0%
	$R_1 = 3900, R_2 = 50$	94.9%	0.32	0.09	0%
standard BT	$B = 50, R_b = 80$	94.9%	0.31	0.06	0%
	$B = 100, R_b = 40$	98.4%	0.33	0.06	0%
	$B = 400, R_b = 10$	99.9%	0.50	0.08	0%
	$B = 1000, R_b = 4$	100%	0.73	0.10	0%
adaptive BT (4 best combinations)	$B = 100, n_0 = 15, r = 1.8, M = 1$	95.0%	0.30	0.06	0%
	$B = 100, n_0 = 5, r = 1.2, M = 7$	95.3%	0.31	0.06	0%
	$B = 100, n_0 = 10, r = 1.8, M = 1$	94.1%	0.31	0.06	0%
	$B = 100, n_0 = 20, r = 1.2, M = 1$	93.7%	0.30	0.06	0%
nonparametric delta method	$R_d = 2000$	93.8%	0.26	0.04	0%

Table 3.5: Stochastic activity network in Figure 3.1b. $n_i = 30$ for $1 \leq i \leq 7$ and 25 for $8 \leq i \leq 14$. Total simulation budget 4000. Run times (second/CI): three EL methods 2.7×10^{-2}, the bootstrap 2.7×10^{-2}, delta method 1.7×10^{-2}.

in Table 3.6 and 94% in Table 3.7). EEL continues to over-cover. Notably, BEL suffers from severe under-coverage issues, while the standard bootstrap suffers from severe over-coverage issues. Though FEL gives accurate CIs in most cases, the simple budget allocation strategy of setting $R_2 = 50$ and investing the remainder to R_1 appears to perform less well than using a larger R_2 such as $100, 250$. This could be because of the highly skewed performance function, which requires a larger R_2 to invoke the central limit behavior needed in the CI construction. Our suggestion is to use R_2 in the range of hundreds in FEL for tail estimation problems. The adaptive bootstrap seems to significantly under-cover the truth in this case, which could be because the relatively small input uncertainty and the high skewness of the simulation output make it more difficult to differentiate the bootstrap resamples in the screening stage.

	methods & parameters	coverage estimate	mean CI length	std. CI length	% of overshoot
BEL	$R_1 = 15000, R_2 = 500$	86.0%	0.064	0.020	0%
	$R_1 = 15500, R_2 = 250$	80.0%	0.064	0.026	0%
	$R_1 = 15800, R_2 = 100$	70.3%	0.064	0.040	0%
	$R_1 = 15900, R_2 = 50$	57.8%	0.062	0.055	0%
EEL	$R_1 = 15000, R_2 = 500$	98.5%	0.110	0.023	0%
	$R_1 = 15500, R_2 = 250$	98.8%	0.130	0.031	1.2%
	$R_1 = 15800, R_2 = 100$	98.7%	0.166	0.046	30%
	$R_1 = 15900, R_2 = 50$	97.5%	0.205	0.067	65%
FEL	$R_1 = 15000, R_2 = 500$	93.2%	0.079	0.020	0%
	$R_1 = 15500, R_2 = 250$	93.0%	0.090	0.027	0%
	$R_1 = 15800, R_2 = 100$	93.2%	0.120	0.044	0%
	$R_1 = 15900, R_2 = 50$	91.4%	0.155	0.062	3.8%
standard BT	$B = 50, R_b = 320$	97.1%	0.090	0.018	0%
	$B = 100, R_b = 160$	99.2%	0.104	0.017	0%
	$B = 400, R_b = 40$	100%	0.170	0.026	0%
	$B = 1000, R_b = 16$	100%	0.230	0.038	0%
adaptive BT (3 best combinations)	$B = 100, n_0 = 80, r = 1.1, M = 5$	89.0%	0.093	0.026	0%
	$B = 100, n_0 = 100, r = 1.1, M = 4$	92.3%	0.089	0.023	0%
	$B = 100, n_0 = 100, r = 1.2, M = 2$	91.4%	0.091	0.024	0%
nonparametric delta method	$R_d = 16000$	93.2%	0.070	0.011	0%

Table 3.6: Tail probability of stochastic activity network in Figure 3.1b. $n_i = 120$ for $1 \leq i \leq 7$ and 100 for $8 \leq i \leq 14$. Total simulation budget 16000. Run times (second/CI): three EL methods 0.11, the bootstrap 0.03, delta method 0.10.

3.5.3 Summary and Comparisons with the Bootstrap

Based on the findings in Sections 3.5.1 and 3.5.2, we provide some general comparisons between our optimization-based approach and the standard bootstrap in terms of the required simulation burden, the ease of implementation and the computation cost.

Because of the nested simulation, the total simulation load of the standard bootstrap is BR_b. To ensure the stochastic noise is negligible relative to input uncertainty, one would need $R_b \gg n$ (where "$\gg$" means "of larger order than"). On the other hand, Theorems 3.3.3 and 3.3.4 suggest that, in the optimization-based approach, one can choose $R_1 \gg n, R_2 \gg 1$. Thus, the bootstrap requires $BR_b \gg Bn$ total simulation load, whereas ours requires $R_1 + 2R_2 \gg n$ simulation load. Since B is typically a big number (in the experiments we use B between 50 and 1000), our method

	methods & parameters	coverage estimate	mean CI length	std. CI length	% of overshoot
BEL	$R_1 = 59000, R_2 = 500$	73.3%	0.032	0.017	0%
	$R_1 = 59500, R_2 = 250$	63.1%	0.033	0.024	0%
	$R_1 = 59800, R_2 = 100$	50.6%	0.032	0.038	0%
	$R_1 = 59900, R_2 = 50$	43.0%	0.032	0.054	0%
EEL	$R_1 = 59000, R_2 = 500$	99.1%	0.078	0.018	0%
	$R_1 = 59500, R_2 = 250$	98.6%	0.097	0.025	0%
	$R_1 = 59800, R_2 = 100$	97.9%	0.132	0.040	15%
	$R_1 = 59900, R_2 = 50$	94.9%	0.172	0.061	58%
FEL	$R_1 = 59000, R_2 = 500$	93.4%	0.055	0.017	0%
	$R_1 = 59500, R_2 = 250$	94.1%	0.071	0.025	0%
	$R_1 = 59800, R_2 = 100$	94.0%	0.104	0.041	0%
	$R_1 = 59900, R_2 = 50$	93.2%	0.141	0.061	28%
standard BT	$B = 50, R_b = 1200$	97.6%	0.047	0.007	0%
	$B = 100, R_b = 600$	99.3%	0.054	0.006	0%
	$B = 400, R_b = 150$	100%	0.090	0.007	0%
	$B = 1000, R_b = 60$	100%	0.134	0.012	0%
adaptive BT (3 best combinations)	$B = 100, n_0 = 200, r = 1.2, M = 3$	84.0%	0.048	0.011	0%
	$B = 200, n_0 = 150, r = 1.2, M = 3$	82.6%	0.050	0.013	0%
	$B = 100, n_0 = 200, r = 1.2, M = 5$	84.2%	0.048	0.010	0%
nonparametric delta method	$R_d = 60000$	94.3%	0.035	0.003	0%

Table 3.7: Tail probability of stochastic activity network in Figure 3.1b. $n_i = 480$ for $1 \leq i \leq 7$ and 400 for $8 \leq i \leq 14$. Total simulation budget 60000. Run times (second/CI): three EL methods 1.4, the bootstrap 0.08, delta method 1.3.

seems to be more efficient in terms of simulation cost. In Tables 3.1 and 3.2, we have observed that under the same total simulation budget FEL consistently possess coverage probabilities close to the benchmark coverage while the bootstrap very often significantly exceeds the benchmark level.

We also notice that our optimization-based approach is more robust with respect to the algorithmic parameter configuration. Given a fixed total simulation budget, it could be challenging to figure out a good choice of B and R_b for the bootstrap, as it can highly depend on the input data sizes and the magnitude of the simulation error. Indeed, our experiments indicate that the coverage of the bootstrap CIs is quite sensitive to the allocations of B and R_b. When B and R_b are not appropriately chosen, the bootstrap CI tends to over-cover the truth. On the other hand, in the optimization-based method, particularly FEL, setting R_2 to be a fixed moderately large number (say 50) and investing the remaining budget to R_1 seems to be quite stable regardless of the data

size. Nonetheless, we have seen that if the performance measure is a small probability, choosing a larger R_2 would improve the coverages.

Despite the simulation savings and stability, the optimization-based approach calls for a heavier computation overhead than the bootstrap beyond the simulation effort. In the bootstrap, the extra numerical computation other than simulation runs is negligible. In our approach, we need to estimate gradient information (the influence function) in (3.2) in Step 1, and solve the optimization pair in Step 2. Computation of the score function $S_{i,j}(\mathbf{X}_i^r)$ for all i, j and $r = 1, \ldots, R_1$ requires $O((\sum_{i=1}^m T_i)R_1)$ time, by counting the occurrence of each $X_{i,j}$ in the generated input variates. The sample covariance between the output h and the score function is computed in $O((\sum_{i=1}^m n_i)R_1)$ time. Thus the total computation in Step 1 has a complexity $O((\sum_{i=1}^m n_i + \sum_{i=1}^m T_i)R_1)$. Using the approach suggested by Proposition 3.3.1, the optimization pair (3.3) can be solved in $O(c^{bi}(\sum_{i=1}^m n_i c_i^{nt}))$ time, where c^{bi} is the number of bisection iterations on β and c_i^{nt} is the number of Newton iterations to obtain each $\lambda_i(\beta)$. The global linear convergence of bisection and Newton's method in our setting suggest that, to achieve a given tolerance level, typically c^{bi} and each c_i^{nt} only need to be logarithmically large. Ignoring logarithmic factors, we see that the computation cost of Step 2 is roughly $O(\sum_{i=1}^m n_i)$. Thus the cost of Step 1 dominates Step 2, leading to a total overhead cost $O((\sum_{i=1}^m n_i + \sum_{i=1}^m T_i)R_1)$. In the case of large data size, these overhead costs of our method can be substantial, which is reflected by the significantly longer run times of the EL methods compared to the bootstrap in Tables 3.6 and 3.7.

3.6 Conclusion

We have proposed an optimization-based approach to construct CIs for simulation output performance measures that account for the input uncertainty from finite data. This approach relies on solving a pair of optimization programs posited over distributions supported on the data, with a constraint expressed in terms of the weighted average of empirically defined Burg-entropy divergences. It then uses the solutions to define probability weights that subsequently drive simulation runs. We present several related procedures under this approach and analyze their statistical per-

formances using a generalization of the EL method. Compared to the bootstrap, our approach is less sensitive to the allocation choices and achieves more stable coverages, especially under small data sizes, as explained both theoretically and shown by our numerical experiments. The numerical results also reveal that our approach tends to curb the under-coverage issues encountered in the delta method. The last of our procedures, FEL, seems particularly attractive compared to both the bootstrap and the delta method in terms of finite-data finite-simulation performance.

Chapter 4

Bounding Optimality Gap in Stochastic Optimization via Bagging

4.1 Introduction

Consider a stochastic optimization problem

$$Z^* = \min_{x \in \mathcal{X}} \{Z(x) = E_F[h(x, \xi)]\} \tag{4.1}$$

where $\xi \in \Xi$ is generated under some distribution F, and $E_F[\cdot]$ denotes its expectation. We focus on the situations where F is not known, but instead a collection of i.i.d. data for ξ, say $\boldsymbol{\xi}_{1:n} = (\xi_1, \ldots, \xi_n)$, are available. Obtaining a good solution for (4.1) under this setting has been under active investigation both from the stochastic and the optimization communities. Common methods include the sample average approximation (SAA) (Shapiro et al. (2014), Kleywegt et al. (2002)), stochastic approximation (SA) or gradient descent (Kushner and Yin (2003), Borkar (2009), Nemirovski et al. (2009)), and (distributionally) robust optimization (Delage and Ye (2010), Bertsimas et al. (2018b), Wiesemann et al. (2014), Ben-Tal et al. (2013)). These methods aim to find a solution that is nearly optimal, or in some way provide a safe approximation. Applications of the generic problem (4.1) and its data-driven solution techniques span from operations research, such

as inventory control, revenue management, portfolio selection (see, e.g., Shapiro et al. (2014), Birge and Louveaux (2011)) to risk minimization in machine learning (e.g., Friedman et al. (2001)).

This chapter concerns the estimation of Z^* using limited data. Moreover, given a solution, say $\hat{x}$, a closely related problem is to estimate the optimality gap

$$\mathcal{G}(\hat{x}) = Z(\hat{x}) - Z^* \tag{4.2}$$

This allows us to assess the quality of $\hat{x}$, in the sense that the smaller $\mathcal{G}(\hat{x})$ is, the closer is the solution $\hat{x}$ to the true optimum in terms of achieved objective value. More precisely, we will focus on inferring a lower confidence bound for Z^*, and, correspondingly, an upper bound for $\mathcal{G}(\hat{x})$ - noting that its first term $Z(\hat{x})$ can be treated as a standard population mean of $h(\hat{x}, \xi)$ that is estimable using a sample independent of the given $\hat{x}$, or that $\mathcal{G}(\hat{x})$ can be represented as the max of the expectation of $h(\hat{x}, \xi) - h(x, \xi)$ whose estimation is structurally the same as Z^*.

This problem is motivated by the fact that many state-of-the-art solution methods mentioned before are only amenable to crude, worst-case performance bounds. For instance, Shapiro and Nemirovski (2005) and Kleywegt et al. (2002) provide large deviations bounds on the optimality gap of SAA in terms of the diameter or cardinality of the decision space and the maximum variance of the function h. Nemirovski et al. (2009) and Ghadimi and Lan (2013) provide bounds on the expected value and deviation probabilities of the SA iterates in terms of the strong convexity parameters, space diameter and maximum variance. These bounds can be refined under additional structural information (e.g., Shapiro and Homem-de-Mello (2000)). While they are very useful in understanding the behaviors of the optimization procedures, using them as a precise assessment on the quality of an obtained solution may be conservative. Because of this, a stream of work study approaches to validate solution performances by statistically bounding optimality gaps. Mak et al. (1999), Bayraksan and Morton (2006), Love and Bayraksan (2015) and Shapiro (2003) investigate the use of SAA to estimate these bounds. Lan et al. (2012) validate the performances of SA iterates by using convexity conditions. Stockbridge and Bayraksan (2013) and Partani et al. (2006) study approaches like the jackknife and probability metric minimization to reduce the bias in the resulting

gap estimates. Bayraksan and Morton (2011) utilize gap estimates to guide sequential sampling. Duchi et al. (2016), Blanchet et al. (2019) and Lam and Zhou (2017) investigate the use of empirical and profile likelihoods to estimate optimal values. Our investigation in this chapter follows the above line of work on solution validation, focusing on the situation when data are limited and hence the statistical efficiency becomes utmost important. We also point out a related series of work that validate feasibility under uncertain constraints (e.g., Luedtke and Ahmed (2008), Pagnoncelli et al. (2009), Wang and Ahmed (2008), Carè et al. (2014), Calafiore (2017)), though their problem of interest is beyond the scope of this chapter, as we focus on deterministically constrained problems and objective value performances.

More precisely, we introduce a bootstrap aggregating, or commonly known as bagging (Breiman (1996)), approach to estimate a lower confidence bound for Z^*. This comprises repeated resampling of data to construct SAAs, and ultimately averaging the resampled optimal SAA values. We demonstrate how this approach applies under very general conditions on the cost function h and decision space $\mathcal{X}$, while enjoys high statistical efficiency and stability. Compared to procedures based on batching (e.g., Mak et al. (1999)), which also have documented benefits in wide applicability and stability, the data recycling in our approach breaks free a tradeoff between the tightness of the resulting bound and the statistical accuracy/correctness exhibited by batching. In cases where sufficient smoothness is present and central limit theorem (CLT) for SAA (e.g., Shapiro et al. (2014), Bayraksan and Morton (2006)) can be directly applied, we also see that our approach gains stability regarding standard error estimation, thanks to the smoothing effect brought by bagging. Nonetheless, our approach generally requires higher computational load than these previous methods due to the need to solve many resampled programs. While we focus primarily on statistical performances, towards the end of this chapter we will discuss some computational implications.

The theoretical justification of our bagging scheme comes from viewing SAA as a kernel in an infinite-order symmetric statistic (Frees (1989)), and an established optimistic bound for SAA as its asymptotic limit. A symmetric statistic is a generalization of sample mean in which each summand consists of a function (i.e., kernel) acting on more than one observation (Serfling (2009), Lee (2019)). In particular, the size of the SAA program can be seen as precisely the kernel "order"

(or "degree"), which depends on the data size and is consequently of an infinite-order nature. Our bagging scheme serves as a Monte Carlo approximation for this symmetric statistic. As a main methodological contribution, we analyze the asymptotic behaviors of the statistic and the resulting bounds as the SAA size grows, and translate them into efficient performances of our bagging scheme. Finally, we note that the notion of infinite-order symmetric statistics has been used in analyzing ensemble machine learning predictors like random forests (Wager and Athey (2018)); our SAA kernels are, from this view, in parallel to the base learners in the latter context.

Finally, we mention that Eichhorn and Römisch (2007) has also studied the resampling of SAA programs to construct confidence intervals for the optimal values of stochastic programs. Our approach connects with, but also differs substantially from Eichhorn and Römisch (2007) in several regards. In terms of scope of applicability, Eichhorn and Römisch (2007) focuses on mixed-integer linear programs, while we consider cost functions that can be generally non-Donsker. However, we instead require an additional "non-degeneracy" condition that depends on the cost function and the underlying probability distribution. In terms of methodology, Eichhorn and Römisch (2007) utilizes the quantiles of the resampled distribution to generate confidence intervals, by observing the same limiting distribution between an original CLT and the bootstrap CLT. The resampling in Eichhorn and Römisch (2007) requires a "two-layer" extended bootstrap where each resample is drawn from a new sample of the true distribution (as opposed to some bootstrap methods that allows repeated resample from the same original sample, with the availability of a conditional bootstrap CLT). Thus the approach requires substantial data size or otherwise resorting to subsampling. Our bagging approach, in contrast, is based on a direct use of Gaussian limit and standard error estimation in the CLT for the optimistic bound. Our burden lies on the bootstrap size requirement to obtain consistent standard error estimate, and less on the data size requirement.

We summarize our contributions as follows:

1. Motivated from the challenges of existing techniques (Section 4.2), we introduce a bagging procedure to estimate a lower confidence bound for Z^*, correspondingly an upper confidence bound for $\mathcal{G}(\hat{x})$ (Section 4.3). We present the idea of our procedure that views SAA as a kernel in a symmetric statistic, and an optimistic bound for SAA as its associated limiting

quantity (Section 4.4).

2. We analyze the asymptotic behaviors of the infinite-order symmetric statistic generated from the SAA kernel, under minimal smoothness requirements on the optimization problem. Moreover, when smoothness conditions are introduced, we demonstrate how these behaviors recover the classical CLT on SAA. These results are presented in Section 4.5. The mathematical developments without smoothness conditions utilize a combination of probabilistic coupling arguments and a new hypergeometric representation associated with the Hajek projection (Van der Vaart (2000)) (Appendices C.1 and C.2). The developments to recover the classical CLT use another analysis-of-variance (ANOVA) decomposition and a maximum deviation bound for empirical processes (Appendix C.3).

3. Building on the above results, we demonstrate how the bounds generated from our bagging procedure exhibit asymptotically correct coverages, and improve a tradeoff between the bound tightness and the statistical accuracy in existing batching schemes. This efficiency gain can be seen by an asymptotic comparison of the standard error in our estimator and an interpretation using conditional Monte Carlo. These developments are in Sections 4.6 and 4.7, with mathematical details in Appendices C.4-C.8.

4. We explain the stability in our generated bounds brought by the smoothing effect of bagging in estimating standard error. This compares favorably with the direct use of CLT in situations where the objective function is smooth. This property is supported by our numerical experiments (Section 4.8).

4.2 Existing Challenges and Motivation

We discuss some existing methods and their challenges, to motivate our investigation. We start the discussion with the direct use of asymptotics from sample average approximation (SAA).

4.2.1 Using Asymptotics of Sample Average Approximation

When the cost function h in (4.1) is smooth enough, it is known classically that a central limit theorem (CLT) governs the behavior of the estimated optimal value in SAA, namely

$$\hat{Z}_n = \min_{x \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} h(x, \xi_i). \tag{4.3}$$

We first introduce the following Lipschitz condition:

Assumption 4.2.1 (Lipschitz continuity) *The cost function $h(x, \xi)$ is Lipschitz continuous in the decision x, in the sense that*

$$|h(x_1, \xi) - h(x_2, \xi)| \leq M(\xi)\|x_1 - x_2\|$$

for any $x_1, x_2 \in \mathcal{X}$, where $M(\xi)$ satisfies $E[M^2(\xi)] < \infty$.

Denote "$\Rightarrow$" as convergence in distribution. The following result is taken from Shapiro et al. (2014):

Theorem 4.2.1 (Extracted from Theorem 5.7 in Shapiro et al. (2014)) *Suppose that Assumption 4.2.1 holds, $E[h(\tilde{x}, \xi)^2] < \infty$ for some point $\tilde{x} \in \mathcal{X}$, and $\mathcal{X}$ is compact. Given i.i.d. data $\boldsymbol{\xi}_{1:n} = (\xi_1, \ldots, \xi_n)$, consider the SAA problem (4.3). The SAA optimal value $\hat{Z}_n$ satisfies*

$$\sqrt{n}(\hat{Z}_n - Z^*) \Rightarrow \inf_{x \in \mathcal{X}^*} Y(x) \tag{4.4}$$

where $\mathcal{X}^*$ is the set of optimal solutions for (4.1), and $Y(x)$ is a centered Gaussian process on $\mathcal{X}^*$ that has a covariance structure defined by $Cov(h(x_1, \xi), h(x_2, \xi))$ between any $x_1, x_2 \in \mathcal{X}^*$.

Roughly speaking, Theorem 4.2.1 stipulates that, under the depicted conditions, one can use (4.4) to obtain

$$\hat{Z}_n - \frac{\hat{q}}{\sqrt{n}} \tag{4.5}$$

as a valid lower confidence bound for Z^* (and analogously for $\mathcal{G}(\hat{x})$ given $\hat{x}$), where $\hat{q}$ is some suitable error term that captures the quantile of the limiting distribution in (4.4). Indeed, in the case of estimating $\mathcal{G}(\hat{x})$, Bayraksan and Morton (2006) provides an elegant argument that shows that, to achieve $1-\alpha$ confidence, one can take $\hat{q} = z_{1-\alpha}\hat{\sigma}$ where $z_{1-\alpha}$ is the standard normal critical value and $\hat{\sigma}$ is a standard deviation estimate, regardless of whether the limit in (4.4) is a Gaussian distribution (or in other words the solution is unique). Bayraksan and Morton (2006) calls this the single-replication procedure. More precisely, $\hat{\sigma}^2$ is obtained from

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(h(\hat{x},\xi_i) - h(\hat{x}_n^*,\xi_i) - (\bar{h}(\hat{x}) - \bar{h}(\hat{x}_n^*)))^2$$

where $\hat{x}_n^*$ is the solution from (4.3), and $\bar{h}(\hat{x}) - \bar{h}(\hat{x}_n^*) = (1/n)\sum_{i=1}^{n}(h(\hat{x},\xi_i) - h(\hat{x}_n^*,\xi_i))$.

Though Theorem 4.2.1 (and other related work, e.g., Dentcheva et al. (2017), Kleywegt et al. (2002)) is very useful, there are at least two reasons why one would need more general methods:

1. When the decision space contains discrete elements (e.g., combinatorial problems), Assumption 4.2.1 does not hold anymore. There is no guarantee in using the bound (4.5), i.e., it may still be correct but conservative, or it may simply possess incorrect coverages. We note, however, that for some class of problems (e.g., two-stage mixed-integer linear programs), extensions to Theorem 4.2.1 and approaches such as quantile-based bootstrapping (e.g., Eichhorn and Römisch (2007)) are useful alternatives.

2. If the SAA solutions have a "jumping" behavior, namely that program (4.1) has several near-optimal solutions with hugely differing objective variances, then the standard deviation estimate $\hat{\sigma}$ needed in the bound (4.5) can be unreliable. This is because $\hat{\sigma}$ depends heavily on $\hat{x}_n^*$, which can fall close to any of the possible near-optimal solutions with substantial chance and make the resulting estimation noisy. This issue is illustrated in, e.g., Examples 1 and 2 in Bayraksan and Morton (2006).

We should also mention that, as an additional issue, the bias in $\hat{Z}_n$ relative to Z^* can be quite large in any given problem, i.e., arbitrarily close to order $1/\sqrt{n}$ described in the CLT, even if all the

conditions in Theorem 4.2.1 hold (Partani (2007)). Note that this bias is in the optimistic direction (i.e., the resulting bound is still correct, but conservative), and it also appears in the "optimistic bound" approach that we discuss next. There have been techniques such as jackknife (Partani (2007), Partani et al. (2006)) and probability metric minimization (Stockbridge and Bayraksan (2013)) in reducing this bias effect.

4.2.2 Batching Procedures

An alternate approach is to use the optimistic bound (Mak et al. (1999), Shapiro (2003), Glasserman (2013))

$$E[\hat{Z}_n] \leq Z^* \tag{4.6}$$

where $E[\cdot]$ in (4.6) is taken with respect to the data in constructing the SAA value $\hat{Z}_n$. The bound (4.6) holds for any $n \geq 1$, as a direct consequence from exchanging the expectation and the minimization operator in the SAA, and holds as long as $\boldsymbol{\xi}_{1:n}$ are i.i.d.

The bound (4.6) offers a simple way to construct a lower bound for Z^* under great generality. Note that the left hand side of (4.6) is a mean of SAA. Thus, if one can "sample" a collection of SAA values, then a lower confidence bound for Z^* can be constructed readily by using a standard estimate of population mean. To "sample" SAA values, an approach suggested by Mak et al. (1999) is to batch the i.i.d. data set $\boldsymbol{\xi}_{1:n}$ into say m batches, each batch consisting of k observations, so that $mk = n$ (we ignore rounding issues). For each $j = 1, \ldots, m$, solve an SAA using the k observations in the j-th batch; call this value $\hat{Z}_k^j$. Then use

$$\tilde{Z}_k - z_{1-\alpha}\frac{\tilde{\sigma}}{\sqrt{m}} \tag{4.7}$$

where $\tilde{Z}_k = (1/m)\sum_{j=1}^m \hat{Z}_k^j$ and $\tilde{\sigma}^2 = (1/(m-1))\sum_{j=1}^m (\hat{Z}_k^j - \tilde{Z}_k)^2$ are the sample mean and variance from $\hat{Z}_k^j, j = 1, \ldots, m$, and $z_{1-\alpha}$ is the $(1-\alpha)$-level standard normal quantile.

The bound (4.7) does not rely on any continuity of h, and $\tilde{\sigma}/\sqrt{m}$ is simply the sample standard deviation for a sample mean. In these regards, the bound largely circumvents the two concerns described before.

Nonetheless, there is an intrinsic tradeoff between tightness and statistical accuracy in this batching approach. On one hand, m must be chosen big enough (e.g., roughly > 30) so that one can use the CLT to justify the approximation (4.7). Moreover, the larger is m, typically the smaller is the magnitude of the standard error in the second term of (4.7). On the other hand, the larger is k, the closer is $E[\hat{Z}_k^j]$ to Z^* in (4.6), leading to a tighter lower bound for Z^*. This is thanks to a monotonicity property in that $E[\hat{Z}_n]$ is non-decreasing in n Mak et al. (1999). Therefore, there is a tradeoff between the statistical accuracy controlled by m (in terms of the validity of the CLT and the magnitude of the standard error term) and the tightness controlled by k (in terms of the position of $E[\hat{Z}_k^j]$ in (4.6)). In the batching or the so-called multiple-replication approach of Mak et al. (1999), this tradeoff is confined to the relation $mk = n$. There have been suggestions to improve this tradeoff, e.g., by using overlapping batches (Love and Bayraksan (2015, 2011)), but their validity requires uniqueness or exponential convergence of the solution (e.g., in discrete decision space).

4.2.3 Motivation and Overview of Our Approach

Thus, in general, when the sample size n is small, the batching approach appears to necessarily settle for a conservative bound in order to retain statistical validity/accuracy. The starting motivation for the bagging procedure that we propose next is to break free this tightness-accuracy tradeoff. In particular, we offer a bound roughly in the form

$$Z_k^{bag} - \frac{q^{bag}}{\sqrt{n}} \tag{4.8}$$

where Z_k^{bag} is a point estimate obtained from bagging many resampled SAA values, and k signifies the size of the resampled SAA (i.e., the "bags"). The quantity q^{bag} relies on a standard deviation estimate of Z_k^{bag}. Our method operates at a similar level of generality as batching and handles the two concerns Points 1 and 2 in Section 4.2.1: The estimate q^{bag} does not succumb to the "jumping" solution behavior, and the bound holds regardless of the continuity to the decision. Moreover, compared to the batching bound (4.7), our bound has a standard error term shrunk to

order $1/\sqrt{n}$ from $1/\sqrt{m}$ (and relies on an asymptotic on n, not m), thus gaining higher statistical precision. In fact, this term regains the same order of precision level as the bound (4.5) that uses SAA asymptotics directly.

On the other hand, we will show that the choice of k in (4.8), which affects the tightness, can be taken as roughly $o(\sqrt{n})$ in general. Compared with the direct-CLT bound (4.5), our bound appears less tight. However, we consider conditions more general than when (4.5) is applicable. We will see that if we re-impose Lipschitz continuity on the decision (i.e., Assumption 4.2.1), then k can be set arbitrarily close to the order of n. This means that our approach is almost as statistically efficient as the bound (4.5), with the extra benefit of stability in estimating q^{bag}.

Nonetheless, we point out that our approach requires solving resampled SAA programs many times, and is thus computationally more costly than batching and direct-CLT methods. The higher computation cost is the price to pay to elicit our benefits depicted above. Our approach is thus most recommended when statistical performance is of higher concern than computation efficiency, prominently in small-sample situations.

The next section will explain our procedure in more detail. A key insight is to view SAA as a symmetric kernel and the optimistic bound (4.6) as a limiting quantity of an associated symmetric statistic, which can be estimated by bagging. On a high level, the stability in estimating the standard error q^{bag} can be attributed to the nature of bagging as a smoother (Büchlmann and Yu (2002), Efron (2014)).

4.3 Bagging Procedure to Estimate Optimal Values

This section presents our approach. Instead of batching the data, we uniformly resample k observations from $\boldsymbol{\xi}_{1:n}$ for many, say B, times. We use each resample to form an SAA problem and solve it. We then average all these resampled SAA optimal values. The resampling can be done with or without replacement (we will discuss some differences between the two). We summarize our procedure in Algorithm 7.

In the output of Algorithm 7, the first term $\tilde{Z}_k^{bag}$ is the average of many bootstrap resampled

Algorithm 7 Bagging Procedure for Bounding Optimal Values

Given n observations $\boldsymbol{\xi}_{1:n} = \{\xi_1, \ldots, \xi_n\}$, select a positive integer k

for $b = 1$ **to** B **do**

Randomly sample $\boldsymbol{\xi}_k^b = (\xi_1^b, \ldots, \xi_k^b)$ uniformly from $\boldsymbol{\xi}_{1:n}$ (with or without replacement), and solve

$$\hat{Z}_k^b = \min_{x \in \mathcal{X}} \frac{1}{k} \sum_{i=1}^{k} h(x, \xi_i^b)$$

end for

Compute $\tilde{Z}_k^{bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{Z}_k^b$ and

$$\tilde{\sigma}_{IJ}^2 = \begin{cases} \sum_{i=1}^{n} \widehat{Cov}_*(N_i^*, \hat{Z}_k^*)^2, & \text{if resampling is with replacement} \\ \left(\frac{n}{n-k}\right)^2 \sum_{i=1}^{n} \widehat{Cov}_*(N_i^*, \hat{Z}_k^*)^2, & \text{if resampling is without replacement} \end{cases} \tag{4.9}$$

where

$$\widehat{Cov}_*(N_i^*, \hat{Z}_k^*) = \frac{1}{B} \sum_{b=1}^{B} (N_i^b - \frac{k}{n})(\hat{Z}_k^b - \tilde{Z}_k^{bag}) \tag{4.10}$$

and N_i^b is the number of ξ_i that shows up in the b-th resample

Output $\tilde{Z}_k^{bag} - z_{1-\alpha}\tilde{\sigma}_{IJ}$

SAA values, which resembles a bagging predictor by viewing each SAA as a "base learner" (Breiman (1996)). The quantity $\widehat{Cov}_*(N_i^*, \hat{Z}_k^*)$ in (4.10) is the covariance between the count of a specific observation ξ_i in a bootstrap resample, denoted N_i^*, and the resulting resampled SAA value $\hat{Z}_k^*$. The quantity $\tilde{\sigma}_{IJ}^2 = \sum_{i=1}^n \widehat{Cov}_*(N_i^*, \hat{Z}_k^*)^2$ is an empirical version of the so-called infinitesimal jackknife (IJ) estimator (Efron (2014)), which has been used to estimate the standard deviation of bagging schemes, including in random forests or tree ensembles (Wager et al. (2014)). The additional constant factor $(n/(n-k))^2$ in the second line of (4.9) ensures the validity of the IJ estimator under resampling without replacement in certain asymptotic regimes that we will consider.

4.4 SAA as Symmetric Kernel

We explain how Algorithm 7 arises. In short, the $\tilde{Z}_k^{bag}$ in Algorithm 7 acts as a point estimator for $E[\hat{Z}_k]$ in the optimistic bound (4.6), whereas $\tilde{\sigma}_{IJ}^2$ captures the standard error in using this point estimator.

To be more precise, let us introduce a functional viewpoint and write

$$W_k(F) = E_{F^k}[H_k(\xi_1, \ldots, \xi_k)] \tag{4.11}$$

where

$$H_k(\xi_1, \ldots, \xi_k) = \min_{x \in \mathcal{X}} \frac{1}{k} \sum_{i=1}^k h(x, \xi_i)$$

is the SAA value, expressed more explicitly in terms of the underlying data used. Here, the expectation $E_{F^k}[\cdot]$ is generated with respect to i.i.d. variables $(\xi_1, \ldots, \xi_k)$, i.e., F^k denotes the product measure of k F's. For convenience, we denote $E[\cdot]$ as the expectation either with respect to F or the product measure of F's when no confusion arises. Also, we denote $W_k = W_k(F)$.

With these notations, the optimistic bound (4.6) can be expressed as

$$W_k(F) \leq Z^*$$

with the best bound being $W_\infty = \lim_{k \to \infty} W_k \leq Z^*$ thanks to the monotonicity property of the

expected SAA value mentioned before.

Suppose that we have used sampling with replacement in Algorithm 7. Also say we use infinitely many bootstrap replications, i.e., $B = \infty$. Then, the estimator $\tilde{Z}_k^{bag}$ in Algorithm 7 becomes precisely

$$\tilde{Z}_k^{bag} = W_k(\hat{F})$$

where $\hat{F}$ is the empirical distribution formed by $\boldsymbol{\xi}_{1:n}$, i.e., $\hat{F}(\cdot) = (1/n)\sum_{i=1}^n \delta_{\xi_i}(\cdot)$ where $\delta_{\xi_i}(\cdot)$ is the delta measure at ξ_i. If $W_k(\cdot)$ is "smooth" in some sense, then one would expect $W_k(\hat{F})$ to be close to $W_k(F)$. Indeed, when k is fixed, $W_k(F)$, which is expressible as the k-fold expectation under F in (4.11), is multi-linear, i.e.,

$$W_k(F) = E_{F^k}[H_k(\xi_1, \ldots, \xi_k)] = \int \cdots \int H_k(\xi_1, \ldots, \xi_k) \prod_{j=1}^k dF(\xi_j)$$

and is always differentiable with respect to F (in the Gateaux sense) from the theory of von Mises statistical functionals (Serfling (2009)). This ensures that $W_k(\hat{F})$ is close to $W_k(F)$ probabilistically, as elicited by a CLT (Theorem 4.4.1 below).

Note that $W_k(\hat{F})$ is exactly the average of $H_k(\xi_{i_1}, \ldots, \xi_{i_k})$ over all possible combinations of $\{\xi_{i_1}, \ldots, \xi_{i_k}\}$ drawn with replacement from $\boldsymbol{\xi}_{1:n}$. This is equivalent to

$$V_{n,k} = \frac{1}{n^k} \sum_{i_j \in \{1,\ldots,n\}, j=1,\ldots,k} H_k(\xi_{i_1}, \ldots, \xi_{i_k}) \tag{4.12}$$

which is the so-called V-statistic. If we have used sampling without replacement in Algorithm 7, we arrive at the estimator (assuming again $B = \infty$)

$$U_{n,k} = \frac{1}{\binom{n}{k}} \sum_{(i_1,\ldots,i_k) \in \mathcal{C}_k} H_k(\xi_{i_1}, \ldots, \xi_{i_k}) \tag{4.13}$$

where $\mathcal{C}_k$ denotes the collection of all subsets of size k in $\{1, \ldots, n\}$. The quantity (4.13) is known as the U-statistic. The V and U estimators in (4.12) and (4.13) both belong to the class of symmetric statistics (Serfling (2009), Van der Vaart (2000), De la Pena and Giné (2012)), since the estimator is

unchanged against a shuffling of the ordering of the data $\boldsymbol{\xi}_{1:n}$. Correspondingly, the $H_k(\cdot)$ function is known as the symmetric kernel. Symmetric statistics generalize the sample mean, the latter corresponding to the case when $k = 1$.

When $B < \infty$, then $V_{n,k}$ and $U_{n,k}$ above are approximated by a random sampling of the summands on the right hand side of (4.12) and (4.13). These are known as incomplete V- and U-statistics (Lee (2019), Blom (1976), Janson (1984)), and are precisely our $\tilde{Z}_k^{bag}$. As B is chosen large enough, $\tilde{Z}_k^{bag}$ will well approximate $V_{n,k}$ and $U_{n,k}$.

To discuss further, we make the following assumptions:

Assumption 4.4.1 (L_2-boundedness) *We have*

$$E \sup_{x \in \mathcal{X}} |h(x, \xi)|^2 < \infty$$

Denote $g_k(\xi) = E[H_k(\xi_1, \ldots, \xi_k)|\xi_1 = \xi]$. Denote $Var(\cdot) = Var_F(\cdot)$ as the variance under F.

Assumption 4.4.2 (Finite non-zero variance) *We have* $0 < Var(g_k(\xi)) < \infty$.

We have the following asymptotics of $U_{n,k}$ and $V_{n,k}$:

Theorem 4.4.1 *Suppose $k \geq 1$ is fixed, and Assumptions 4.4.1 and 4.4.2 hold. Then*

$$\sqrt{n}(U_{n,k} - W_k) \Rightarrow N(0, k^2 Var(g_k(\xi))) \tag{4.14}$$

and

$$\sqrt{n}(V_{n,k} - W_k) \Rightarrow N(0, k^2 Var(g_k(\xi))) \tag{4.15}$$

as $n \to \infty$, where $N(0, k^2 Var(g_k(\xi)))$ is a normal distribution with mean zero and variance $k^2 Var(g_k(\xi))$.

Proof. Assumption 4.4.1 implies that $EH_k(\xi_{i_1}, \ldots, \xi_{i_k})^2 < \infty$ for any (possibly identical) indices

$i_1, \ldots, i_k$, since

$$EH_k(\xi_{i_1}, \ldots, \xi_{i_k})^2 \leq \frac{1}{k^2} E \sup_{x \in \mathcal{X}} \left(\sum_{j=1}^{k} h(x, \xi_{i_j}) \right)^2 \leq E \sup_{x \in \mathcal{X}} |h(x, \xi)|^2 < \infty \tag{4.16}$$

by the Minkowski inequality. Then, under (4.16) and Assumption 4.4.2, (4.14) follows from Theorem 12.3 in Van der Vaart (2000), and (4.15) follows from Section 5.7.3 in Serfling (2009). □

Theorem 4.4.1 is a consequence of the classical CLT for symmetric statistics. The expression $kg_k(\xi)$, as a function defined on the space $\mathcal{X}$, is the so-called influence function of $W_k(F)$, which can be viewed as its functional derivative with respect to F (Hampel (1974)). Alternately, for a U-statistic $U_{n,k}$, the expression is the so-called Hajek projection (Van der Vaart (2000)), which is the projection of the statistic onto the subspace generated by the linear combinations of $f_i(\xi_i), i = 1, \ldots, n$ and any measurable function f_i. It turns out that these two views coincide, and the U- and V-statistics (whose approximation uses the projection viewpoint and the functional derivative viewpoint respectively) obey the same CLT as depicted in Theorem 4.4.1.

The output of Algorithm 7 is now evident given Theorem 4.4.1. When $B = \infty$, $\tilde{Z}_k^{bag}$ is precisely $U_{n,k}$ under sampling without replacement or $V_{n,k}$ under sampling with replacement. The quantity $\tilde{\sigma}_{IJ}^2$ in Algorithm 7, an empirical IJ estimator, can be shown to approximate the asymptotic variance $k^2 Var(g_k(\xi))/n$ as $n, B \to \infty$, by borrowing recent results in bagging (Efron (2014), Wager and Athey (2018)) (Theorems 4.7.1 and 4.7.2 below show stronger results). Then the procedural output is the standard CLT-based lower confidence bound for W_k.

The discussion above holds for a fixed k, the sample size used in the resampled SAA. It also shows that, at least asymptotically, using with or without replacement does not matter. However, using a fixed k regardless of the size of n is restrictive and leads to conservative bounds. The next subsection will relax this requirement and present results on a growing k against n, which in turn allows us to get a tighter $W_k = E[\hat{Z}_k]$ in the optimistic bound (4.6).

4.5 Asymptotic Behaviors with Growing Resample Size

We first make the following strengthened version of Assumption 4.4.1:

Assumption 4.5.1 ($L_{2+\delta}$-bounded modulus of continuity) *We have*

$$E \sup_{x \in \mathcal{X}} |h(x, \xi) - h(x, \xi')|^{2+\delta} < \infty$$

where ξ, ξ' are i.i.d. generated from F.

Assumption 4.5.1 holds quite generally, for instance under the following sufficient conditions:

Assumption 4.5.2 (Uniform boundedness) *$h(\cdot, \cdot)$ is uniformly bounded over $\mathcal{X} \times \Xi$.*

Assumption 4.5.3 (Uniform Lipschitz condition) *$h(x, \xi)$ is Lipschitz continuous with respect to ξ, where the Lipschitz constant is uniformly bounded in $x \in \mathcal{X}$, i.e.,*

$$|h(x, \xi) - h(x, \xi')| \leq L\|\xi - \xi'\|$$

where $\|\cdot\|$ is some norm in Ξ. Moreover, $E\|\xi\|^{2+\delta} < \infty$.

Assumption 4.5.4 (Majorization)

$$|h(x, \xi) - h(x, \xi')| \leq f(\xi) + f(\xi')$$

where $Ef(\xi)^{2+\delta} < \infty$.

That Assumption 4.5.2 implies Assumption 4.5.1 is straightforward. To see how Assumption 4.5.3 implies Assumption 4.5.1, note that, if the former is satisfied, we have

$$E \sup_{x \in \mathcal{X}} |h(x, \xi) - h(x, \xi')|^{2+\delta} \leq L^{2+\delta} E\|\xi - \xi'\|^{2+\delta} < \infty$$

Similarly, Assumption 4.5.4 implies Assumption 4.5.1 because the former leads to

$$E \sup_{x \in \mathcal{X}} |h(x, \xi) - h(x, \xi')|^{2+\delta} \leq E(f(\xi) + f(\xi'))^{2+\delta} < \infty$$

Next, we also make the following assumption:

Assumption 4.5.5 (Non-degeneracy) *We have*

$$P\left(\min_{x\in\mathcal{X}}\{h(x,\xi)-Z(x)\}>0\right)+P\left(E\left[\min_{x\in\mathcal{X}}\{h(x,\xi)-h(x,\xi')\}\Big|\xi'\right]>0\right)>0 \tag{4.17}$$

where $\xi,\xi' \overset{i.i.d.}{\sim} F$.

Roughly speaking, Assumption 4.5.5 means that ξ is sufficiently mixed so that the optimal value of a data-driven optimization problem with only one (or two) data point can deviate away from its mean. This assumption holds, e.g., when $\mathcal{X}$ lies in a positive region in the real space that is bounded away from the origin. The assumption can be further relaxed in practical problems. For example, one can replace $\mathcal{X}$ in (4.17) by a smaller region that can possibly contain any candidates of optimal solutions. Moreover, if the cost function is Lipschitz (i.e., Assumption 4.2.1 holds), it suffices to replace the entire decision space $\mathcal{X}$ in (4.17) with the set of optimal solutions $\mathcal{X}^*$, namely:

Assumption 4.5.6 (A weaker non-degeneracy condition) *We have*

$$P\left(\min_{x\in\mathcal{X}^*}\{h(x,\xi)-Z^*\}>0\right)+P\left(E\left[\min_{x\in\mathcal{X}^*}\{h(x,\xi)-h(x,\xi')\}\Big|\xi'\right]>0\right)>0 \tag{4.18}$$

where $\mathcal{X}^*$ *is the set of optimal solutions for* (4.1). *In particular, when the optimal solution is unique, i.e.,* $\mathcal{X}^*=\{x^*\}$, *this assumption is reduced to* $Var(h(x^*,\xi))>0$.

An important implication of the above two assumptions is to ensure that $k^2Var(g_k(\xi))$ is bounded away from 0 even as k grows, thus leading to a behavior similar to Assumption 4.4.2 for the finite k case.

Lemma 4.5.1 (Non-degenerate asymptotic variance) *Suppose Assumption 4.4.1 holds. Also suppose either Assumption 4.5.5 holds, or that Assumptions 4.2.1 and 4.5.6 hold jointly and* $\mathcal{X}$ *is compact. Then* $k^2Var(g_k(\xi))>\epsilon>0$ *for some constant* ϵ, *when* k *is sufficiently large.*

The proof of Lemma 4.5.1 uses a coupling argument between $g_k(\xi)=E[H_k(\xi_1,\ldots,\xi_k)|\xi_1=\xi]$, which is a conditional expectation on H_k, and $E[H_k(\xi_1,\ldots,\xi_k)]$, the full expectation on H_k, by

assigning the same random variables $\xi_2, \ldots, \xi_k$. This coupling is used to bound the difference $g_k(\xi) - E[H_k(\xi_1, \ldots, \xi_k)]$ used in calculating the variance $Var(g_k(\xi))$, which then combines with the non-degeneracy condition (Assumption 4.5.5 or 4.5.6) to get a lower bound for $Var(g_k(\xi))$. See Appendix C.1 for the detailed proof.

We have the following asymptotics:

Theorem 4.5.2 (CLT for growing resample size without replacement) *Suppose Assumptions 4.4.1, 4.5.1 and 4.5.5 hold. If the resample size $k = o(\sqrt{n})$, then*

$$\frac{\sqrt{n}(U_{n,k} - W_k)}{k\sqrt{Var(g_k(\xi))}} \Rightarrow N(0,1)$$

where $N(0,1)$ is the standard normal variable.

Theorem 4.5.3 (CLT for growing resample size with replacement) *Suppose Assumptions 4.4.1, 4.5.1 and 4.5.5 hold. If the resample size $k = O(n^\gamma)$ for some constant $\gamma < \frac{1}{2}$, then*

$$\frac{\sqrt{n}(V_{n,k} - W_k)}{k\sqrt{Var(g_k(\xi))}} \Rightarrow N(0,1)$$

where $N(0,1)$ is the standard normal variable.

Theorems 4.5.2 and 4.5.3 are analogs of Theorem 4.4.1 when $k \to \infty$. In both theorems, we see that there is a limit in how large k we can take relative to n, which is thresholded at roughly order $\sqrt{n}$. A symmetric statistic with a growing k is known as an infinite-order symmetric statistic (Frees (1989)), and has been harnessed in analyzing random forests (Mentch and Hooker (2016), Wager et al. (2014), Wager and Athey (2018)). Theorems 4.5.2 and 4.5.3 give the precise conditions under which the SAA kernel results in an asymptotically converging infinite-order symmetric statistic.

The proof of Theorem 4.5.2 utilizes a general projection theorem, in which one can translate the convergence of a projected statistic into convergence of the beginning statistic, if the ratio of their variances tends to 1 (Theorem 11.2 in Van der Vaart (2000); restated in Theorem C.2.1 in Appendix C.2). In our case, the considered projection is the Hajek projection of the infinite-order U-statistic. To execute this theorem, we approximate the variance ratios between the projection

and the remaining orthogonal component. This requires using a further coupling argument among the higher-order conditional expectations, and combining with a representation of the variance ratio in terms of moments of hypergeometric random variables. Then, the CLT for the U-statistic follows by verifying the Lyapunov condition of the Hajek-projected U-statistic.

From Theorem 4.5.2, the conclusion of Theorem 4.5.3 follows by using a relation between U- and V-statistics in the form

$$n^k(U_{n,k} - V_{n,k}) = (n^k - {}_nP_k)(U_{n,k} - R_{n,k}) \tag{4.19}$$

where ${}_nP_k = n(n-1)\cdots(n-k+1)$ and $R_{n,k}$ is the average of all $H_k(\xi_{i_1}, \ldots, \xi_{i_k})$ with at least two of $i_1, \ldots, i_k$ being the same (see, e.g., Section 5.7.3 in Serfling (2009)). By carefully controlling the difference between $U_{n,k}$ and $V_{n,k}$, one can show an asymptotic for $V_{n,k}$ under a similar growth rate of k as that for $U_{n,k}$. This leads to a slightly less general result for $V_{n,k}$ in Theorem 4.5.3. We mention that the growth rates of k in both Theorems 4.5.2 and 4.5.3 are sufficient conditions. We will also see in the next section that, under further conditions, the growth of k can be allowed bigger.

The proofs of Theorems 4.5.2 and 4.5.3 are both in Appendix C.2. These two theorems conclude that $U_{n,k}$ and $V_{n,k}$ continue to well approximate the optimistic bound W_k even as $k \to \infty$, under the depicted assumptions and bounds on the growth rate.

Taking one step further, the following shows that bagging under sampling without replacement achieves almost the same efficiency as the direct use of CLT for SAA in (4.5).

Theorem 4.5.4 *If Assumptions 4.2.1, 4.4.1, 4.5.1 and 4.5.6 hold, and the decision space $\mathcal{X}$ is compact, then the conclusion of Theorem 4.5.2 holds by choosing $k = o(n)$.*

Theorem 4.5.4 implies that, asymptotically, we can use almost the full data set to construct the resampled SAA in $U_{n,k}$. This implies that its standard error is of order close to $1/\sqrt{n}$, and also the point estimate is approximately the SAA with full size n. Hence both the tightness and statistical accuracy of the resulting bound reach the level of (4.5). Moreover, the standard error of our bagging estimator is stabler than the one in (4.5), as it does not rely on the quality of only one

particular SAA solution.

Next we show yet another refinement when, in addition to Lipschitzness, the optimal solution is also unique. Under this additional assumption, our bagging scheme elicits essentially the same CLT as Theorem 4.2.1, and thus recovers the direct-CLT bound in (4.5).

Theorem 4.5.5 (Recovery of the classical CLT for SAA under solution uniqueness) *In addition to the conditions in Theorem 4.5.4, if we further assume that* (4.1) *has a unique optimal solution* $x^* \in \mathcal{X}$*, then the conclusion of Theorem 4.5.2 holds for any* $k \leq n$*. Moreover we have* $k^2 Var(g_k(\xi)) \to Var(h(x^*, \xi))$ *and* $W_k - Z^* = o(1/\sqrt{k})$ *as* $k \to \infty$*. In particular, if* $k \geq \epsilon n$ *for some constant* $\epsilon > 0$*, then*

$$\sqrt{n}(U_{n,k} - Z^*) \Rightarrow N(0, Var(h(x^*, \xi)))$$

where $N(0, Var(h(x^*, \xi)))$ *is the normal variable with mean zero and variance* $Var(h(x^*, \xi))$*.*

Note that, compared with Theorems 4.5.2 and 4.5.3, the centering quantity in Theorem 4.5.5 is changed from W_k to Z^*. The asymptotic distribution is Gaussian with variance precisely the objective variance at x^*. This recovers Theorem 4.2.1 in the special case where $\mathcal{X}^* = \{x^*\}$. If the uniqueness condition does not hold, there could be a discrepancy between the optimistic bound W_∞ and Z^* (This can be hinted by observing the different types of limits between Theorems 4.5.2, 4.5.3 and Theorem 4.2.1, namely Gaussian versus the minimum of a Gaussian process).

We obtain Theorems 4.5.4 and 4.5.5 from a different path than Theorem 4.5.2, in particular by looking at the variance of $U_{n,k}$ via an analysis-of-variance (ANOVA) decomposition (Efron and Stein (1981)) of the symmetric kernel H_k. Thanks to the uncorrelatedness among the ANOVA terms, we can control the variance of $U_{n,k}$ by using a bound from Wager and Athey (2018), which can be shown to depend on the maximum deviation of an empirical process generated by the centered cost function indexed by the decision, i.e., $\mathcal{F} := \{h(x, \cdot) - Z(x) : x \in \mathcal{X}\}$. The Lipschitz assumption allows us to estimate this maximum deviation using empirical process theory. Appendix C.3 shows the proof details.

4.6 Statistical Properties of Bagging Bounds and Comparisons with Batching

We analyze the properties of our confidence bounds implied from Theorems 4.5.2 and 4.5.3, namely consisting of a point estimator $U_{n,k}$ or $V_{n,k}$ and a standard error $k\sqrt{Var(g_k(\xi))/n}$. We first show that the latter is of order $1/\sqrt{n}$, thus reconciling with our claim in (4.8) and demonstrating an asymptotically higher statistical precision compared to the batching bound in (4.7).

Proposition 4.6.1 (Magnitude of the standard error) *Under Assumption 4.4.1, it holds that $k^2 Var(g_k(\xi)) \leq C$ for some constant $C > 0$, as $k \to \infty$. Consequently, the asymptotic standard deviation of $U_{n,k}$ or $V_{n,k}$, namely $k\sqrt{Var(g_k(\xi))/n}$, is of order $O(1/\sqrt{n})$.*

Note that Proposition 4.6.1 is quite general in that it does not impose any growth rate restriction on k. We also note that, under conditions that provide a CLT for the SAA (i.e., Theorem 4.2.1), the $\tilde{\sigma}$ in the batching bound (4.7) can be of order $O(1/\sqrt{k})$ as the data size per batch k grows, and thus the resulting error term there can be controlled to be $O(1/\sqrt{n})$ like ours (and also the direct-CLT bound (4.5)). Nonetheless, Proposition 4.6.1 is free of such type of assumptions. Its proof uses the coupling argument in bounding the variance that appears in the proof of Theorem 4.5.2. The proof details are in Appendix C.4.

The following shows a more revealing result on the higher statistical efficiency of our bagging procedure compared to batching:

Theorem 4.6.2 (Asymptotic variance reduction) *Recall that $\tilde{Z}_k$ is the point estimate in the bound (4.7) given by the batching procedure. Assume the same conditions and resample sizes of either Theorem 4.5.2 or 4.5.4 in the case of resampling without replacement, or Theorem 4.5.3 in the case of resampling with replacement. With the same batch size and resample size, both denoted by k, we define the asymptotic ratios of variance*

$$r_U := \limsup_{n,k\to\infty} \frac{Var(U_{n,k})}{Var(\tilde{Z}_k)},\ r_V := \limsup_{n,k\to\infty} \frac{Var(V_{n,k})}{Var(\tilde{Z}_k)}. \tag{4.20}$$

We have $r_U = r_V = \limsup_{k\to\infty} \frac{kVar(g_k(\xi))}{Var(H_k)} \leq 1$, and in particular

1. $r_U = r_V = 0$ *when* $\lim_{k\to\infty} kVar(H_k) = \infty$

2. $r_U = r_V < 1$ *when the conditions of Theorem 4.5.4 hold and the weak limit* $\inf_{x\in\mathcal{X}^*} Y(x)$ *in* (4.4) *is non-Gaussian*

3. $r_U = r_V = 1$ *when the conditions of Theorem 4.5.4 hold and* $\mathcal{X}^*$ *is a singleton.*

The second case $r_U = r_V < 1$ in Theorem 4.6.2 corresponds to the case of multiple optimal solutions in $\mathcal{X}^*$, which induces a weak scaled limit of the SAA value H_k as the infimum of a Gaussian process that is in general non-Gaussian. The following example shows that in such a setting the asymptotic ratio of variance not only is strictly less than 1 but also can be arbitrarily close to 0.

Example 4.6.1 *Consider the cost function*

$$h(x,\xi) = \begin{cases} (2-x)\xi_1 + (x-1)\xi_2 & \text{if } 1 \le x \le 2 \\ \vdots & \vdots \\ (j+1-x)\xi_j + (x-j)\xi_{j+1} & \text{if } j < x \le j+1 \\ \vdots & \vdots \\ (d-x)\xi_{d-1} + (x-(d-1))\xi_d & \text{if } d-1 < x \le d \end{cases}$$

for $x \in [1,d]$ *and uncertain quantity* $\xi = (\xi_1,\ldots,\xi_d)$ *where* $\xi_j, j = 1,\ldots,d$ *are independent standard normal variables. In other words, at* $x = j$ *the cost* $h(x,\xi)$ *is set to* ξ_j *and everywhere else given by a linear interpolation between the two neighboring integer points. In this case, the objective is constantly zero over the entire decision space so* $\mathcal{X}^* = [1,d]$*. The SAA value* $H_k = \min_{j=1,\ldots,d} \bar{\xi}_j$ *where* $\bar{\xi}_j$ *is the sample mean of the* j*-th component* ξ_j*, hence* $\sqrt{k}H_k$ *is the minimum of* d *independent standard normal variables. A direct application of Corollary 1.9 in Ding et al. (2015) leads to* $kVar(H_k) \ge C/\log d$ *for some universal constant* $C > 0$*. In Appendix C.5 we show that* $\lim_{k\to\infty} k^2 Var(g_k(\xi)) = 1/d$*. Therefore* $r_U = r_V \le \log d/(Cd)$*.*

Furthermore, the following shows that the point estimator under sampling without replacement always has a smaller variance than the batching estimator, for any n and k:

Theorem 4.6.3 (Variance reduction under any finite sample) *Recall that $\tilde{Z}_k$ is the point estimate in the bound* (4.7) *given by the batching procedure. Denote $\{\xi'_1, \ldots, \xi'_n\}$ as the (unordered) collection of values of the data set $\xi_1, \ldots, \xi_n$. With the same batch size and resample size, both denoted by k, we have*

$$Var(\tilde{Z}_k) = Var(U_{n,k}) + E[Var(\tilde{Z}_k | \{\xi_1, \ldots, \xi_n\} = \{\xi'_1, \ldots, \xi'_n\})]$$

and hence $Var(\tilde{Z}_k) \geq Var(U_{n,k})$ for any $k \geq 1$.

Proof. By the law of total variance we have

$$Var(\tilde{Z}_k) = E[Var(\tilde{Z}_k | \{\xi_1, \ldots, \xi_n\} = \{\xi'_1, \ldots, \xi'_n\})] + Var(E[\tilde{Z}_k | \{\xi_1, \ldots, \xi_n\} = \{\xi'_1, \ldots, \xi'_n\}]).$$

The desired conclusion follows from noticing that $E[\tilde{Z}_k | \{\xi_1, \ldots, \xi_n\} = \{\xi'_1, \ldots, \xi'_n\}] = U_{n,k}$. □

Theorem 4.6.3 reinforces the smaller standard error in bagging compared to batching from asymptotic to *any* finite sample, provided that we use sampling without replacement. The key reasoning behind Theorem 4.6.3 is that the batching estimate depends on the ordering of the data; if the data are reordered, then the batching estimate changes. Bagging eliminates the variability due to the ordering of the data by averaging over all the possible combinations. Alternately, one can also interpret bagging as a conditional Monte Carlo scheme applied on the batching estimator given the unordered collection of values realized by the data.

Next, the following result concerns the biases of $U_{n,k}$ and $V_{n,k}$:

Theorem 4.6.4 (Bias) *Under the same assumptions and resample sizes as Theorems 4.5.2 and 4.5.3, the bias of $U_{n,k}$ in estimating W_k is 0, whereas the bias of $V_{n,k}$ in estimating W_k is $O((k^2/n)^l + k/n)$ where l is any fixed positive integer.*

The zero-bias property of $U_{n,k}$ is trivial: Each summand in its definition is an SAA value with distinct i.i.d. data, and thus has mean exactly W_k. On the other hand, the summands in $V_{n,k}$ are SAA values constructed from potentially repeated observations, which induces bias relative to W_k. The proof of the latter again utilizes the relation (4.19), and is left to Appendix C.6.

From Theorem 4.6.4, we see that $U_{n,k}$ outperforms $V_{n,k}$ in terms of bias control. When k is fixed, such an advantage for $U_{n,k}$ is relatively mild, since the bias of $V_{n,k}$ in estimating the optimistic bound W_k is of order $1/n$. However, as k grows, this advantage becomes more significant, and the bias of $V_{n,k}$ can be arbitrarily close to $O(1)$ (when $k \approx \sqrt{n}$).

Theorems 4.5.4, 4.6.3 and 4.6.4 together justify that, in terms of both standard error and bias, sampling without replacement, i.e., $U_{n,k}$, seems to be the more recommendable choice for our bagging procedure. However, in our numerical experiments in Section 4.8, $U_{n,k}$ and $V_{n,k}$ appear to perform quite similarly.

Lastly, we should mention that the biases depicted in Theorem 4.6.4 concern the estimators of W_k, but do not capture the discrepancy between W_k and Z^*. The latter quantity is of separate interest. As discussed at the end of Section 4.2.1, it can be generally reduced by existing methods like the jackknife or probability metric minimization (Partani et al. (2006), Stockbridge and Bayraksan (2013)).

4.7 Error Estimates and Coverages

Finally, we analyze the use of the IJ estimator in approximating the standard error and the error coming from the Monte Carlo noise in running the bootstrap. Together with the results in Section 4.5 and 4.6, these will give us an overall CLT on the output from Algorithm 7. First, we have the following consistency of the IJ variance estimator, relative to the magnitude of the target standard error:

Theorem 4.7.1 (Consistency of IJ estimator without replacement) *Consider the case of resampling without replacement. In any of the following three settings:*

1. *Assumptions 4.4.1, 4.5.1 and 4.5.5 hold and* $k = o(\sqrt{n})$
2. *Assumptions 4.2.1, 4.4.1, 4.5.1 and 4.5.6 hold, the decision space* $\mathcal{X}$ *is compact and* $k = o(n)$
3. *In addition to the assumptions in* 2*, further assume the problem* (4.1) *has a unique optimal solution, and use resample size* $k \leq \theta n$ *for some constant* $\theta < 1$

the IJ variance estimator is relatively consistent, i.e.

$$\frac{n^2}{(n-k)^2}\sum_{i=1}^{n}\mathrm{Cov}_*^2(N_i^*, H_k^*)\Big/\frac{k^2}{n}Var(g_k(\xi)) \xrightarrow{p} 1.$$

Theorem 4.7.2 (Consistency of IJ estimator with replacement) *Consider resampling with replacement. If Assumptions 4.4.1, 4.5.1 and 4.5.5 hold, and $k = O(n^\gamma)$ for some constant $\gamma < \frac{1}{2}$, the IJ variance estimator is relatively consistent, i.e.*

$$\sum_{i=1}^{n}\mathrm{Cov}_*^2(N_i^*, H_k^*)\Big/\frac{k^2}{n}Var(g_k(\xi)) \xrightarrow{p} 1.$$

The three sets of assumptions and resample size in Theorem 4.7.1 are precisely those of Theorem 4.5.2 (the general case), Theorem 4.5.4 (the Lipschitz case) and Theorem 4.5.5 (the Lipschitz and unique solution case) respectively, except a slight tightening on the choice of k in the last case that can only be arbitrarily close to but not exactly n. The assumptions in Theorem 4.7.2 are precisely those of Theorem 4.5.3. Theorem 4.7.1 is justified by adopting the arguments for random forests in Wager and Athey (2018) and a weak law of large numbers, and Theorem 4.7.2 follows from analyzing the difference between U- and V-statistics as in the proof of Theorem 4.5.3. Appendix C.7 shows the details.

When a large enough bootstrap size B is used in Algorithm 7, the Monte Carlo errors in estimating the point estimator and its variance both vanish. This gives an overall CLT for the output of our bagging procedure, as in the next theorem:

Theorem 4.7.3 (CLT for Algorithm 7) *Under the same conditions and resample sizes of Theorem 4.7.1 in the case of resampling without replacement, or those of Theorem 4.7.2 in the case of resampling with replacement, if the bootstrap size B in Algorithm 7 is such that $B/(kn) \to \infty$, then the output of Algorithm 7 satisfies*

$$\frac{\tilde{Z}_k^{bag} - W_k}{\tilde{\sigma}_{IJ}} \Rightarrow N(0,1)$$

where $N(0,1)$ is the standard normal variable.

An immediate consequence of Theorem 4.7.3 is the correct coverage of the true optimal value:

Corollary 4.7.4 (Correct coverage from Algorithm 7) *Under the same assumptions, growth rates of the resample size k and the bootstrap size B in Theorem 4.7.3, the output of Algorithm 7 satisfies*

$$P\left(\tilde{Z}_k^{bag} - z_{1-\alpha}\tilde{\sigma}_{IJ} \leq Z^*\right) \geq P\left(\tilde{Z}_k^{bag} - z_{1-\alpha}\tilde{\sigma}_{IJ} \leq W_k\right) \to 1-\alpha$$

where P is generated under the data $\boldsymbol{\xi}_{1:n}$.

Theorem 4.7.3 and Corollary 4.7.4 thus close our analyses by showing an exact asymptotic coverage of our bagging bound for the optimistic bound W_k, and a correct asymptotic coverage for Z^*, where the exactness of the later depends on the discrepancy between W_k and Z^*. Additionally, Theorem 4.5.5 stipulates that this discrepancy vanishes under the same setting as when the classical SAA CLT has a normal limit, and thus hints that our bound for Z^* is close to having exact coverage in this case.

Lastly, note that B needs to be taken to have order greater than kn to wash away the Monte Carlo error under the considered conditions. To achieve the best result regarding the tightness of the bound, in the case of non-Lipschitzness (Theorems 4.5.2 and 4.5.3) we would choose k to be close to $\sqrt{n}$, which means the need of roughly order $n^{3/2}$ bootstrap size or optimization programs to solve, whereas under Lipschitzness (Theorems 4.5.4 and 4.5.5) we would choose k to be close to n, giving a bootstrap size of order n^2. As discussed previously, because of the computational load, our bagging scheme is most recommended for small-sample situation where n is relatively small. If computation is a concern, one can always use a smaller k in our scheme to speed up computation, with the price of generating a more conservative bound.

4.8 Numerical Experiments

In this section we provide numerical tests to demonstrate the validity of our bagging-based procedures and compare them to the batching procedure given in (4.7) and the single-replication procedure given in (4.5).

Four stochastic optimization problems are tested. The first problem we consider is the $(1-\alpha_1)$-level conditional value at risk (CVaR) of a standard normal variable ξ

$$\min_{x\in\mathbb{R}} x + \frac{1}{\alpha_1} E[(\xi - x)_+] \tag{4.21}$$

where $(\cdot)_+ := \max\{\cdot, 0\}$ denotes the positive part. We set $\alpha_1 = 0.1$, namely, we are solving for the 90%-level CVaR of the standard normal, whose true value can be calculated to be 1.755.

The second one is a portfolio optimization problem where one seeks to minimize the $(1-\alpha_2)$-level CVaR risk measure of an investment portfolio subject to that the expected return of the investment exceeds some target level. Let $\xi = (\xi_1, \ldots, \xi_5)^T$ be the vector of random returns of five different assets whose joint distribution follows a multivariate normal, $x = (x_1, \ldots, x_5)^T$ be the holding proportions of the assets, and b be the target level of expected return. The optimization is described by

$$\begin{aligned} \min_{c,x} \quad & c + \frac{1}{\alpha_2} E[(-\xi^T x - c)_+] \\ \text{s.t.} \quad & E[\xi^T x] \geq b \\ & \sum_{i=1}^{5} x_i = 1 \\ & x_i \geq 0 \text{ for } i = 1, \ldots, 5. \end{aligned} \tag{4.22}$$

In particular, the random return vector ξ follows $N(\mu, \Sigma)$ where the mean $\mu = (1, 2, 3, 4, 5)^T$ and the covariance Σ is randomly generated, $\alpha_2 = 0.05$ and $b = 3$. Note that the cost function here, as well as that in (4.21), is piecewise linear hence Lipschitz continuous, and the optimal solution is unique. Therefore we expect all the methods to perform well for these two problems. Note that, to avoid feasibility complications that divert our focus, in (4.22) we assume knowledge of the expected return μ so the constraint becomes $\mu^T x \geq b$.

To describe the third problem, suppose there are ten different items labeled as #1 through #10 each of which incurs a random loss ξ_i, and one is required to pick at least one out of the ten items and at most two items among $\#7, \#8, \#9, \#10$ in such a way that the total expected loss is minimized. Mathematically, the problem can be formulated as the following stochastic linear

integer program

$$\begin{aligned} \min_{x} \quad & E[\xi^T x] \\ \text{s.t.} \quad & Ax \leq b \\ & x_i \in \{0,1\} \text{ for } i = 1,2,\ldots,10 \end{aligned} \tag{4.23}$$

where ξ follows $N(\mu, \Sigma)$ with mean $\mu = (-1, -7/9, -5/9, \ldots, 7/9, 1)^T \in \mathbb{R}^{10}$ and covariance Σ randomly generated, $b = (-1, 2)^T$ and

$$A = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

It is straightforward to see that picking the items with negative expected losses, i.e., #1 through #5, gives the minimum total loss, hence the unique optimal solution is $x^* = (1,1,1,1,1,0,0,0,0,0)^T$ with a total loss -2.78. Because of the integrality requirement the single-replication procedure is not theoretically justified and can exhibit incorrect coverage. When implementing the methods we solve the SAA problems by a direct enumeration (feasible thanks to the relatively low dimensionality).

The fourth optimization problem is the following simple stochastic linear program

$$\begin{aligned} \min_{x} \quad & E[-0.05x + (3 - 2x)\xi] \\ \text{s.t.} \quad & -1 \leq x \leq 1 \end{aligned} \tag{4.24}$$

where the uncertain quantity ξ is a standard normal and the decision x is a scalar. It is clear that the optimal value is -0.05 at $x^* = 1$. This problem serves to highlight that, although the optimization is highly smooth, using past methods may give subpar finite-sample performances due to a delicate interplay between the variance and jumping behavior of the estimated solution. It then illustrates how bagging can be a resolution in such a scenario.

4.8.1 Lower Bounds of Optimal Values

In this subsection we use bagging without replacement, bagging with replacement, the batching procedure (4.7) and the single-replication procedure (4.5) to compute lower confidence bounds for

the optimal value $Z^* = \min_{x\in\mathcal{X}} Z(x)$. Specifically, we first simulate an i.i.d. data set $\xi_1, \ldots, \xi_n$ of size n, and then compute a 95% lower bound of the optimal value using each of the four methods. As suggested by Theorem 4.7.3, we set B, the number of resamples, in Algorithm 7 to be roughly $5nk$ to wash out the effect of Monte Carlo error in estimating the covariances. This is in accordance with our focus on statistical efficiencies, under the presumed adequate resources in solving SAA problems. In the batching procedure we use the quantile of t-distribution with $m-1$ degrees of freedom when there are less than 30 batches, so as to enhance finite-sample performances as suggested in Mak et al. (1999), whereas in other procedures we use the normal quantile.

Tables 4.1 and 4.2 summarize the results for problem (4.21) when the data size $n = 50$ and $n = 300$, whereas Table 4.3 shows those for problem (4.24). We compute 1000 confidence bounds from 1000 independently generated data sets, and then average the results to estimate coverage probability (c.p.(%)), mean of the lower bound (mean) and standard deviation of the lower bound (std.). We use k to denote either batch size in the batching procedure or resample size in our bagging procedures. The "NA" entries in the tables correspond to the cases where $n/k < 2$ hence the batching procedure is not tested. The "Single-replication" column of each table has only one row because all the n data are used to form the SAA in the single-replication procedure.

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
10	99.4	1.00	0.29	99.4	1.16	0.22	99.7	1.16	0.20	95.7	1.33	0.24
25	97.1	0.36	0.96	98.9	1.23	0.22	99.6	1.23	0.21			
40	NA	NA	NA	98.6	1.26	0.23	98.5	1.26	0.23			

Table 4.1: Problem (4.21), $n = 50$. Lower bounds of optimal values.

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
10	100	1.37	0.10	100	1.38	0.09	100	1.38	0.08	96.9	1.57	0.10
30	99.1	1.48	0.11	99.8	1.50	0.09	99.5	1.51	0.09			
50	97.8	1.50	0.12	98.7	1.53	0.10	98.9	1.52	0.10			
100	96.9	1.44	0.19	97.9	1.55	0.10	98.0	1.55	0.09			
150	96.1	1.20	0.42	98.0	1.55	0.10	97.6	1.55	0.10			
250	NA	NA	NA	96.9	1.55	0.10	98.1	1.56	0.10			

Table 4.2: Problem (4.21), $n = 300$. Lower bounds of optimal values.

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
5	100	-1.23	0.38	100	-1.21	0.35	100	-1.23	0.37	95.5	-0.63	0.59
10	98.9	-1.08	0.48	99.8	-0.99	0.39	100	-1.01	0.37			
25	94.0	-0.93	0.63	98.9	-0.79	0.44	99.0	-0.82	0.40			
50	95.0	-1.57	1.55	97.1	-0.67	0.47	97.5	-0.72	0.44			
70	NA	NA	NA	95.1	-0.62	0.51	97.5	-0.69	0.43			

Table 4.3: Problem (4.24), $n = 100$. Lower bounds of optimal values.

Tables 4.1-4.3 show that for a wide range of resample sizes, namely from 10 to more than half of the data size, our bagging procedures generate statistically valid lower bounds in the sense that the coverage probabilities are equal to or above the nominal value 95%. The batching and single-replication procedures also generate valid confidence bounds. The results across different values of k verify the relation between the resample size and tightness of the optimistic bound (4.6). To be specific, in all the tables, as the resample size k grows, the mean lower bound gets closer to the true optimal value 1.755 in Tables 4.1 and 4.2 and -0.05 in Table 4.3. In particular, in the case of problem (4.21) and $n = 300$ (Table 4.2) bagging (with or without replacement) provides a lower bound as good as 1.55 with coverage probability 97%-98% by using $k = 100, 150, 250$. It therefore appears that, with the bagging procedures, one can obtain a relatively tight bound for the optimal value and in the meantime retain good statistical accuracy, by using a resample size k that is roughly half the data size.

Although the bounds generated from all considered methods are statistically valid, they differ in tightness and stability. We observe that our bagging procedures appear to output tighter and stabler bounds on the optimal value than batching. In each of Tables 4.1-4.3, under the same batch size or resample size k, the bounds given by bagging (with or without replacement) are always larger in terms of the mean, and meanwhile less variable as measured by the standard deviation, than those by batching. The difference in tightness and stability becomes more noticeable as k increases. This is in accordance with benefit of reducing variance in using bagging procedures as illustrated by Theorems 4.6.2 and 4.6.3.

The results also show the tradeoff between tightness and statistical accuracy in the batching procedure. According to the monotonicity property of the optimistic bound, the confidence bound

should exhibit a monotonic trend of becoming tighter as the batch size k increases. However, in all the tables the mean lower bound first gets tighter for relatively small batch size but then becomes looser again as the size further increases. For example, in Table 4.3 the tightest bound (in terms of the mean) is -0.93 at $k = 25$ and in Table 4.2 the tightest is 1.50 at $k = 50$. This non-monotonic behavior appears since, as the batch size gets large, too few batches are available for the procedure to maintain the desired statistical accuracy (i.e. a coverage probability above 95%). To mitigate this issue, we resort to using t-quantile in place of normal which loosens the bound in exchange for correct coverages. In fact, if we change the t-quantile to normal the coverage probability drops to 92% in Table 4.1 and 86% in Table 4.2 in our experiment. Note that such kind of tradeoff no longer appears in our bagging procedures as the bound always gets tighter and at the same time has the desired coverage level even for large k.

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
2	99.7	3.94	1.79	99.8	3.82	1.79	99.9	3.79	1.76	93.0	1.89	1.65
5	99.5	3.17	1.79	99.4	3.01	1.78	99.3	3.07	1.88			
10	99.5	3.10	1.96	98.7	2.53	1.75	97.6	2.50	1.74			
20	99.0	6.94	5.36	97.1	2.09	1.71	97.6	2.02	1.69			
27	NA	NA	NA	96.9	2.10	1.67	95.1	2.00	1.61			

Table 4.4: Problem (4.22), $n = 40, n_1 = 20, n_2 = 20$. Upper bounds of optimality gaps by BC.

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
2	99.9	4.32	2.41	100	4.36	2.41	100	4.24	2.35	91.4	1.20	1.31
5	99.3	3.66	2.71	99.1	3.18	2.14	99.5	3.24	2.28			
10	99.3	4.92	4.59	97.1	2.43	2.02	98.7	2.73	2.21			
15	NA	NA	NA	94.7	2.13	1.87	98.2	2.53	2.22			
20	NA	NA	NA	NA	NA	NA	96.7	2.09	1.81			

Table 4.5: Problem (4.22), $n = 40, n_1 = 20, n_2 = 20$. Upper bounds of optimality gaps by CRN.

4.8.2 Upper Bounds of Optimality Gaps

Now we test our methods in bounding optimality gaps of solutions. In our experiments we first solve the SAA formed by n_1 data points $\xi_1, \ldots, \xi_{n_1}$ to obtain a solution $\hat{x}$. We then generate

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
5	100	5.77	1.84	100	5.60	1.69	100	5.58	1.74	100	4.31	1.67
10	100	5.44	1.94	100	4.93	1.61	100	5.10	1.74			
25	100	5.57	2.16	100	4.61	1.69	100	4.78	1.67			
50	100	11.04	6.76	100	4.49	1.64	100	4.45	1.60			
70	NA	NA	NA	100	4.33	1.62	100	4.42	1.63			
90	NA	NA	NA	100	4.25	1.63	100	4.47	1.64			

Table 4.6: Problem (4.23), $n = 100, n_1 = 64, n_2 = 36$. Upper bounds of optimality gaps by BC.

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
6	100	2.14	1.04	100	1.83	0.76	100	1.94	0.74	86.1	0.93	0.84
9	99.3	1.88	1.17	100	1.44	0.75	100	1.63	0.78			
18	96.7	2.46	2.59	99.4	1.17	0.83	99.7	1.27	0.72			
30	NA	NA	NA	94.0	0.97	0.84	99.6	1.23	0.82			

Table 4.7: Problem (4.23), $n = 100, n_1 = 64, n_2 = 36$. Upper bounds of optimality gaps by CRN.

n_2 independent data points $\xi_{n_1+1}, ..., \xi_{n_1+n_2}$. These (and possibly the first n_1 data points) are then used to compute an upper confidence bound for the optimality gap $\mathcal{G}(\hat{x}) = Z(\hat{x}) - Z^*$. For convenience we denote $n = n_1 + n_2$ as the total sample size in the experiments.

We consider two approaches to bounding the gap, one reusing the first n_1 data points, and the other not. The first approach is to use the Bonferroni Correction (BC). Specifically, we use the second group of n_2 data to compute $U = \bar{h} + z_{0.975}\hat{\sigma}/\sqrt{n_2}$ as a 97.5% upper confidence bound of $Z(\hat{x})$, where $\bar{h}, \hat{\sigma}^2$ are the sample mean and variance of $h(\hat{x}, \xi_{n_1+1}), \ldots, h(\hat{x}, \xi_n)$, and compute a 97.5% lower confidence bound L of the true optimal value Z^* using all the n data as in the previous section. In the end we output $U - L$ as a confidence bound for the gap $\mathcal{G}(\hat{x})$. By BC we know

$$P(U - L \geq Z(\hat{x}) - Z^*) \geq P(U \geq Z(\hat{x})) + P(L \leq Z^*) - 1 \approx 97.5\% + 97.5\% - 1 = 95\%$$

hence $U - L$ is an asymptotically valid 95% confidence bound for the gap.

The second approach is a Common Random Numbers (CRN) variance-reduction technique proposed by Mak et al. (1999) in this context. Consider minimizing a different objective $E[h(x, \xi) - h(\hat{x}, \xi)]$ as a whole, where $\hat{x}$ is viewed as fixed, whose optimal value is exactly $-\mathcal{G}(\hat{x})$. We use the

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
5	100	2.25	1.06	100	2.14	1.03	100	2.22	1.05	100	1.57	1.17
10	100	2.06	1.11	100	1.95	1.06	100	2.01	1.06			
25	100	2.02	1.30	100	1.72	1.07	100	1.75	1.06			
50	100	3.75	3.15	100	1.59	1.07	100	1.70	1.09			
70	NA	NA	NA	100	1.54	1.11	100	1.66	1.10			

Table 4.8: Problem (4.24), $n = 100, n_1 = 64, n_2 = 36$. Upper bounds of optimality gaps by BC.

	Batching			Bagging w/o replacement			Bagging w/ replacement			Single-replication		
k	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.	c.p.(%)	mean	std.
3	100	1.61	0.55	100	1.53	0.50	100	1.53	0.53	79.5	0.80	0.85
9	97.8	1.29	0.89	99.9	1.10	0.57	99.9	1.12	0.55			
18	90.8	2.09	2.13	98.1	0.93	0.66	99.3	0.98	0.63			
30	NA	NA	NA	92.1	0.87	0.79	97.7	0.91	0.67			

Table 4.9: Problem (4.24), $n = 100, n_1 = 64, n_2 = 36$. Upper bounds of optimality gaps by CRN.

second group of n_2 data to compute a 95% lower confidence bound for this new optimization problem, and then negate the lower bound to obtain a valid upper bound for $\mathcal{G}(\hat{x})$.

Tables 4.4, 4.6 and 4.8 summarize the results for problems (4.22)(4.23)(4.24) using BC, while Tables 4.5, 4.7 and 4.9 display those using CRN. Note that, in either approach, in order to guarantee the statistical accuracy of the confidence bound a relatively small number of data (e.g., around 30) would suffice. In view of this, we choose n_2 around 30 in all the experiments.

We see a few similar observations as in Section 4.8.1 where we compute lower bounds for optimal values. The two bagging procedures generate statistically valid upper bounds in almost all the cases (mildly undercover in the case $k = 30$ of Table 4.9). The bounds by batching also possess the desired coverage probability in most cases, but are looser (i.e., larger) than those given by bagging. It can be seen that the tightest bound by batching can be twice that by bagging (e.g., in Tables 4.5 and 4.7). Like in Tables 4.1-4.3, the batching bounds are also more variable, as measured by the standard deviation, than the bagging-based bounds under the same resample size k.

Some new observations are as follows. First, we see that the single-replication procedure suffers from severe under-cover issues in problems (4.23) and (4.24) (86.1% in Tables 4.7 and 79.5% in Table 4.9). In problem (4.23) this can be attributed to the integrality requirement on the decision.

In problem (4.24) the optimization itself is smooth, and the issue lies in the delicate relation between the variance and the jumping behavior of the estimated solution. We find that with high probability the candidate solution $\hat{x}$ is -1 (with optimality gap 0.1), and, given $\hat{x} = -1$, solving the SAA associated with the new cost function $h(x, \xi) - h(-1, \xi)$ gives the solution -1 again with high probability. However this way the estimated variance $\hat{\sigma}^2$ will be zero (because the new cost function is constantly 0 at $x = -1$) which causes the under-cover issue. Similar observations have been discussed in Section 6 of Bayraksan and Morton (2006). On the contrary, our bagging procedures mitigate this by estimating the variance using all the resampled SAA solutions.

Second, in general the CRN approach enjoys the benefit of generating tighter and stabler confidence bounds than the BC approach thanks to variance reduction. By comparing Table 4.6 with Table 4.7 or Table 4.8 with Table 4.9, we see that this benefit of CRN becomes more significant when one invests more data in obtaining $\hat{x}$, i.e. when n_1 is chosen larger. This is because, the closer the estimated solution $\hat{x}$ gets to the true optimum x^*, the smaller is the variance of the gap function $h(x, \xi) - h(\hat{x}, \xi)$ at the optimum (i.e., x^*) due to the continuity of its variance (as a function of x), which in turn leads to a smaller standard error. We also observe that the BC approach tends to over-cover the optimality gap, potentially because of the looseness of the union bound.

4.9 Conclusion

We have studied a bagging approach to estimate bounds for the optimal value, and consequently the optimality gap for a given solution in stochastic optimization. We demonstrate how our approach works under minimal regularity conditions, including for non-smooth problems, and exhibits competitive statistical efficiency and stability. Compared to batching, our approach generates a new tradeoff between bound tightness and statistical accuracy that is especially beneficial in small-sample situations. Compared to approaches based on direct SAA asymptotics, our approach requires less smoothness conditions on the objectives and gives more stable estimates thanks to the smoothing effect of bagging. These benefits, however, are offset by the price of more computation in repeatedly solving SAA programs. We have developed the theoretical properties of our approach

by viewing SAA as a kernel in infinite-order symmetric statistics, and have illustrated our findings with numerical results.

Chapter 5

Combating Conservativeness in Optimization with Uncertain Constraints

5.1 Introduction

In Chapter 4 we have studied a bagging-based approach to quantifying uncertainties in the objective, whereas in this chapter we consider the case where the uncertainty lies in the constraint. We focus on optimization with stochastic or probabilistic constraints that, on a high level, can be written in the form

$$\min_x f(x) \;\; \text{subject to} \;\; H(x) \geq \gamma \tag{5.1}$$

where $H(x)$ is an expectation of a random function of the decision variable x. Formulation (5.1) is ubiquitous in decision-making problems under multiple tradeoffs, where the constraint $H(x) \geq \gamma$ signifies a restriction on the risk level or resource capacity (e.g., Atlason et al. (2004), Krokhmal et al. (2002)). Moreover, when the random function is an indicator of an event, formulation (5.1) is a so-called probabilistically constrained or chance-constrained problem (CCP) (Prékopa (2003)). This important formulation posits the decision to tolerate a small probability on catastrophic events

such as system failures and big losses, and comprises a primary approach for safe decision-making when facing uncertainty.

We are interested in the situation where the probability distribution governing H is unknown but only observed through data. Finding good solutions under this setting has been studied prominently in the data-driven optimization literature, harnessing various tools from (distributionally) robust optimization (e.g., Bertsimas et al. (2011), Ben-Tal et al. (2009), Wiesemann et al. (2014)) to sample average approximation (e.g., Shapiro et al. (2014)) and scenario optimization (e.g., Campi and Garatti (2008)). From a statistical viewpoint, the problem challenge and the focus of these studies can be cast as a balancing between feasibility and optimality. Due to data noise, feasibility is at best guaranteed with a high statistical confidence, and accounting for this uncertainty incurs a price on the achieved objective value – resulting in conservativeness. This impact on optimality from ensuring feasibility depends heavily on the efficiency in assimilating statistical information into the data-driven formulation. In the following, we first explain how the established estimation frameworks can face severe "looseness" in this regard and lead to over-conservative solutions. This motivates our study that, on a high level, aims to investigate a strategy to substantially tighten the feasibility-optimality tradeoff compared to the previous methods.

5.1.1 Existing Frameworks and Motivation of Our Approach

To facilitate discussion, suppose for concreteness that the decision variable x lies in a d-dimensional deterministic space $\mathcal{X} \subset \mathbb{R}^d$. Denote $H(x) := \mathbb{E}_F[h(x, \xi)]$ where $\mathbb{E}_F[\cdot]$ is the expectation under $\xi \sim F$, and $h(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is a function of $x \in \mathcal{X}$ controlled by the randomness $\xi \in \mathbb{R}^m$. Also, since our focus is on handling uncertain constraints, we assume that the objective function f is deterministic (this can be relaxed with proper modifications of our subsequent discussion). Suppose we have i.i.d. observations $\xi_1, \ldots, \xi_n$.

Let us first consider a natural idea to replace the unknown $H(\cdot)$ with some point estimate, say the sample average $\hat{H}(\cdot) = (1/n) \sum_{i=1}^{n} h(x, \xi_i)$, in the constraint. Though simple, this approach is typically inadequate to ensure feasibility in any statistical sense. To explain, suppose the "true" optimal solution x^* is at the boundary of the feasible region, i.e., $H(x^*) = \gamma$. If we use

$(1/n)\sum_{i=1}^{n} h(x,\xi_i) \geq \gamma$ as the constraint, then, with significant probability an obtained solution $\hat{x}^*$ (conceivably also at the boundary of the data-driven constraint) can have $H(\hat{x}^*)$ below γ (when $(1/n)\sum_{i=1}^{n} h(\hat{x}^*,\xi_i) > \mathbb{E}_F[h(\hat{x}^*,\xi)]$), which is infeasible for the original problem. This issue may not arise if x^* or $\hat{x}^*$ is in the interior of the feasible region, but a priori we do not know our decision. In other words, the nature of constrained optimization enforces us to put some "safety" margin in addition to the point estimate, in order to achieve any reasonable confidence in feasibility. Here, we can plausibly use a data-driven constraint $\hat{H}(x) - \epsilon(x) \geq \gamma$, where $\epsilon(x)$ is a properly chosen positive function such that $H(x) \geq \hat{H}(x) - \epsilon(x)$ for any x with high confidence (such as the scheme in Wang and Ahmed (2008), among others).

We place the above discussion in a more general framework. Let $\mathcal{F}$ be the (unknown) feasible region of (5.1). Given the data $\xi_1,\ldots,\xi_n$, by a valid procedure we mean one that is able to output a solution $\hat{x}^*$ that is truly feasible with a given high confidence level, say $1-\beta$ (e.g., 95%). That is,

$$P_{data}(\hat{x}^* \in \mathcal{F}) \geq 1-\beta \tag{5.2}$$

where P_{data} refers to the probability with respect to the data. By a data-driven reformulation, we mean replacing $\mathcal{F}$ with $\hat{\mathcal{F}}$ that is constructed solely from the data $\xi_1,\ldots,\xi_n$. This gives

$$\min_{x\in\mathcal{X}} \; f(x) \;\; \text{subject to} \;\; x \in \hat{\mathcal{F}} \tag{5.3}$$

which outputs solution $\hat{x}^*$. If we can choose $\hat{\mathcal{F}}$ such that

$$P_{data}(\hat{\mathcal{F}} \subset \mathcal{F}) \geq 1-\beta \tag{5.4}$$

then we clearly have (5.2) since $P_{data}(\hat{x}^* \in \mathcal{F}) \geq P_{data}(\hat{\mathcal{F}} \subset \mathcal{F})$. In the example above, we have used $\hat{\mathcal{F}} = \{x \in \mathcal{X} : \hat{H}(x) - \epsilon(x) \geq \gamma\}$, in the hope that (5.4) holds in order to achieve (5.2).

We contend that most approaches in data-driven optimization rely on the above reasoning and are based on (5.4). In particular, (5.4) provides a convenient way to certify feasibility, by requiring that *all* solutions feasible for (5.3) are also feasible for (5.1) with high confidence. This set-level

guarantee generally hinges on a simultaneous estimation task across all x in the decision space $\mathcal{X}$, for which a proper control of the statistical error can lead to a substantial shrinkage of the size of $\hat{\mathcal{F}}$ that exacerbates with problem dimension (either of the decision space or the probability space).

We provide several examples to illustrate the phenomenon above. Some of these examples apply most relevantly to CCP, where $H(x)$ is in the form $\mathbb{P}_F(G(x,\xi) \le b)$ with $G(x,\xi) : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$.

Example 5.1.1 (Sample average approximation (SAA)) *In the case of CCP, the SAA approach sets* $\hat{\mathcal{F}} = \{x \in \mathcal{X} : \frac{1}{n}\sum_{i=1}^n \mathbf{1}(G(x,\xi_i) + \epsilon \le b) \ge \gamma + \delta\}$, *where* ϵ *and* δ *are suitably tuned parameters. For example, when* G *is Lipschitz continuous in* x, *selecting* $\delta = \Omega(\sqrt{(d/n)\log(1/\epsilon)})$ *can guarantee* (5.4) *(Luedtke and Ahmed (2008)), and similar relations also hold in discrete decision space (Luedtke and Ahmed (2008)) and expected value constraints (Wang and Ahmed (2008)). These estimates come from concentration inequalities in which union bounds are needed and give rise to the dependence on the dimension* d. *Note that the resulting margin* δ *scales in order* $\sqrt{d}$, *and to get any reasonably small* δ, n *must be of higher order than* d. □

Example 5.1.2 (Robust optimization (RO) and safe convex approximation (SCA)) *For the case of CCP, RO sets*

$$\hat{\mathcal{F}} = \{x \in \mathcal{X} : G(x,\xi) \le b, \text{ for all } \xi \in \mathcal{U}\} \tag{5.5}$$

where $\mathcal{U}$ *is known as the uncertainty set, and* ξ *in* (5.5) *is viewed as a deterministic unknown (Bertsimas et al. (2011), Ben-Tal et al. (2009)). A common example of* $\mathcal{U}$ *is an ellipsoidal set* $\{\xi : (\xi - \hat{\mu})'\hat{\Sigma}^{-1}(\xi - \hat{\mu}) \le \rho\}$ *where* $\hat{\mu} \in \mathbb{R}^d$, $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ *a positive semidefinite matrix, and* $\rho \in \mathbb{R}$. *Here the center* $\hat{\mu}$ *and shape* $\hat{\Sigma}$ *typically correspond to the mean and covariance of the data, and* ρ *controls the set size. A duality argument shows that, in the case of linear chance constraint in the form* $G(x,\xi) = x'\xi$, (5.5) *is equivalent to the quadratic constraint* $\hat{\mu}'x + \sqrt{\rho}\|\hat{\Sigma}^{1/2}x\|_2 \le b$. *Using such type of convex constraints as inner approximations for intractable chance constraints is also known as SCA (e.g., Ben-Tal and Nemirovski (2000), Nemirovski (2003), Nemirovski and Shapiro (2006)).*

It is known that, if for instance the random variable ξ *has a known bounded support, the above*

approach guarantees an obtained solution has a satisfaction probability of order $1-e^{-\rho/2}$ *via Hoeffding's inequality, and* ρ *is chosen by matching this expression with the tolerance level* γ*. Although* ρ *calibrated this way may not explicitly depend on the problem dimension, its tightness varies heavily based on problem instance (due to the worst-case nature of concentration bounds), and its validity relies on a priori distributional information (e.g., support) rather than an efficient utilization of data. Another viewpoint that has been taken recently in data-driven RO (Bertsimas et al. (2018a); Tulabandhula and Rudin (2014); Goldfarb and Iyengar (2003); Hong et al. (2017)) is to select* $\mathcal{U}$ *to be a set that contains* γ*-content of the distribution of* ξ*, i.e.,* $\mathbb{P}_F(\xi \in \mathcal{U}) \geq \gamma$*, with a confidence level* $1-\beta$*. In this case, any solution* $\hat{x}$ *feasible for* (5.5) *would satisfy* $\mathbb{P}_F(G(\hat{x},\xi) \leq b) \geq \mathbb{P}_F(\xi \in \mathcal{U}) \geq \gamma$ *with at least* $1-\beta$ *confidence, thus achieving* (5.4) *as well. Such generated uncertainty set however typically has a size that scales with the dimension of the probability space. For example, consider* $G(x,\xi) = x'\xi$ *with* $\xi \in \mathbb{R}^m$ *being standard multivariate Gaussian and the uncertainty set* $\mathcal{U}$ *is an ellipsoid with* $\hat{\mu}$ *and* $\hat{\Sigma}$ *being the true mean and covariance, i.e.,* $\mathcal{U} = \{\xi \in \mathbb{R}^m : \|\xi\|_2^2 \leq \rho\}$*. Then, in order to make* $\mathcal{U}$ *a* γ*-content set the radius* ρ *has to be at least of order* m *since* $\|\xi\|_2^2$ *has a mean* m*, resulting in the robust counterpart* $\sqrt{\rho}\,\|x\|_2 = \Theta(\sqrt{m})\,\|x\|_2 \leq b$*. However, the exact chance constraint in this case can be rewritten as* $z_\gamma \|x\|_2 \leq b$*, where* z_γ *is the* γ*-quantile of the univariate standard normal, which is independent of the dimension.* □

Example 5.1.3 (Distributionally robust optimization (DRO)) *DRO sets*

$$\hat{\mathcal{F}} = \{x \in \mathcal{X} : \inf_{Q \in \mathcal{U}} \mathbb{E}_Q[h(x,\xi)] \geq \gamma\} \tag{5.6}$$

where $\mathcal{U}$ *is a set in the space of probability measures that is constructed from data, and is often known as the ambiguity set or uncertainty set. The rationale here is similar to RO, but views the uncertainty in terms of the distribution. If* $\mathcal{U}$ *is constructed such that it contains the true distribution* F *with high confidence, i.e.,* $P_{data}(F \in \mathcal{U}) \geq 1-\beta$*, then any solution* $\hat{x}$ *feasible for the DRO constraint* (5.6) *would satisfy* $\mathbb{P}_F(G(\hat{x},\xi) \leq b) \geq \gamma$ *with at least* $1-\beta$ *confidence so that* (5.4) *holds.*

Popular choices of $\mathcal{U}$ *include moment sets, i.e., specifying the moments of* Q *(to be within a*

range for instance) (Ghaoui et al. (2003); Delage and Ye (2010); Xu et al. (2012); Wiesemann et al. (2014); Goh and Sim (2010); Natarajan et al. (2008); Van Parys et al. (2016); Doan et al. (2015); Hanasusanto et al. (2015)), and distance-based sets, i.e., specifying Q in the neighborhood ball surrounding a baseline distribution, where the ball size is measured by a statistical distance such as ϕ-divergence (Petersen et al. (2000); Ben-Tal et al. (2013); Glasserman and Xu (2014); Lam (2016b); Lam (2018); Hu and Hong (2013); Jiang and Guan (2016); Gotoh et al. (2018); Dupuis et al. (2016); Bayraksan and Love (2015)) or Wasserstein distance (Esfahani and Kuhn (2018); Blanchet and Murthy (2019); Gao and Kleywegt (2016); Xie (2019)).

Ensuring $P_{data}(F \in \mathcal{U}) \geq 1-\beta$ means that $\mathcal{U}$ is a confidence region for F. In the moment set case, this boils down to finding confidence regions for the moments whose sizes in general scale with the probability space dimension. To explain, when only the mean $\mathbb{E}_F[\xi]$ is estimated, the confidence region constructed from, say the delta method (Marandi et al. (2019)), takes the form $\{\hat{\mu} + \hat{\Sigma}^{\frac{1}{2}} v : v \in \mathbb{R}^m, \|v\|_2^2 \leq \chi^2_{m,1-\beta}\}$, where $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance and $\chi^2_{m,1-\beta}$ (which is of order m) is the $1-\beta$ quantile of the χ^2 distribution with degree of freedom m, therefore the diameter of the confidence region scales as $\sqrt{m}$. When the mean and covariance are jointly estimated, the dimension dependence scales up further. In the distance-based set case, one needs to estimate statistical distances. If the Wasserstein distance is used to construct the ball surrounding the empirical distribution, results from measure concentration (Fournier and Guillin (2015)) indicate that the ball size needs to be of order $n^{-\frac{1}{m}}$ to ensure $P_{data}(F \in \mathcal{U}) \geq 1-\beta$. Alternatively, if $\mathcal{U}$ is constructed as a ϕ-divergence ball surrounding some nonparametric kernel-type density estimate, results from kernel density estimation (see Section 4.3 in Wand and Jones (1994)) suggest that the estimation error is of order $n^{-\frac{4}{m+4}}$. In either case, the required size of the uncertainty set exhibits exponential dependence on the dimension. Recently, the empirical or the profile likelihood method has also been proposed to calibrate the ball size such that $\mathcal{U}$ can be (much) smaller than what is needed in being a confidence region for F, while at the same time (5.4) *still holds (Lam and Zhou (2017); Duchi et al. (2016); Lam (2019); Blanchet and Kang (2020)). However, the ball size in this approach scales as the supremum of a so-called χ^2-process over the decision space (e.g., Lam (2019)). An analysis using metric entropy (e.g., Example 2 in Section*

14 in Lifshits (2013)) shows that the χ^2-process supremum can scale linearly in the decision space dimension d, a much better but still considerable dependence on the dimension. □

Finally, we discuss the only two exceptional paradigms, to our best knowledge, in providing guarantee (5.2) using (5.4). First, Gupta (2019) studies a Bayesian framework to define feasibility guarantees for (stochastic) constraints with unknown parameters, focusing on DRO formulations. The idea is to ensure the obtained data-driven solution satisfies the constraints with a high posterior probability on the unknown parameters. This definition of feasibility does not utilize the concept of experimental repetitions in the frequentist sense as we have considered, but views the unknown parameters as random and considers the frequency of feasibility from the posterior belief, thus bypassing the set-level guarantee in (5.4). Indeed, Gupta (2019) shows that under suitable convexity assumption on the constraints (with respect to the unknown parameter) and discreteness of the underlying distribution, the size of the uncertainty set in DRO can be chosen lightly dependent on the problem dimension.

The second exceptional paradigm that we are aware of is scenario optimization (SO) (e.g., Calafiore and Campi (2005), Campi and Garatti (2008)), which applies to the case of CCP. In its basic form, this approach sets

$$\hat{\mathcal{F}} = \{x \in \mathcal{X} : G(x, \xi_i) \leq b \ \text{ for all } \ i = 1, \ldots, n\}$$

i.e., using sampled constraints formed from the data. As the number of constraints increases, $\hat{\mathcal{F}}$ is postulated to populate the decision space in some sense and ensure the obtained solution $\hat{x}^*$ lies in $\mathcal{F}$. While the sample size required in the basic SO is linear in the decision dimension d, recent works reduce this dependence by an array of generalizations, including using regularization (Campi and Carè (2013)), tighter support rank estimates (Schildbach et al. (2013), Campi and Garatti (2018)) and sequential and validation-based schemes (Carè et al. (2014), Calafiore (2017)).

The approach that we propose in this chapter aims to avoid using the set-level guarantee in (5.4) and the need to control its simultaneous estimation errors, which can cause over-conservativeness as discussed. Our approach operates under a frequentist framework, nonparametric assumptions

on the underlying distributions, and applies to all the exemplified methods mentioned above (SAA, RO, DRO and SO). It is thus different from the Bayesian parametric framework in Gupta (2019). Our idea is closest to some of the validation-type schemes suggested for SO, but more general as it applies to stochastic constraints beyond CCP and to data-driven reformulations beyond SO. Akin to these SO studies, our main results concern the power of our validation procedures in guaranteeing feasibility, which informs the required sample size in relation to the problem dimension. Our results also introduce a notion of optimality with respect to the chosen reformulation class, and deduce joint optimality-feasibility guarantees. In these regards, one main contribution of our work can be viewed as a rigorous construction of the first general-purpose validation framework for data-driven constrained optimization to systematically reduce conservativeness.

5.2 Overview of Our Framework and Rationale

Our framework, as discussed, aims to bypass the set-level guarantee in (5.4) and the need to control its simultaneous estimation errors. Our starting observation is the following. In all the described approaches above, the data-driven reformulation involves a key parameter that controls the level of conservativeness:

1. SAA: safety margin δ
2. RO and SCA: uncertainty set size ρ
3. DRO: divergence ball size or moment set size
4. SO: number of constraints

These parameters have the properties that setting it to one extreme (e.g., 0) would signal no uncertainty in the formulation, leading to a solution very likely infeasible, while setting it to another extreme (e.g., ∞) would cover the entire decision space, leading to a solution that is very conservative. In the established approaches, the parameter value is chosen to ensure (5.4), which tend to locate towards the latter extreme.

On the other hand, given a specific data-driven reformulation, it is easy to see that no matter how we choose this "conservativeness" parameter, the solution must lie in a low-dimensional manifold. More precisely, denote a given data-driven reformulation as

$$\min_{x\in\mathcal{X}} \; f(x) \;\; \text{subject to} \;\; x \in \hat{\mathcal{F}}(s) \tag{5.7}$$

where $s \in S$ denotes the conservativeness parameter, and we highlight the dependence of the data-driven feasible region $\hat{\mathcal{F}}(s)$ on s. We denote the obtained solution from (5.7) as $x^*(s)$. The *solution path* $\{x^*(s) : s \in S\}$ contains all possible obtainable solutions from the data-driven reformulation (5.7). Intuitively, any statement on feasibility suffices to focus on this solution path, instead of the whole decision space.

Nonetheless, besides the conservativeness parameter, a data-driven reformulation could have other parameters playing various roles (e.g., center and shape of an ellipsoidal uncertainty set in RO, baseline distribution in distance-based DRO etc.). The flexibility of these parameter values can enlarge the obtainable solution space and elevate its dimensionality. Suppose we want to contain this enlargement, and at the same time be able to select the optimal candidate within the low-dimensional manifold $\{x^*(s) : s \in S\}$. We propose the following two-phase framework to achieve this rigorously.

Our procedure (Algorithm 8) splits the data into two groups. With the first group of data, we construct a given data-driven reformulation parametrized by a conservativeness parameter s that varies over a space S, which we call $OPT(s)$. We obtain the optimal solution $x^*(s)$ for a range of values $s = s_j, j = 1, \ldots, p$. This step assumes the availability of an efficient solver for $OPT(s)$. Next, the second group of data is fed into a *validator* V that aims to identify the best feasible solution $x^*(\hat{s}^*)$ among $\{x^*(s_j) : j = 1, \ldots, p\}$. The number of points p required to validate depends on the size of S, which is constructed to be low-dimensional. There are multiple ways to set up the validator V, each with its own benefits and requirements. In the next two sections, we will introduce two classes of validators, one we call *Gaussian supremum validator* (Section 5.3), and another one we call *univariate Gaussian validator* (Section 5.4). We will present their rationales, theoretical

Algorithm 8 The Two-Phase Framework

Input: data $\boldsymbol{\xi}_{1:n} = \{\xi_1, \ldots, \xi_n\}$; numbers of data n_1, n_2 allocated to each phase ($n_1 + n_2 = n$); a confidence level $1 - \beta$; a given method to construct data-driven reformulation with a (possibly multi-dimensional) parameter $s \in S$; a discrete mesh $\{s_1, s_2, \ldots, s_p\} \subseteq S$.

Phase one:

1. Use n_1 observations, which we index as $\{\xi_{n_2+1}, \ldots, \xi_n\}$ for convenience, to construct the data-driven reformulation $OPT(s)$ in the form (5.7) parameterized by $s \in S$.

2. For each $j = 1, \ldots, p$, compute the optimal solution $x^*(s_j)$ of $OPT(s_j)$.

Phase two:

Use a validator V to select $(\hat{s}^*, x^*(\hat{s}^*)) = V(\{\xi_1, \ldots, \xi_{n_2}\}, \{x^*(s_1), \ldots, x^*(s_p)\}, 1 - \beta)$, where $x^*(\hat{s}^*)$ is a solution and $\hat{s}^*$ is the associated parameter value.

Output: $x^*(\hat{s}^*)$.

statistical guarantees, and implications on the feasibility and optimality of the obtained solution. Section 5.5 will then tie back the applicability of these validators to the exemplified approaches in Section 5.1.1.

5.3 Validation via Multivariate Gaussian Supremum

Our first validator uses a simultaneous estimation of $H(x)$ in the constraint in (5.1) to assess feasibility over the discretized solution path of $x^*(s)$. More precisely, given the solution set $\{x^*(s_j) : j = 1, \ldots, p\}$, we use a sample average with an appropriately calibrated safety margin, i.e., $\frac{1}{n_2}\sum_{i=1}^{n_2} h(x, \xi_i) - \epsilon$, to replace the unknown $H(\cdot)$ in (5.1) and output the best solution among the set. The margin ϵ is calibrated via the limiting distribution of $(\frac{1}{n_2}\sum_{i=1}^{n_2} h(x^*(s_j), \xi_i))_{j=1,\ldots,p}$ which captures the estimation error of $H(\cdot)$ and is multivariate Gaussian. It contains a critical value $q_{1-\beta}$ that is the quantile of a Gaussian supremum. Algorithms 9 and 10 describe two variants of this validator, one unnormalized while another one normalized by the standard deviation at each s_j. In the following, we denote $N_p(0, \Sigma)$ as a p-dimensional Gaussian vector with mean zero and

covariance Σ.

Algorithm 9 V: Unnormalized Gaussian Supremum Validator

Input: $\{\xi_1, \ldots, \xi_{n_2}\}, \{x^*(s_1), \ldots, x^*(s_p)\}, 1-\beta$

1. For each $j = 1, \ldots, p$ compute the sample mean $\hat{H}_j = (1/n_2)\sum_{i=1}^{n_2} h(x^*(s_j), \xi_i)$ and sample covariance matrix $\hat{\Sigma}$ with $\hat{\Sigma}(j_1, j_2) = (1/n_2)\sum_{i=1}^{n_2}(h(x^*(s_{j_1}), \xi_i) - \hat{H}_{j_1})(h(x^*(s_{j_2}), \xi_i) - \hat{H}_{j_2})$.

2. Compute $q_{1-\beta}$, the $(1-\beta)$-quantile of $\max\{Z_1, \ldots, Z_p\}$ where $(Z_1, \ldots, Z_p) \sim N_p(0, \hat{\Sigma})$, and let

$$\hat{s}^* = \operatorname{argmin}\left\{ f(x^*(s_j)) : \hat{H}_j \geq \gamma + \frac{q_{1-\beta}}{\sqrt{n_2}}, 1 \leq j \leq p \right\}. \tag{5.8}$$

Output: $\hat{s}^*, x^*(\hat{s}^*)$.

Algorithm 10 V: Normalized Gaussian Supremum Validator

Input: $\{\xi_1, \ldots, \xi_{n_2}\}, \{x^*(s_1), \ldots, x^*(s_p)\}, 1-\beta$

1. Same as in Algorithm 9.

2. Denote $\hat{\sigma}_j^2 = \hat{\Sigma}(j, j)$. Compute $q_{1-\beta}$, the $(1-\beta)$-quantile of $\max\{Z_j/\hat{\sigma}_j : \hat{\sigma}_j^2 > 0, 1 \leq j \leq p\}$ where $(Z_1, \ldots, Z_p) \sim N_p(0, \hat{\Sigma})$, and let

$$\hat{s}^* = \operatorname{argmin}\left\{ f(x^*(s_j)) : \hat{H}_j \geq \gamma + \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}}, 1 \leq j \leq p \right\}. \tag{5.9}$$

Output: $\hat{s}^*, x^*(\hat{s}^*)$.

The first Gaussian supremum validator (Algorithm 9) is reasoned from a joint central limit theorem (CLT) that governs the convergence of $\sqrt{n_2}(\hat{H}_1 - H(x^*(s_1)), \ldots, \hat{H}_p - H(x^*(s_p)))$ to $N_p(0, \Sigma)$, where $\Sigma(j_1, j_2) = \mathrm{Cov}_F(h(x^*(s_{j_1}), \xi), h(x^*(s_{j_2}), \xi))$. Using the sample covariance $\hat{\Sigma}$ from Step 1 of Algorithm 9 as an approximation of Σ, we have, by the continuous mapping theorem,

$$\max_{1 \leq j \leq p} \sqrt{n_2}(\hat{H}_j - H(x^*(s_j))) \approx \max_{1 \leq j \leq p} Z_j \text{ in distribution}$$

where $(Z_1, \ldots, Z_p) \sim N_p(0, \hat{\Sigma})$. Therefore using the $1-\beta$ quantile $q_{1-\beta}$ of the Gaussian supremum in the margin leads to

$$H(x^*(s_j)) \geq \hat{H}_j - \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p, \text{ with probability} \approx 1-\beta.$$

The second validator (Algorithm 10) uses an alternate version of the CLT that is normalized by the componentwise standard deviation σ_j, i.e., $\sqrt{n_2}((\hat{H}_1 - H(x^*(s_1)))/\sigma_1, \ldots, (\hat{H}_p - H(x^*(s_p)))/\sigma_p)$ converges to $N_p(0, D\Sigma D)$, where D is a diagonal matrix of $1/\sigma_j, j = 1, \ldots, p$. Note that the quantile $q_{1-\beta}$ in both validators can be computed to high accuracy via Monte Carlo.

Let us make the above reasoning precise. We present our results for two cases that need separate treatments: When $H(x) \geq \gamma$ is a "light-tailed" stochastic constraint, and when it is a chance constraint.

5.3.1 Performance Guarantees for General Stochastic Constraints

Recall that $H(x) = \mathbb{E}_F[h(x, \xi)]$. Denote

$$\sigma^2(x) := \mathrm{Var}_F(h(x, \xi))$$

as the variance of h for each decision $x \in \mathcal{X}$. We assume the following on optimization problem (5.1):

Assumption 5.3.1 (Light-tailedness) *There exists a constant $D_1 \geq 1$ such that for all $x \in \mathcal{X}$ with $\sigma^2(x) > 0$, we have*

$$\mathbb{E}_F\Big[\exp\Big(\frac{|h(x,\xi) - H(x)|^2}{D_1^2 \sigma^2(x)}\Big)\Big] \leq 2 \quad \textit{and} \quad \mathbb{E}_F\Big[\Big(\frac{|h(x,\xi) - H(x)|}{\sigma(x)}\Big)^{2+k}\Big] \leq D_1^k \textit{ for } k = 1, 2.$$

This assumption stipulates that the distribution of $h(x, \xi)$ after being centered and normalized by its standard deviation is sufficiently light-tailed at each x. Note that no other regularity property, e.g., convexity or continuity, is assumed for the function h itself. We have the following finite-sample feasibility guarantees for the solution output by Algorithm 9 or 10:

Theorem 5.3.1 (Finite-sample feasibility guarantee for unnormalized validator) *Denote* $\overline{H} = \max_{1\le j\le p} H(x^*(s_j))$ *and* $\bar{\sigma}^2 = \max_{1\le j\le p}\sigma^2(x^*(s_j))$. *Suppose Assumption 5.3.1 holds. For every solution set* $\{x^*(s_j) : 1 \le j \le p\}$, *every* n_2, *and* $\beta \in (0, \frac{1}{2})$, *the solution output by Algorithm 9 satisfies*

$$\begin{aligned} & P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \\ \ge\ & 1-\beta-C\left(\Big(\frac{D_1^2\log^7(pn_2)}{n_2}\Big)^{\frac{1}{6}} + \exp\big(-\frac{cn_2\epsilon^2}{D_1^2\bar{\sigma}^2}\big) + p\exp\big(-\frac{cn_2}{D_1^4}\big)\right) \end{aligned}$$

with

$$\epsilon = \left(\overline{H} - \gamma - C\bar{\sigma}\sqrt{\frac{\log(p/\beta)}{n_2}}\right)_+ \tag{5.10}$$

where C *and* c *are universal constants, and* $P_{\boldsymbol{\xi}_{1:n_2}}$ *denotes the probability with respect to Phase two data* $\{\xi_1,\ldots,\xi_{n_2}\}$ *and conditional on Phase one data* $\{\xi_{n_2+1},\ldots,\xi_n\}$.

Theorem 5.3.2 (Finite-sample feasibility guarantee for normalized validator) *Denote by* $\overline{s} \in \operatorname{argmax}\{H(x^*(s_j)) : j = 1,\ldots,p\}$ *the parameter giving rise to the highest constraint value, i.e.,* $H(x^*(\overline{s})) = \overline{H}$. *Under the same conditions of Theorem 5.3.1, the solution output by Algorithm 10 satisfies*

$$\begin{aligned} & P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \\ \ge\ & 1-\beta-C\left(\Big(\frac{D_1^2\log^7(pn_2)}{n_2}\Big)^{\frac{1}{6}} + \frac{D_1^2\log^2(pn_2)}{\sqrt{n}_2} + \exp\big(-\frac{cn_2\epsilon^2}{D_1^2\sigma^2(x^*(\overline{s}))}\big) + p\exp\big(-\frac{cn_2^{2/3}}{D_1^{10/3}}\big)\right) \end{aligned}$$

with

$$\epsilon = \left(\overline{H} - \gamma - C\sigma(x^*(\overline{s}))\sqrt{\frac{\log(p/\beta)}{n_2}}\right)_+ \tag{5.11}$$

where C *and* c *are universal constants.*

In both Theorems 5.3.1 and 5.3.2, the finite-sample coverage probability consists of two sources of errors. The first source comes from the CLT approximation that decays polynomially in the Phase 2 sample size n_2. The second error arises from the possibility that none of the solutions

$\{x^*(s_1), \ldots, x^*(s_p)\}$ satisfies the criterion in (5.8) or (5.9), which vanishes exponentially fast. When ϵ in (5.10) or (5.11) is of constant order, the CLT error dominates. In this case the finite-sample error depends logarithmically on p, the number of candidate parameter values, and the bounds dictate a coverage tending to $1 - \beta$ when p is as large as $\exp(o(n_2^{1/7}))$.

The derivation of the logarithmic dependence on p in Theorem 5.3.1 builds on a high-dimensional CLT and an associated multiplier bootstrap approximation recently developed in Chernozhukov et al. (2017) (Appendix D.2.1). The proof of Theorem 5.3.2 further requires a Hoeffding-type inequality for U-statistics to control the errors of the sample variance estimates, as well as the so-called Nazarov's inequality, an anti-concentration inequality for multivariate Gaussian, to control the coverage errors when using estimated standard deviations in the margin (Appendix D.2.2). Appendices D.2.3 and D.2.4 detail the proofs of Theorems 5.3.1 and 5.3.2 that put together the above mathematical developments.

We explain the implication on the dimensionality of the problem. Note that to sufficiently cover the whole solution path, p is typically exponential in the dimension of S, denoted $\dim(S)$ (this happens when we uniformly discretize the parameter space S). The discussion above thus implies a requirement that n_2 is of higher order than $\dim(S)^7$. Here the low dimensionality of S is crucial; for instance, a one-dimensional conservativeness parameter s would mean $\dim(S) = 1$, so that a reasonably small n_2 can already ensure adequate feasibility coverage. Moreover, the margin adjustments in Algorithms 9 and 10 both depend only on $\dim(S)$. Thus, choosing $\hat{s}^*$ relies only on $\dim(S)$, but not the dimension of the whole decision space. Note that Theorems 5.3.1 and 5.3.2 provide guarantee conditional on Phase one data. However, the universality of the involved constants means that analogous unconditional feasibility guarantees also hold if Assumption 5.3.1 can be verified uniformly or with high probability with respect to Phase one data, an observation that persists for other subsequent results.

Comparing between the two validators, we also see that the normalized one (Algorithm 10) is statistically more efficient than the unnormalized one (Algorithm 9) when the variance $\sigma^2(x)$ exhibits high variability across solutions. More specifically, in order to make the exponential error non-dominant, one needs at least $\epsilon > 0$. In the case of Algorithm 9, expression (5.10) suggests

that, after ignoring the logarithmic factor $\log(p/\beta)$, this requires an n_2 to be of order $\overline{\sigma}^2/(\overline{H}-\gamma)^2$. In contrast, for Algorithm 10 this becomes $\sigma^2(x^*(\overline{s}))/(\overline{H}-\gamma)^2$, where the maximum variance is replaced with the variance at the solution that optimizes the H-value, which in general does not have the maximum variance.

Theorems 5.3.4 and 5.3.5 also give immediately the following asymptotic feasibility guarantee (proof in Appendix D.2.4):

Corollary 5.3.3 (Asymptotic feasibility guarantee) *Let $\overline{H} = \max_{1\le j\le p} H(x^*(s_j))$. Suppose Assumption 5.3.1 holds. For every solution set $\{x^*(s_j) : 1 \le j \le p\}$ such that $\overline{H} > \gamma$ and every $\beta \in (0, \frac{1}{2})$, the solution output by Algorithm 9 or 10 satisfies*

$$\liminf_{n_2\to\infty \text{ and } p\exp(-n_2^{1/7})\to 0} P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \ge 1-\beta.$$

5.3.2 Performance Guarantees for Chance Constraints

Underlying the finite-sample bounds in Theorems 5.3.1 and 5.3.2 is the light-tailedness condition in Assumption 5.3.1. However, in a CCP that takes the form

$$\min_{x\in\mathcal{X}} f(x) \text{ subject to } P(x) := \mathbb{P}_F((x,\xi)\in A) \ge 1-\alpha \tag{5.12}$$

where $A \subseteq \mathbb{R}^d \times \mathbb{R}^m$ is a deterministic set and $1-\alpha$ is a tolerance level for the satisfaction probability, the tail of the normalized indicator function $\mathbf{1}((x,\xi)\in A)$ can be arbitrarily heavy as the satisfaction probability approaches 0 or 1 and hence violates Assumption 5.3.1. Thus, instead, we present different finite-sample error bounds for (5.12) than Theorems 5.3.1 and 5.3.2 whose derivations rely on the Bernoulli nature of the underlying function:

Theorem 5.3.4 (Finite-sample CCP feasibility guarantee for unnormalized validator) *Let $\bar{\alpha} = 1 - \max_{1\le j\le p} P(x^*(s_j))$. For every solution set $\{x^*(s_j) : 1 \le j \le p\}$, every n_2, and $\beta \in (0, \frac{1}{2})$, the solution output by Algorithm 9 satisfies*

$$P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.12)}) \ge 1-\beta-C\left(\left(\frac{\log^7(pn_2)}{n_2\alpha}\right)^{\frac{1}{6}} + \exp\left(-cn_2\min\{\epsilon, \frac{\epsilon^2}{\bar{\alpha}}\}\right)\right)$$

with

$$\epsilon = \left(\alpha - \bar{\alpha} - C\sqrt{\frac{\log(p/\beta)}{n_2}} \right)_+ \tag{5.13}$$

where C and c are universal constants.

Theorem 5.3.5 (Finite-sample CCP feasibility guarantee for the normalized validator) *Under the same conditions of Theorem 5.3.4, the solution output by Algorithm 10 satisfies*

$$\begin{aligned} & P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.12)}) \\ \geq \; & 1 - \beta - C\left(\left(\frac{\log^7(pn_2)}{n_2\alpha} \right)^{\frac{1}{6}} + \frac{\log^2(pn_2)}{\sqrt{n_2\alpha}} + \exp\big(- cn_2 \min\{\epsilon, \frac{\epsilon^2}{\bar{\alpha}}\} \big) \right) \end{aligned}$$

with

$$\epsilon = \left(\alpha - \bar{\alpha} - C\sqrt{\frac{\big(\bar{\alpha} + \log(n_2\alpha)/n_2\big)\log(p/\beta)}{n_2}} \right)_+ \tag{5.14}$$

where C and c are universal constants.

A comparison between Theorems 5.3.4 and 5.3.5 again reveals the higher statistical efficiency of Algorithm 10 than Algorithm 9 which, in the CCP context, applies to the case when the satisfaction probability is large (i.e., the common case). Suppose that $1 - \alpha$ approaches 1. In order to make $\epsilon > 0$ in (5.13), we need a sample size n_2 of order $(\alpha - \bar{\alpha})^{-2}$ (after ignoring the logarithmic factor $\log(p/\beta)$), whereas in (5.14) it can be seen to need only an n_2 of order $\alpha(\alpha - \bar{\alpha})^{-2}$, a much smaller size when $1 - \alpha$ is close to 1.

Lastly, we have the following asymptotic feasibility guarantee in the case of CCP in parallel to Corollary 5.3.3:

Corollary 5.3.6 (Asymptotic chance constraint feasibility guarantee) *For every solution set $\{x^*(s_j) : 1 \leq j \leq p\}$ such that $\alpha > 1 - \max_{1\leq j\leq p} P(x^*(s_j))$ and every $\beta \in (0, \frac{1}{2})$, the solution output by Algorithm 9 or 10 satisfies*

$$\liminf_{n_2\to\infty \text{ and } p\exp(-n_2^{1/7})\to 0} P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.12)}) \geq 1 - \beta.$$

Appendix D.2.4 details the proofs of Theorem 5.3.4, Theorem 5.3.5 and Corollary 5.3.6.

To close this section, we note that our Gaussian supremum validators also enjoy a notion of asymptotic solution-path optimality under additional assumptions. To streamline our presentation, we defer this discussion to the next section and combine it with the discussion of our next validator.

5.4 Validation via Univariate Gaussian Margin

We offer an alternate validator that can perform more efficiently than Algorithms 9 and 10, provided that further regularity assumptions are in place. This is a scheme that simply uses a standard univariate Gaussian critical value to calibrate the margin (Algorithm 11).

Algorithm 11 outputs a solution with objective value no worse than Algorithms 9 and 10. Comparing the criteria to choose $\hat{s}^*$, we see that, due to the stochastic dominance of the maximum among a multivariate Gaussian vector over each of its individual components, the margin in (5.8) satisfies $q_{1-\beta} \geq z_{1-\beta}\hat{\sigma}_j$ for all j, and similarly the margin in (5.9) satisfies $q_{1-\beta}\hat{\sigma}_j \geq z_{1-\beta}\hat{\sigma}_j$, so that both are bounded from below by the margin in (5.15). Consequently the solution from (5.15) achieves an objective value no worse than the other two.

Algorithm 11 V: Univariate Gaussian Validator

Input: $\{\xi_1, \ldots, \xi_{n_2}\}, \{x^*(s_1), \ldots, x^*(s_p)\}, 1-\beta$

1. For each $j = 1, \ldots, p$ compute the sample mean $\hat{H}_j = (1/n_2)\sum_{i=1}^{n_2} h(x^*(s_j), \xi_i)$ and sample variance $\hat{\sigma}_j^2 = (1/n_2)\sum_{i=1}^{n_2}(h(x^*(s_j), \xi_i) - \hat{H}_j)^2$.

2. Compute

$$\hat{s}^* = \operatorname{argmin}\left\{ f(x^*(s_j)) \middle| \hat{H}_j \geq \gamma + \frac{z_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}}, 1 \leq j \leq p \right\} \tag{5.15}$$

where $z_{1-\beta}$ is the $1-\beta$ quantile of the standard Gaussian distribution.

Output: $\hat{s}^*, x^*(\hat{s}^*)$.

The univariate Gaussian critical value used in the margin in Algorithm 11 hints that feasibility needs to be validated at only one value of s instead of the solution path S. The validity of this procedure is based on the statistical consistency of the obtained solution $x^*(\hat{s}^*)$ to some limiting

solution (correspondingly $\hat{s}^*$ to some limiting optimal parameter value) as n_2 increases. Intuitively, this implies that with sufficient sample size one can focus feasibility validation on a small neighborhood of $\hat{s}^*$, which further suggests that we need to control only the statistical error at effectively one solution parametrized at $\hat{s}^*$. For this argument to hold, however, we would need several additional technical assumptions including a low functional complexity of h, and a different line of derivations.

5.4.1 Asymptotic Performance Guarantees

We present the statistical guarantees of Algorithm 11 as Phase two data size $n_2 \to \infty$. We assume continuity for the objective of (5.1):

Assumption 5.4.1 (Continuous objective) *The objective function $f(x)$ is continuous on $\mathcal{X}$.*

For the constraint, we assume the following:

Assumption 5.4.2 (Functional complexity) *The function class $\mathcal{F} := \{h(x,\cdot)|x \in \mathcal{X}\}$ is F-Donsker.*

Assumption 5.4.3 (L_2-boundedness) $\mathbb{E}_F\big[\sup_{x\in\mathcal{X}} |h(x,\xi) - H(x)|^2\big] < \infty$.

Assumption 5.4.4 (L_2-continuity) *For every fixed $x \in \mathcal{X}$ and another $x' \in \mathcal{X}$, we have that $\lim_{x'\to x} \mathbb{E}_F[(h(x',\xi) - h(x,\xi))^2] = 0$.*

To give a sense of the generality of the above assumptions, we identify two general classes of constraints for which these assumptions are guaranteed to hold, one suitable for general h, and another one for CCPs:

Proposition 5.4.1 *Assumptions 5.4.2-5.4.4 hold in each of the following two cases:*

i. There exists some $M(\xi)$ such that $\mathbb{E}_F[M(\xi)^2] < \infty$ and $|h(x_1,\xi) - h(x_2,\xi)| \le M(\xi)\,\|x_1 - x_2\|$ for all $x_1, x_2 \in \mathcal{X}$, there exists some $\tilde{x} \in \mathcal{X}$ such that $\mathbb{E}_F[h(\tilde{x},\xi)^2] < \infty$, and the decision space $\mathcal{X}$ is compact;

ii. $h(x,\xi) = \mathbf{1}(a_k' A_k(x) \le b_k \text{ for } k = 1,\dots,K)$ *for some* $K < \infty$*, where each* $A_k(\cdot) : \mathbb{R}^d \to \mathbb{R}^{m_k}$ *is a continuous mapping and each* $a_k \in \mathbb{R}^{m_k}, b_k \in \mathbb{R}$ *satisfies either (i)* a_k *has a density on* $\mathbb{R}^{m_k}$ *and* b_k *is a non-zero constant or (ii)* (a_k, b_k) *has a density on* $\mathbb{R}^{m_k+1}$*.*

Case (i) in Proposition 5.4.1 follows from standard results in empirical process theory, including in particular the Jain-Marcus Theorem. The proof of Case (ii) involves checking the finite Vapnik-Chervonenkis (VC) dimension and pointwise separability of the function class in order to verify F-Donskerness. Appendix D.3 details the proof of Proposition 5.4.1.

We impose one more assumption on the constraint function regarding its variance:

Assumption 5.4.5 (Non-degeneracy of the variance on the boundary) $\sigma^2(x) > 0$ *for all* $x \in \mathcal{X}$ *such that* $H(x) = \gamma$*.*

In Assumption 5.4.5, non-zero variance is assumed only for those x's at which the stochastic constraint is satisfied with equality, but not necessarily for other x. This is significant in the case of CCP (5.12). While there could exist 100% or 0% safe solutions, i.e., x such that $P(x) = 1$ or 0, and hence non-degeneracy may not be satisfied over the whole $\mathcal{X}$, it holds for those x's with $P(x) = 1 - \alpha$ that have (the same) non-zero variance $\alpha(1-\alpha)$.

Now we present our assumptions on the data-driven reformulation $OPT(s), s \in S$. We focus on formulations with a single parameter (A separate set of results for formulations with multiple parameters can be found in Appendix D.6). We first assume that the solution path is piecewise continuous:

Assumption 5.4.6 (Piecewise continuous solution path) *The parameter space* S *is a finite interval* $[s_l, s_u]$*. The optimal solution* $x^*(s)$ *of* $OPT(s)$ *exists and is unique except for a finite number of parameter values* $\tilde{s}_i, i = 1,\dots,M-1$ *such that* $s_l = \tilde{s}_0 < \tilde{s}_1 < \cdots < \tilde{s}_{M-1} < \tilde{s}_M = s_u$*, and the parameter-to-solution mapping* $x^*(s)$ *is uniformly continuous on each piece* $[\tilde{s}_0, \tilde{s}_1)$*,* $(\tilde{s}_{M-1}, \tilde{s}_M]$*, and* $(\tilde{s}_{i-1}, \tilde{s}_i)$ *for* $i = 2,\dots,M-1$*.*

Continuity of the solution path allows approximating the whole solution curve by discretizing the parameter space S. Also note that under Assumption 5.4.6 the solution $x^*(s)$ exists and is unique

for almost surely every $s \in S$ with respect to the Lebesgue measure. Therefore, if one discretizes the parameter space by randomizing via a continuous distribution over S, then with probability one the solution $x^*(s)$ is unique at all sampled parameter values. This provides an easy way to ensure the assumption that none of the parameter values $\{s_1, \ldots, s_p\}$ used in Phase one of Algorithm 8 belongs to the discontinuity set $\{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$.

To explain the superior performance of Algorithm 11, we introduce a notion of optimality within the solution path $\{x^*(s) : s \in S\}$. First, since the parameter-to-solution mapping $x^*(s)$ is not defined at the discontinuities under Assumption 5.4.6, we need to fill in these holes in the solution path. Thanks to uniform continuity, the mapping $x^*(s)$ on each piece $(\tilde{s}_{i-1}, \tilde{s}_i)$ can be continuously extended to the closure $[\tilde{s}_{i-1}, \tilde{s}_i]$ by taking left and right limits. Specifically, we define:

Definition 5.4.1 *Under Assumption 5.4.6, the parameter-to-solution mapping $x^*(\cdot)$ at each discontinuity $\tilde{s}_i, i = 1, \ldots, M-1$ is defined in an extended fashion as*

$$x^*(\tilde{s}_i) = \{x^*(\tilde{s}_i-), x^*(\tilde{s}_i+)\} \text{ where } x^*(\tilde{s}_i-) := \lim_{s \to \tilde{s}_i-} x^*(s) \text{ and } x^*(\tilde{s}_i+) := \lim_{s \to \tilde{s}_i+} x^*(s).$$

Note that the two solutions $x^*(\tilde{s}_i-)$ and $x^*(\tilde{s}_i+)$ are different if the i-th and $(i+1)$-th pieces are disconnected. With the extended parameter-to-solution mapping $x^*(\cdot)$, we now introduce the notions of optimal solution and optimal parameter associated with the solution path:

Definition 5.4.2 *Associated with the solution path $\{x^*(s) : s \in S\}$, the optimal solution set is*

$$\mathcal{X}_S^* := \operatorname*{argmin}_x \left\{ f(x) : \begin{array}{l} H(x) \geq \gamma, \text{ where} \\ x = x^*(s) \text{ for some } s \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\} \text{ or} \\ x \in x^*(\tilde{s}_i) \text{ for some } i = 1, \ldots, M-1 \end{array} \right\} \tag{5.16}$$

and the optimal parameter set is

$$S^* := \{s \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\} : x^*(s) \in \mathcal{X}_S^*\} \cup \{\tilde{s}_i : x^*(\tilde{s}_i) \cap \mathcal{X}_S^* \neq \emptyset, i = 1, \ldots, M-1\}. \tag{5.17}$$

We need several additional technical assumptions. The first is that the stochastic constraint is not

binding at the endpoints of each piece of the solution path:

Assumption 5.4.7 *$H(x^*(\tilde{s}_i-)) \neq \gamma$ and $H(x^*(\tilde{s}_i+)) \neq \gamma$ for all $i = 1, \ldots, M-1$, $H(x^*(s_l)) \neq \gamma$, $H(x^*(s_u)) \neq \gamma$, and $\sup_{s \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}} H(x^*(s)) > \gamma$.*

Since the solution path $\{x^*(s) : s \in S\}$ depends on Phase one data $\boldsymbol{\xi}_{n_2+1:n}$, the path and hence the endpoints $x^*(\tilde{s}_i-), x^*(\tilde{s}_i+)$ are random objects, and so the first part of Assumption 5.4.7 is expected to hold almost surely provided that the set $\{x \in \mathcal{X} : H(x) = \gamma\}$ is a null set under the Lebesgue measure on $\mathbb{R}^d$. The second part states that the solution path contains a strictly feasible solution which in turn ensures that the optimal solution set $\mathcal{X}_S^*$ is non-empty. Note that this can typically be achieved by simply including very conservative parameter values in S.

Another property we assume regards the monotonicity of the feasible set size with respect to the parameter s in the reformulation $OPT(s)$:

Assumption 5.4.8 *Denote by $\mathrm{Sol}(s) := \mathcal{X} \cap \hat{\mathcal{F}}(s)$ the feasible set of $OPT(s)$. Assume $\mathrm{Sol}(s)$ is a closed set for all $s \in S$ and $\mathrm{Sol}(s_2) \subseteq \mathrm{Sol}(s_1)$ for all $s_1, s_2 \in S$ such that $s_1 < s_2$.*

Assmption 5.4.8 holds for all common reformulations (all examples in the beginning of Section 5.2) as s controls the conservativeness level. For instance, in RO with ellipsoidal uncertainty set, the RO feasible region shrinks with the radius of the ellipsoid, and similar relations hold for DRO, SAA, and SO. A straightforward consequence of Assumption 5.4.8 is the monotonicity of the parameter-to-objective mapping

$$v(s) := \inf\{f(x) : x \in \mathcal{X} \cup \hat{\mathcal{F}}(s)\}$$

as described in the following proposition (proof in Appendix D.3):

Proposition 5.4.2 *Suppose Assumptions 5.4.6 and 5.4.8 hold. For all $s_1, s_2 \in S$ such that $s_1 < s_2$ it holds $v(s_1) \leq v(s_2)$, and if additionally $s_1, s_2 \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ then $v(s_1) < v(s_2)$ if and only if $x^*(s_1) \neq x^*(s_2)$.*

The assumptions we have made for the formulation $OPT(s)$ give rise to the following uniqueness characterization of the optimal solution set $\mathcal{X}_S^*$ and the optimal parameter set S^* within the solution path, which would be used to establish the feasibility guarantees for Algorithm 11.

Proposition 5.4.3 *Under Assumptions 5.4.1, 5.4.4, and 5.4.6-5.4.8, the optimal solution set $\mathcal{X}_S^*$ is a singleton $\{x_S^*\}$ and the optimal parameter set S^* is a closed interval $[s_l^*, s_u^*]$ for $s_l^*, s_u^* \in S$. In addition, if $v(s)$ is strictly increasing on S, then S^* is a singleton $\{s^*\}$.*

The proof of Proposition 5.4.3, which is in Appendix D.3, involves an exhaustion of all possible structures of the set $\mathcal{X}_S^*$ that contain more than one solution, and showing each of them contradicts with our assumptions (especially Assumption 5.4.8).

Lastly, we assume the following technical assumption for the set of optima:

Assumption 5.4.9 *For any $\epsilon > 0$ there exists an $s \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ such that $H(x^*(s)) > \gamma$ and $\|x^*(s) - x_S^*\|_2 < \epsilon$, where x_S^* is the unique optimal solution from Proposition 5.4.3.*

This assumption trivially holds if $\mathcal{X}_S^* = \{x_S^*\}$ as described in Proposition 5.4.3 and $H(x_S^*) > \gamma$. Otherwise, if $H(x_S^*) = \gamma$, it rules out the case that the solution path $x^*(s)$ passes through x_S^* without entering the interior of the feasible set of (5.1). The latter exceptional case typically happens with zero probability, in view of the fact that the solution path is itself random with respect to Phase one data.

Now we are ready to present the asymptotic performance guarantee for Algorithm 11:

Theorem 5.4.4 (Asymptotic joint feasibility-optimality guarantee) *Suppose Assumptions 5.4.1-5.4.5 hold for* (5.1)*. Also suppose Assumptions 5.4.6-5.4.9 hold for the reformulation $OPT(s)$ constructed in Algorithm 8, and $\{s_1, \ldots, s_p\} \cap \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\} = \emptyset$. Let $\epsilon_S = \sup_{s \in S} \inf_{1 \le j \le p} |s - s_j|$ be the mesh size, and x_S^* be the unique optimal solution from Proposition 5.4.3. Then, with respect to $\{\xi_1, \ldots, \xi_{n_2}\}$, the solution and parameter output by Algorithm 11 satisfy*

$$\lim_{n_2 \to \infty, \epsilon_S \to 0} x^*(\hat{s}^*) = x_S^* \text{ and } \lim_{n_2 \to \infty, \epsilon_S \to 0} d(\hat{s}^*, S^*) = 0 \tag{5.18}$$

almost surely. Moreover, if $H(x_S^*) = \gamma$ we have

$$\liminf_{n_2 \to \infty, \epsilon_S \to 0} P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for } (5.1)) \ge 1 - \beta, \tag{5.19}$$

otherwise if $H(x_S^*) > \gamma$ *we have*

$$\lim_{n_2 \to \infty, \epsilon_S \to 0} P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for } (5.1)) = 1. \tag{5.20}$$

Theorem 5.4.4 states that as the mesh $\{s_1, \ldots, s_p\}$ gets increasingly fine and the data size grows, the solution given by Algorithm 11 enjoys performance guarantees concerning both feasibility and solution-path optimality. In particular, the estimated solution and the conservativeness parameter converge to the optimal solution x_S^* and the optimal parameter set S^* respectively, while simultaneously the obtained solution is feasible with the desired confidence level $1 - \beta$.

The proof of Theorem 5.4.4 is in Appendix D.3. The consistency result in (5.18) is shown via a dense approximation of the discrete parameter set $\{s_1, \ldots, s_p\}$ on the continuum S, through the continuity of the solution path and a uniform law of large numbers. Then, based on this consistency, the feasibility guarantee (5.19) is established by showing $P\big(H(x^*(\hat{s}^*)) \geq \gamma\big) \geq P\big(H(x^*(\hat{s}^*)) \geq \hat{H}(x^*(\hat{s}^*)) - z_{1-\beta}\hat{\sigma}(x^*(\hat{s}^*))/\sqrt{n_2}\big) \approx P\big(H(x_S^*) \geq \hat{H}(x_S^*) - z_{1-\beta}\hat{\sigma}(x_S^*)/\sqrt{n_2}\big) \to 1 - \beta$, where the "$\geq$" follows from our validation criterion (5.15) whereas the "$\approx$" comes from the asymptotic tightness of the empirical process $\{\sqrt{n_2}(\hat{H}(x^*(s)) - H(x^*(s))) : s \in S\}$ and the L_2 continuity of the constraint function $h(x, \xi)$.

Furthermore, under additional smoothness conditions on the constraint function h and the solution path $\{x^*(s) : s \in S\}$, we also establish the finite-sample counterparts for the optimality guarantee (5.18) and feasibility guarantee (5.19) for Algorithm 11. These are presented in Appendix D.5.

Note that the confidence level (5.19) at which Algorithm 11 outputs a feasible solution (and also Algorithms 9 and 10, i.e., Corollaries 5.3.3 and 5.3.6) is generally not tight, i.e., a lower bound instead of an equality is guaranteed. However, with a strict monotonicity condition on the reformulation $OPT(s)$ and a finer discretization mesh for the conservativeness parameter, Algorithm 11 can give a tight confidence guarantee:

Theorem 5.4.5 (Asymptotically tight feasibility guarantee) *In addition to the conditions of Theorem 5.4.4, further assume that the parameter-to-objective mapping* $v(s)$ *is strictly increasing*

on S, and consider the case that $H(x^*_S) = \gamma$. If the mesh $\{s_1, \ldots, s_p\}$ is fine enough so that

$$\max_{i=1,\ldots,M} \max_{j=1,\ldots,p_i-1} \left|H(x^*(s^i_j)) - H(x^*(s^i_{j+1}))\right| = o\big(\frac{1}{\sqrt{n_2}}\big) \tag{5.21}$$

where $s^i_1 < \cdots < s^i_{p_i}$ are the parameter values $\{s_j : s_j \in (\tilde{s}_{i-1}, \tilde{s}_i), j = 1, \ldots, p\}$ (so that $\sum_{i=1}^M p_i = p$), then we must have

$$\lim_{n_2 \to \infty \text{ and } \epsilon_S \to 0 \text{ s.t. } (5.21) \text{ holds}} P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for } (5.1)) = 1 - \beta$$

for the solution output by Algorithm 11.

Roughly speaking, the loose confidence guarantee in (5.19) can be attributed to the one-sided nature of the inequality criterion used in (5.15). The monotonicity of $v(s)$ and the mesh condition (5.21) give rise to a tight confidence guarantee by strengthening this inequality criterion to an equality (with a negligible error) at the chosen parameter value $\hat{s}^*$. Note that, when the expected constraint value $H(x^*(s))$ is Lipschitz continuous in the parameter, the mesh condition (5.21) is guaranteed if $\epsilon_S = o\big(\frac{1}{\sqrt{n_2}}\big)$ or if $\frac{p}{\sqrt{n_2}} \to \infty$ and the mesh is equispaced. The proof of Theorem 5.4.5 is in Appendix D.3.

Relatedly, the following shows that, like Algorithm 11, the supremum-based validators in Algorithms 9 and 10 also exhibit joint asymptotic feasibility and solution-path optimality guarantees. However, their confidence guarantees for feasibility are not as tight. This result complements our discussions at the end of Section 5.3 regarding the optimality property of the supremum-based validators, and also at beginning of Section 5.4 regarding the better objective value of the solution obtained by Algorithm 11, which is consistent with its tighter achievement of the feasibility confidence level.

Theorem 5.4.6 (Asymptotic joint guarantee for Gaussian supremum validators) *Under the same conditions as Theorem 5.4.4, the solution and parameter from Algorithm 9 satisfy the con-*

sistency guarantee (5.18). *In the case* $H(x_S^*) = \gamma$ *it holds*

$$\liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \geq \Phi\big(\frac{\bar{q}_{1-\beta}}{\sigma(x_S^*)}\big) \geq 1-\beta \tag{5.22}$$

where $\bar{q}_{1-\beta}$ *is the* $1-\beta$ *quantile of the supremum of the Gaussian process indexed by the parameter* $s \in S\backslash\{\tilde{s}_1,\ldots,\tilde{s}_{M-1}\}$ *with the covariance structure* $\mathrm{Cov}(s,s') = \mathrm{Cov}_F(h(x^*(s),\xi), h(x^*(s'),\xi))$, *and* Φ *is the distribution function of the standard normal.*

If it is further assumed that $\inf_{x\in\mathcal{X}} \sigma^2(x) > 0$, *then* (5.18) *also holds for Algorithm 10, and in the case* $H(x_S^*) = \gamma$ *we have*

$$\liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \geq \Phi(\tilde{q}_{1-\beta}) \geq 1-\beta \tag{5.23}$$

where $\tilde{q}_{1-\beta}$ *is the* $1-\beta$ *quantile of the supremum of the Gaussian process on* $S\backslash\{\tilde{s}_1,\ldots,\tilde{s}_{M-1}\}$ *with covariance* $\mathrm{Cov}(s,s') = \mathrm{Cov}_F(h(x^*(s),\xi), h(x^*(s'),\xi))/(\sigma(x^*(s))\sigma(x^*(s')))$.

In general, when the Gaussian processes involved in (5.22) and (5.23) have non-constant covariance structures, the asymptotic confidence levels rendered by Algorithms 9 and 10 are strictly higher than the nominal level $1-\beta$. This suggests that supremum-based margins tend to generate more conservative solutions than the univariate Gaussian margin does, although they all approach the same optimal solution x_S^* in the limit.

The proof of Theorem 5.4.6 (in Appendix D.3) involves steps similar to that of Theorem 5.4.4, but furthermore showing the statistical consistency of the critical value $q_{1-\beta}$ calibrated in Algorithm 9 or 10. The latter utilizes the separability of the limiting Gaussian process and a control of errors coming from the associated multiplier bootstrap approximation.

5.5 Applying Our Framework in Data-Driven Reformulations

In this section we showcase various data-driven reformulations of (5.1) or (5.12) to which our proposed framework can be applied. We first comment that our Gaussian supremum validators (Algorithms 9 and 10) are applicable to all formulations considered here, as long as the constraint

function $h(x, \xi)$ is sufficiently light-tailed as described in Assumption 5.3.1 or the constraint is a chance constraint. That is,

Theorem 5.5.1 (Applicability of Gaussian supremum validators) *Consider the stochastically constrained problem* (5.1) *that satisfies Assumption 5.3.1, or CCP* (5.12)*. All the data-driven reformulations $OPT(s)$ presented below, namely SAA, DRO with ϕ-divergence, Wasserstein and moment-based uncertainty sets, RO with polyhedral and ellipsoidal uncertainty sets, and SO (the last two approaches are for CCP only), can be validated by the Gaussian supremum validators in Algorithms 9 and 10 and elicit the conclusions in all theorems and corollaries in Section 5.3.*

The tighter univariate Gaussian validator (Algorithm 11) however requires some extra regularity conditions from the data-driven formulation $OPT(s)$, but still works for many common formulations. We consider decision space $\mathcal{X}$ that has the form:

Assumption 5.5.1 $\mathcal{X} = \{x \in \mathbb{R}^d : f_r(x) \leq 0 \text{ for } r = 1, \ldots, R \text{ and } Wx \leq z\}$, *where each f_r is continuous and convex, and $W = [w_1, w_2, \ldots, w_L]' \in \mathbb{R}^{L \times d}, z \in \mathbb{R}^L$.*

We consider optimization formulations that satisfy the following two assumptions:

Assumption 5.5.2 (Slater's condition) *Slater's condition holds for $OPT(s_u) := \min\{f(x) : x \in \mathcal{X} \cap \hat{\mathcal{F}}(s_u)\}$ where s_u is the maximum parameter value.*

Assumption 5.5.3 (Non-empty and bounded level set) *There exists a constant c such that $\mathcal{X} \cap \hat{\mathcal{F}}(s_l) \cap \{x : f(x) \leq c\}$ is bounded and $\mathcal{X} \cap \hat{\mathcal{F}}(s_u) \cap \{x : f(x) \leq c\}$ is non-empty where s_l, s_u are the minimum and maximum parameter values.*

Slater's condition (Assumption 5.5.2) is a common property that is expected to hold for most optimization problems in practice. Dür et al. (2016) states that Slater's condition is a generic property for linear conic programs by showing that it holds for all problem data except in a null set of Lebesgue measure. Assumption 5.5.3 also trivially holds in many settings, e.g., when $\mathcal{X} \cap \hat{\mathcal{F}}(s_l)$ is compact or $f(x)$ is coercive. Under these two assumptions, stability results from parametric optimization (Proposition 4.4 in Bonnans and Shapiro (2013)) ensure that the solution path $x^*(s)$

is continuous when the optimal solution is unique for each $OPT(s)$, or piecewise continuous when uniqueness fails at only a finite number of parameter values, leading to Assumption 5.4.6. Since other assumptions from Section 5.4.1 regarding $OPT(s)$ can be readily verified to hold in general, for each considered formulation below we focus on identifying the conditions that guarantee the validity of Assumption 5.4.6 in order to ensure the asymptotic feasibility and optimality guarantees. The proofs of all results in this section are presented in Appendix D.4.

We introduce a condition that will appear in the following discussion. Consider the linear objective $f(x) = c'x$ for some deterministic $c \in \mathbb{R}^d$. We say a finite collection of vectors $\{v_1, \ldots, v_k\} \subset \mathbb{R}^d$ with $k \leq d-1$ satisfies the strict cone inclusion (SCI) condition if

$$\text{SCI: } v_1, \ldots, v_k \text{ are linearly independent, and there exist } \lambda_1, \ldots, \lambda_k > 0 \text{ such that } \sum_{i=1}^{k} \lambda_i v_i = -c.$$

SAA: First consider the SAA reformulation for the general stochastic constraint in (5.1) in the form

$$\min_{x \in \mathcal{X}} \ f(x) \ \text{ subject to } \ \frac{1}{n} \sum_{i=1}^{n} h(x, \xi_i) \geq \gamma + s \tag{5.24}$$

where $s > 0$ is the margin to be tuned (and for convenience, in this section only, we use n to represent a generic sample size; in applications this typically refers to the Phase one data size). We have the following result concerning the applicability of Algorithm 11:

Theorem 5.5.2 (Applying univariate Gaussian validator to SAA) *Consider the reformulation $OPT(s)$ using* (5.24)*. Suppose Assumptions 5.5.1-5.5.3 hold. In either of the following two cases:*

i. $f(x)$ is continuous and strictly convex, $h(x, \xi)$ is continuous and concave in x for every ξ.

ii. $f(x) = c'x$ for some non-zero $c \in \mathbb{R}^d$, the functions $f_r, r = 1, \ldots, R$ in Assumption 5.5.1 are strictly convex, any $k \leq d-1$ rows $\{w_{l(1)}, \ldots, w_{l(k)}\}$ of W do not satisfy the SCI condition, and $h(x, \xi)$ either is continuous and strictly concave in x for every ξ or has the form $h(x, \xi) = A(\xi)'x + b(\xi)$ where $A(\xi) \in \mathbb{R}^d$ has a density on $\mathbb{R}^d$.

Assumption 5.4.6 holds with $M = 1$ almost surely in the data $\{\xi_1, \ldots, \xi_n\}$.

The proof of Theorem 5.5.2 (and theorems for other formulations below) mainly consists of establishing the joint continuity of the data-driven constraint (5.24) in x and s, and the uniqueness of $x^*(s)$, two main ingredients that enable us to apply the stability theory from Bonnans and Shapiro (2013) to conclude the continuity of $x^*(s)$. The former is shown by direct verification, whereas the latter is established from either strict convexity or the SCI condition when the formulation has linear objectives and constraints.

In the case of chance constraint (5.12), the SAA formulation has the form

$$\min_{x\in\mathcal{X}} \; f(x) \;\; \text{subject to} \;\; \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}((x,\xi_i)\in A) \geq 1-\alpha+s. \tag{5.25}$$

Note that the left hand side can only take values $\frac{j}{n}, j = 0, 1, \ldots, n$, therefore all s such that $1-\alpha+s \in (\frac{j-1}{n}, \frac{j}{n}]$ lead to the same feasible region and hence the same solution $x^*(s)$. As a result, the solution path $\{x^*(s) : s \in S\}$ consists of at most n constant pieces and Assumption 5.4.6 holds automatically. Thus we have:

Theorem 5.5.3 (Applying univariate Gaussian validator to SAA for CCP) *Consider the reformulation $OPT(s)$ using (5.25). Assumption 5.4.6 holds for some $M \leq n$.*

ϕ-divergence DRO: Given a convex function ϕ on $[0, +\infty)$ such that $\phi(1) = 0$, consider the ϕ-divergence DRO formulation for (5.1) in the form

$$\min_{x\in\mathcal{X}} \; f(x) \;\; \text{subject to} \;\; \inf\Big\{\sum_{i=1}^{n} w_i h(x,\xi_i) : \sum_{i=1}^{n}\frac{1}{n}\phi(nw_i) \leq s, \sum_{i=1}^{n} w_i = 1, w_i \geq 0 \text{ for all } i\Big\} \geq \gamma. \tag{5.26}$$

We have the following result:

Theorem 5.5.4 (Applying univariate Gaussian validator to ϕ-divergence-based DRO) *Consider $OPT(s)$ using (5.26). Suppose Assumptions 5.5.1-5.5.3 hold, and ϕ is continuous and convex on $(0, +\infty)$ with $\phi(1) = 0$. In either of the following three cases:*

i. $f(x)$ is continuous and strictly convex, $h(x, \xi)$ is continuous and concave in x for every ξ.

ii. $f(x) = c'x$ *for some non-zero* $c \in \mathbb{R}^d$*, the functions* $f_r, r = 1, \ldots, R$ *in Assumption 5.5.1 are strictly convex, any* $k \leq d-1$ *rows* $\{w_{l(1)}, \ldots, w_{l(k)}\}$ *of* W *do not satisfy the SCI condition, and* $h(x, \xi)$ *is continuous and strictly concave in* x *for every* ξ*.*

iii. Assume the same conditions as in (ii) except that $h(x, \xi)$ *is only concave (instead of strictly concave) in* x *for every* ξ*. In addition,* ϕ *is differentiable and strictly convex on* $(0, +\infty)$ *with* $\lim_{x \to 0+} \phi(x) = +\infty$*. Let* $\widehat{\text{Corr}}(x_1, x_2) = \widehat{\text{Cov}}(h(x_1, \xi), h(x_2, \xi))/(\hat{\sigma}(x_1)\hat{\sigma}(x_2))$ *be the empirical correlation coefficient between* $h(x_1, \xi)$ *and* $h(x_2, \xi)$ *based on data* $\{\xi_1, \ldots, \xi_n\}$ *for any* $x_1, x_2 \in \mathcal{X}$*.* $\hat{\sigma}^2(x) > 0$ *for all* $x \in \mathcal{X}$*, and there exist no distinct* x_1, x_2 *such that* $\widehat{\text{Corr}}(\lambda x_1 + (1-\lambda)x_2, x_1) = 1$ *for all* $\lambda \in [0, 1]$*.*

Assumption 5.4.6 holds with $M = 1$ *for the* ϕ*-divergence DRO conditioned on the data* $\{\xi_1, \ldots, \xi_n\}$*.*

Wasserstein DRO: Consider the Wasserstein DRO reformulation for the constraint in (5.1) given by

$$\min_{x \in \mathcal{X}} \; f(x) \;\; \text{subject to} \;\; \inf \Big\{ \mathbb{E}_G[h(x, \xi)] : d_p(G, F_n) \leq s \Big\} \geq \gamma \tag{5.27}$$

where $F_n = \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$ is the empirical distribution and $d_p(G, F_n)$ is the Wasserstein distance between an arbitrary probability measure G and F_n which is defined as

$$d_p^p(G, F_n) = \inf \Big\{ \mathbb{E}_\pi[\|\xi - \xi'\|^p] : \pi \text{ is a probability measure on } \Xi^2 \text{ with marginals } G \text{ and } F_n \Big\}$$

where Ξ is the known domain of ξ and $\|\cdot\|$ is an arbitrary norm.

The following theorem gives conditions under which Wasserstein DRO satisfies Assumption 5.4.6:

Theorem 5.5.5 (Applying univariate Gaussian validator to Wasserstein-metric DRO) *Consider* $OPT(s)$ *using* (5.27)*. Suppose Assumptions 5.5.1-5.5.3 hold, the domain* Ξ *of* ξ *is compact, and* $1 \leq p < \infty$*. In either of the following two cases:*

i. $f(x)$ *is continuous and strictly convex,* $h(x, \xi)$ *is jointly continuous in* x, ξ *and concave in* x *for every* ξ*.*

ii. $f(x) = c'x$ *for some non-zero* $c \in \mathbb{R}^d$*, the functions* $f_r, r = 1, \ldots, R$ *in Assumption 5.5.1 are strictly convex, any* $k \leq d-1$ *rows* $\{w_{l(1)}, \ldots, w_{l(k)}\}$ *of* W *do not satisfy the SCI condition, and* $h(x, \xi)$ *is jointly continuous in* x, ξ *and strictly concave in* x *for every* ξ*.*

Assumption 5.4.6 holds with $M = 1$*.*

Proving Theorem 5.5.5 requires utilizing the recently developed strong duality theory for Wasserstein DRO to show the joint continuity of the constraint (5.27) and the existence of a worst-case distribution (e.g., Blanchet and Murthy (2019); Gao and Kleywegt (2016)) to establish its strict convexity.

Moment-based DRO: We restrict our discussion in this case to individual linear chance constraints

$$\mathbb{P}_F(a_i'x \leq b_i) \geq 1 - \alpha_i, \quad \text{for } i = 1, \ldots, K \tag{5.28}$$

where each a_i is random and b_i is a deterministic constant, and α_i is an individual tolerance level. This setup also applies to the case of joint linear chance constraint, say, $\mathbb{P}_F(a_i'x \leq b_i \text{ for } i = 1, \ldots, K) \geq 1 - \alpha_i$, where one uses the Bonferroni correction to safely approximate with K single chance constraints $\mathbb{P}_F(a_i'x \leq b_i) \geq 1 - \frac{\alpha}{K}, i = 1, \ldots, K$. We restrict our discussion to (5.28) as it enables the tractable use of moment-based DRO; other settings are possible, but would lead to much more elaborate technicality that we do not pursue here.

We consider for each single constraint the following moment-based distributionally robust counterpart

$$\inf_{a_i \sim Q \text{ s.t. } (\mathbb{E}_Q[a_i], \mathrm{Cov}_Q[a_i]) \in \mathcal{U}_i(s)} \mathbb{P}_Q(a_i'x \leq b_i) \geq 1 - \alpha_i \quad \text{for } i = 1, \ldots, K$$

where each $\mathcal{U}_i(s)$ is a joint uncertainty set for the mean and covariance of the uncertain quantity a_i, all parametrized by the same s. For a fixed mean μ and covariance Σ, the robust constraint $\inf_{a_i \sim Q \text{ s.t. } \mathbb{E}_Q[a_i] = \mu, \mathrm{Cov}_Q(a_i) = \Sigma} \mathbb{P}_Q(a_i'x \leq b_i) \geq 1 - \alpha_i$ has an analytic expression $\mu'x + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x'\Sigma x} \leq$

b_i (Ghaoui et al. (2003)), therefore this moment-based DRO takes the form

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & f(x) \\ \text{subject to} \quad & \sup_{(\mu,\Sigma)\in\mathcal{U}_i(s)} \mu'x + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x'\Sigma x} \le b_i \ \text{ for } i = 1,\dots,K. \end{aligned} \tag{5.29}$$

Theorem 5.5.6 (Application of univariate Gaussian validator to moment-based DRO) *Consider $OPT(s)$ given by (5.29). Suppose Assumptions 5.5.1-5.5.3 hold, and that for each i the uncertainty set $\mathcal{U}_i(s)$ satisfies either (1)(2)(3) or (1)(2)(4) among: (1) $\mathcal{U}_i(s)$ is compact for all s; (ii) $\mathcal{U}_i(s) \subseteq \mathcal{U}_i(s')$ whenever $s < s'$ and $\overline{\cup_{s'<s}\mathcal{U}_i(s')} = \cap_{s'>s}\mathcal{U}_i(s') = \mathcal{U}_i(s)$ for all s; (3) for every s and every $(\mu,\Sigma) \in \mathcal{U}_i(s)$, Σ is positive definite; (4) $\mathcal{U}_i(s) = \mathcal{U}_i^1(s) \times \mathcal{U}_i^2(s)$, where $\mathcal{U}_i^1(s)$ and $\mathcal{U}_i^2(s)$ are uncertainty sets for the mean and covariance respectively, and there is a positive definite $\Sigma_s \in \mathcal{U}_i^2(s)$ such that $\Sigma \preceq \Sigma_s$ for all $\Sigma \in \mathcal{U}_i^2(s)$ where $\preceq$ is the ordering with respect to the positive semi-definite cone. Then, in either of the following two cases:*

i. $f(x)$ is continuous and strictly convex.

ii. $f(x) = c'x$ for some non-zero $c \in \mathbb{R}^d$, the functions $f_r, r = 1,\dots,R$ in Assumption 5.5.1 are strictly convex, any $k \le d-1$ rows $\{w_{l(1)},\dots,w_{l(k)}\}$ of W do not satisfy the SCI condition, and each $b_i \neq 0$.

Assumption 5.4.6 holds with $M = 1$.

Conditions (1) and (2) in Theorem 5.5.6 hold for common choices of moment-based uncertainty sets. We discuss some examples where (3) and (4) arise. (3) holds when $\mathcal{U}_i(s)$ is constructed to be a joint confidence region from, e.g., the delta method (Marandi et al. (2019)), for the mean and covariance of a_i whose covariance component converges to the true positive definite covariance as data size grows. (4) happens if the mean and covariance are treated separately and the uncertainty set for covariance takes the form $\mathcal{U}_i^2(s) = \{\Sigma : \Sigma_l(s) \preceq \Sigma \preceq \Sigma_u(s)\}$ (e.g., Delage and Ye (2010)).

RO with polyhedral uncertainty set: Consider the same linear chance constraint (5.28), and for

each i we use the robust counterpart $\sup_{a_i \in \mathcal{U}_i(s)} a_i' x \leq b_i$ where

$$\mathcal{U}_i(s) = \{a_i : \mathcal{W}_i a_i \leq z_i + s e_i\}$$

for some $\mathcal{W}_i \in \mathbb{R}^{l_i \times d}, z_i \in \mathbb{R}^{l_i}$ and $e_i \in \mathbb{R}^{l_i}_+ := [0, \infty)^{l_i}$. This robust counterpart can be expressed as a set of linear constraints, leading to the following formulation

$$\begin{array}{lll} \min\limits_{x \in \mathcal{X}} & f(x) & \\ \text{subject to} & (z_i + s e_i)' y_i \leq b_i & \\ & W_i' y_i = x & \\ & y_i \geq 0 \ \text{ for all } i = 1, \ldots, K & \end{array} \tag{5.30}$$

where each $y_i \in \mathbb{R}^{l_i}$ is an auxiliary variable.

Theorem 5.5.7 (Applying univariate Gaussian validator to polyhedral RO) *Let the formulation $OPT(s)$ be given by* (5.30)*. Suppose Assumptions 5.5.1-5.5.3 hold. If $f(x)$ is continuous and strictly convex, then Assumption 5.4.6 holds with $M = 1$. Otherwise, if $f(x) = c'x$ for some non-zero $c \in \mathbb{R}^d$, $R = 0$ in Assumption 5.5.1, the uncertainty set $\mathcal{U}_i(s_u)$ of maximum size is bounded for each i, and every $k \leq d-1$ element in $\{w_1, \ldots, w_L\} \cup \big(\cup_{i=1}^K \{\widetilde{\mathcal{W}}_i^{-1} z_i + s\widetilde{\mathcal{W}}_i^{-1} e_i : \widetilde{\mathcal{W}}_i \in \mathbb{R}^{d \times d}$ is an invertible submatrix of $\mathcal{W}_i\}\big)$ satisfies the SCI condition at only finitely many s values, then Assumption 5.4.6 holds with some finite M.*

The proof of Theorem 5.5.7 involves some technical developments to show that $x^*(s)$ has left and right limits at each discontinuity. This consists of transforming (5.30) into an equivalent parametric linear program whose constraints correspond to the vertices of the uncertainty sets, and then showing that its optimal basis stays constant in a neighborhood of each discontinuity. Lastly, we use the Jordan decomposition of the optimal basis matrix to establish the existence of left and right limits.

RO with ellipsoidal uncertainty set: Consider (5.28) again, and now for each constraint we consider

using $\sup_{a_i \in \mathcal{U}_i(s)} a_i' x \leq b_i$ with

$$\mathcal{U}_i(s) = \{a_i : a_i = \mu_i + \Sigma_i v, \|v\|_2 \leq s\}$$

for some positive definite $\Sigma_i \in \mathbb{R}^{d \times d}$, and $\mu_i \in \mathbb{R}^d$. This robust formulation has the following second-order cone representation

$$\begin{array}{ll} \min\limits_{x \in \mathcal{X}} & f(x) \\ \text{subject to} & \mu_i' x + s \left\|\Sigma_i x\right\|_2 \leq b_i \ \text{ for all } i = 1, \ldots, K. \end{array} \tag{5.31}$$

Theorem 5.5.8 (Applying univariate Gaussian validator to ellipsoidal RO) *Consider the formulation $OPT(s)$ given by* (5.31)*. Suppose Assumptions 5.5.1-5.5.3 hold, and each Σ_i is positive definite. In either of the following two cases:*

i. $f(x)$ is continuous and strictly convex.

ii. $f(x) = c'x$ for some non-zero $c \in \mathbb{R}^d$, the functions $f_r, r = 1, \ldots, R$ in Assumption 5.5.1 are strictly convex, any $k \leq d-1$ rows $\{w_{l(1)}, \ldots, w_{l(k)}\}$ of W do not satisfy the SCI condition, and each $b_i \neq 0$.

Assumption 5.4.6 holds with $M = 1$.

SO: Consider the CCP (5.12). Given the data $\{\xi_1, \ldots, \xi_n\}$, consider the following sequence $OPT(s)$ of programs

$$\begin{array}{ll} \min\limits_{x \in \mathcal{X}} & f(x) \\ \text{subject to} & (x, \xi_i) \in A \text{ for all } i = 1, \ldots, s \end{array} \tag{5.32}$$

for $1 \leq s \leq n$, i.e., each $OPT(s)$ uses only the first s sampled constraints. Although s takes integer values only, we can artificially extend the solution path to the continuum $[1, n]$ without introducing new solutions, by letting $x^*(s) = x^*(i)$ for all $s \in [i, i+1)$. Like the SAA formulation for chance constraints, the solution path $x^*(s)$ can now be viewed as piecewise constant in $s \in [1, n]$ hence Assumption 5.4.6 holds. Therefore we have:

Theorem 5.5.9 (Applying univariate Gaussian validator to SO) *Consider the formulation $OPT(s)$ given by* (5.32)*. Assumption 5.4.6 holds for some $M \leq n$.*

Lastly, our univariate Gaussian validator also works on a variant of SO called FAST (Carè et al. (2014)), in a sense that we will detail in Section 5.6.3. FAST differs from the formulations we have discussed so far in that its solution path does not come as solutions of a parametrized optimization problem, but from a line segment connecting two suitably chosen solutions. Nonetheless, the notion of solution-path optimality still applies. In particular, the solution-path optimum is unique if the objective is strictly convex or linear, and all the statistical guarantees in Theorem 5.4.4 can be established using the same proof.

5.6 Numerical Experiments

We present numerical results to demonstrate the performances of our framework in several data-driven reformulations. We consider the following linear CCP

$$\min\ c'x\ \text{ subject to }\ \mathbb{P}_F(\xi' x \leq b) \geq 1-\alpha \tag{5.33}$$

where $c \in \mathbb{R}^d, b \in \mathbb{R}$ are deterministic, the distribution F of the randomness $\xi \in \mathbb{R}^d$ is multivariate Gaussian with mean μ and covariance Σ, and the tolerance level $1-\alpha$ is set to 90%.

We consider a range of data-driven reformulations, including RO (or relatedly SCA), DRO (moment-based), and SO (including its variant FAST). In our experiments, we generate i.i.d. data $\xi_1, \ldots, \xi_n$ from the underlying true distribution F. Then, using a chosen reformulation, we compute a solution $\hat{x}$ of (5.33) that attempts to satisfy the chance constraint with a 95% confidence level, while attain an objective value $c'\hat{x}$ as low as possible. For each reformulation, we compare the performance of an existing benchmark with unnormalized and normalized Gaussian supremum validators (Algorithms 9 and 10) and univariate Gaussian validator (Algorithm 11), in terms of both feasibility and optimality. Moreover, we also test a naive validator that directly compares the sample mean to γ when checking feasibility, i.e., without the Gaussian margin $\frac{z_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}}$ in (5.15), in addition to the three proposed validators, which serves to demonstrate the necessity of the proposed Gaussian

margins in the validation procedure. The "plain average" column of each table displays results of this extra validator. "unnorm. GS" denotes the unnormalized Gaussian supremum validator (Algorithm 9), "norm. GS" denotes the normalized Gaussian supremum validator (Algorithm 10), and "uni. Gaussian" denotes the univariate Gaussian validator (Algorithm 11). When applying these validators in all experiments, we use the simple allocation rule of dividing the overall data size into Phases 1 and 2 equally, except only in the case of basic SO where a too small Phase 1 data size is provably subpar in guaranteeing feasibility.

To collect statistically meaningful estimates, for each formulation we repeat the experiments 1000 times each with an independently generated data set and a data-driven solution output. We take down the average objective value achieved by these solutions (the "mean obj. val." row of each table) and the proportion of feasible solutions as the empirical feasibility coverage (the "feasibility level" row of each table). Therefore, the smaller the "mean obj. val." is, the better is the solution in terms of optimality, and "feasibility level" $\geq 95\%$ indicates that the desired feasibility confidence level is achieved and otherwise not.

5.6.1 RO and SCA

We first test the proposed framework on RO. We use the ellipsoid uncertainty set that leads to a robust counterpart in the form described in Example 5.1.2, i.e., $\hat{\mu}'x + \sqrt{s}\|\hat{\Sigma}^{1/2}x\|_2 \leq b$ where $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance for ξ computed from Phase one data. The benchmark ("SCA" in the tables) is set to an SCA (equation 2.4.11 of Ben-Tal et al. (2009)) for unbounded ξ, which in our case can be expressed as

$$\mu'x + \sqrt{2\log\frac{1}{\alpha}}\sqrt{\sum_{k=1}^{d}(z^{k\prime}x)^2} = \mu'x + \sqrt{2\log\frac{1}{\alpha}}\|\Sigma^{1/2}x\|_2 \leq b \tag{5.34}$$

where μ is the true mean, and z^k is the k-th column of the square root $\Sigma^{1/2}$ of the true covariance matrix Σ. Note that (5.34) is equivalent to the RO formulation with true mean and covariance and parameter value $s = 2\log\frac{1}{\alpha}$. Here, we give this SCA or RO the advantage of knowing the true mean and covariance of the randomness.

To implement our validator, we need to provide a set of parameter values $\{s_1, \ldots, s_p\}$ at which the RO is solved. We take the $(1-\alpha)n_1$-th order statistic $\hat{s}_{1-\alpha}$ of $\{(\xi_{n_2+i} - \hat{\mu})'\hat{\Sigma}^{-1}(\xi_{n_2+i} - \hat{\mu}) : i = 1, \ldots, n_1\}$, where $\xi_{n_2+i}, i = 1, \ldots, n_1$ are the Phase one data, so that $\{\xi : (\xi - \hat{\mu})'\hat{\Sigma}^{-1}(\xi - \hat{\mu}) \leq \hat{s}_{1-\alpha}\}$ is roughly a $(1-\alpha)$-content set for ξ (such type of quantile-based selection has been used in Hong et al. (2017)). We then set the values $s_j = (\hat{s}_{1-\alpha} + 20)\frac{j}{50}$ for $j = 1, \ldots, 50$ ($p = 50$). Tables 5.1, 5.2 and 5.3 summarize the results under different problem dimensions and data sizes.

	SCA	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-3.57	-3.68	-4.20	-4.43	-5.15
feasibility level	100%	99.9%	98.5%	97.5%	76.9%

Table 5.1: RO with ellipsoidal uncertainty set. $d = 10, n = 200$. Data are split to $n_1 = 100, n_2 = 100$.

	SCA	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-3.57	-4.42	-4.58	-4.80	-5.34
feasibility level	100%	99.8%	99.6%	98.8%	77.9%

Table 5.2: RO with ellipsoidal uncertainty set. $d = 10, n = 500$. Data are split to $n_1 = 250, n_2 = 250$.

	SCA	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-16.70	-17.59	-17.33	-17.71	-20.31
feasibility level	100%	98.4%	99.6%	98.4%	82.7%

Table 5.3: RO with ellipsoidal uncertainty set. $d = 50, n = 500$. Data are split to $n_1 = 250, n_2 = 250$.

We highlight a few observations. First, our framework with the three proposed validators outperforms the SCA benchmark. In terms of the objective performance, both our unnormalized and normalized Gaussian supremum validators, and univariate Gaussian validators, achieve lower objective value than SCA (with a difference ≥ 0.6), while at the same time retain the feasibility confidence to above 95% in all the three tables. In particular, as the dimension grows from 10 (Tables 5.1 and 5.2) to 50 (Table 5.3), the feasibility confidence level remains above 95%, consistent with the dimension-free feasibility guarantee of our methods. Second, among the three proposed

validators, the univariate Gaussian validator appears less conservative than the Gaussian supremum counterparts in achieving better objective values, and relatedly tighter feasibility confidence levels (i.e., closer to 95%). Specifically, the univariate Gaussian validator gives a feasibility confidence level around 98% in all the three tables, whereas the Gaussian supremum validators give a level between 99%-100% (and also 0.1-0.4 higher mean objective values). Finally, we comment that the "plain average" scheme does not have the desired feasibility confidence level even when the data size is as large as 500 (Table 5.2), which shows that margin adjustments to the naive sample average in the validators is necessary to ensure feasibility.

5.6.2 Moment-Based DRO

The second formulation we consider is a moment-based DRO. We use the formulation

$$\inf_{\xi\sim Q \text{ s.t. } (\mathbb{E}_Q[\xi],\mathrm{Cov}_Q(\xi))\in\mathcal{U}_s} \mathbb{P}_Q(\xi' x \le b) \ge 1-\alpha \tag{5.35}$$

where $\mathcal{U}_s$ is a confidence region for the true mean and covariance of ξ obtained via the delta method described in Example 5.1.3 (see Section 6 of Marandi et al. (2019) for details). According to (5.29), (5.35) can be expressed as $\sup_{(\mu,\Sigma)\in\mathcal{U}_s} \mu' x + \sqrt{\frac{1-\alpha}{\alpha}} \left\|\Sigma^{1/2} x\right\|_2 \le b$, which can be further reformulated as a conic constraint (see Theorem 1 of Marandi et al. (2019)). In the benchmark case "DRO (χ^2 quantile)" we choose s to be the 95% quantile of the limiting χ^2 distribution as suggested in Marandi et al. (2019) so that $\mathcal{U}_s$ is a valid 95% confidence region. In our framework, we solve the DRO formulation at parameter values $s_j = 1.5\hat{s}_{0.95}\frac{j}{50}$ for $j = 1,\ldots,50$ where $\hat{s}_{0.95}$ is the χ^2 quantile used in the benchmark. Tables 5.4 and 5.5 show the experimental results under different data sizes.

	DRO (χ^2 quantile)	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-1.83	-2.73	-2.73	-2.73	-2.73
feasibility level	100%	100%	100%	100%	100%

Table 5.4: Moment-based DRO. $d = 10, n = 200$. Data are split to $n_1 = 100, n_2 = 100$.

The comparisons between the benchmark and our framework here share similarities with the

	DRO (χ^2 quantile)	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-2.00	-2.62	-2.62	-2.62	-2.62
feasibility level	100%	100%	100%	100%	100%

Table 5.5: Moment-based DRO. $d = 10, n = 500$. Data are split to $n_1 = 250, n_2 = 250$.

RO setting. The solutions output from our validators possess superior objective performance (with a difference of 0.6-0.9) than simply setting s to be the 95%-level χ^2 quantile, while still attain the desired feasibility confidence level. Note that all validators (including the "plain average") give the same objective value (-2.73 in Table 5.4 and -2.62 in Table 5.5), and have a 100% feasibility confidence. This is because the chosen parameter s turns out to be 0 for all of them. In other words, setting the moment constraints as equalities (to the estimated moments from Phase one data) is statistically feasible and achieves the best objective value, and any relaxation from this would lead to a deterioration of solution quality. This hints that the conventional choices of moment set size suggested in the literature could be very conservative.

5.6.3 SO

Given the Phase one data $\{\xi_{n_2+1}, \ldots, \xi_n\}$, we consider the data-driven feasible region specified by the first s sampled constraints, $\xi'_{n_2+i}x \leq b$ for $i = 1, \ldots, s$, and tune the number of satisfied constraints $s \in \{1, 2, \ldots, n_1\}$. The benchmark "SO" in this case is to impose all the constraints given by the whole data set $\{\xi_1, \ldots, \xi_n\}$. Tables 5.6 and 5.7 summarize the results for data size $n = 200, 500$ respectively.

	SO	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-3.90	-4.24	-4.31	-4.46	-4.91
feasibility level	99.7%	95.2%	94.0%	85.1%	44.7%

Table 5.6: SO. $d = 10, n = 200$. Data are split to $n_1 = 150, n_2 = 50$.

We observe the gain in objective performance of our validators compared to SO (a difference of 0.3-0.6 in Table 5.6 and 0.6-1.0 in Table 5.7). We also note the drastic failure of "plain average" in rendering the desired 95% feasibility confidence, thus showing that a margin adjustment to the

	SO	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-3.28	-3.86	-4.10	-4.30	-4.69
feasibility level	100%	99.7%	98.7%	95.6%	62.0%

Table 5.7: SO. $d = 10, n = 500$. Data are split to $n_1 = 250, n_2 = 250$.

validators is necessary. Our validators maintain feasibility in all cases, except the univariate Gaussian validator for $n = 200$. This deficiency is attributed to two potential reasons. First is that with $n_1 = 150$ there is a non-negligible chance that none of the n_1 solutions $x^*(s), s = 1, \ldots, n_1$, produced in Phase one is feasible, thus violating Assumption 5.4.7. In fact, the infeasibility probability of the solution derived by an SO using all the n_1 constraints can be computed to be 6% (Campi and Garatti (2008)), leaving the actual confidence of obtaining a feasible solution at most 94%. The second possible cause is the finite-sample coverage error of the univariate Gaussian validator, seeing that the validation data size $n_2 = 50$ is relatively small. When both n_1 and n_2 increase to 250 in Table 5.7, the desired feasibility confidence level is recovered for the univariate Gaussian validator as the chance of all solution candidates being infeasible decreases to $< 0.2\%$ and the finite-sample error is reduced due to a larger validation data size. Finally, although we do not pursue here, we should mention that the performances of the basic SO considered in the tables can plausibly be boosted by using techniques such as sampling-and-discarding (Campi and Garatti (2011)) and wait-and-judge (Campi and Garatti (2018)). Comprehensive comparisons with these enhanced techniques would be left as important future work.

Lastly, we consider a variant of SO called FAST (Carè et al. (2014)), designed originally to tone down the sample size requirement in basic SO. Our comparison with FAST here is motivated by its similarity with our framework in that it also splits the data into two portions and uses a validation-based idea. With the first portion of data $\{\xi_{n_2+i}, i = 1, \ldots, n_1\}$, FAST computes a solution $\hat{x}$ by imposing all the n_1 constraints $\xi'_{n_2+i}x \leq b$ as in the basic SO, and then uses the second portion to obtain the final solution $\hat{x}^*$ by solving the following program

$$\min\ c'((1-s)x_o + s\hat{x})\ \text{ subject to }\ \xi_i'((1-s)x_o + s\hat{x}) \leq b \text{ for all } i = 1, \ldots, n_2 \text{ and } 0 \leq s \leq 1$$

where x_o is a feasible solution of (5.33) with $\mathbb{P}_F(\xi' x_o \leq b) = 1$. One particular choice of x_o for problem (5.33) is the vector of all zeros and is used in the experiment. When applying our framework to FAST, we search for the best feasible solution along the line segment $\{x^*(s) = (1-s)x_o + s\hat{x} : s \in [0,1]\}$ by validating solutions $x^*(s_j)$ at parameter values $s_j = \frac{j-1}{10}$ for $j = 1, \ldots, 11$ ($p = 11$). Tables 5.8 and 5.9 show the results under different dimensions and data sizes.

	FAST	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-2.54	-3.55	-3.68	-3.87	-4.44
feasibility level	100%	98.9%	98.9%	97.3%	79.6%

Table 5.8: FAST. $d = 10, n = 200$. Data are split to $n_1 = 100, n_2 = 100$.

	FAST	unnorm. GS	norm. GS	uni. Gaussian	plain average
mean obj. val.	-8.92	-14.11	-15.06	-15.80	-18.14
feasibility level	100%	99.8%	99.3%	98.0%	76.7%

Table 5.9: FAST. $d = 50, n = 500$. Data are split to $n_1 = 250, n_2 = 250$.

Similar phenomena persist from our previous settings. Our three validators give tighter feasibility confidence levels and better objective performances (with a difference of ≥ 1 in Table 5.8 and ≥ 5 in Table 5.9) compared to FAST. Among them, univariate Gaussian validator gives the tightest feasibility confidence level and best objective value. The naive "plain average" validator fails in attaining the desired feasibility confidence. Here we have used a rather coarse mesh with only 11 parameter values, and expect a sharper improvement should a finer mesh be used.

5.7 Conclusion

We have studied a validation-based framework to combat the conservativeness in data-driven optimization with uncertain constraints. We have demonstrated how the conventional approaches in several optimization paradigms, including SAA, RO and DRO, implicitly estimate the whole feasible region. This in turn leads to over-conservativeness caused by the need to control huge simultaneous estimation errors, especially for high-dimensional problems. On the other hand, we have also

demonstrated that the solution output from these reformulation classes can often be represented in a low-dimensional manifold parametrized by key conservativeness parameters. Our framework leverages this low dimensionality by extracting the parametrized solution path and selecting the best parameter value. We have proposed two types of validators for this parameter selection, one utilizing a multivariate Gaussian supremum (unnormalized or normalized) and another utilizing a univariate Gaussian, to set the margin in a sample average constraint when optimizing over the solution path. We have shown that the obtained solutions enjoy asymptotic and finite-sample performance guarantees on feasibility that scale lightly with the problem dimension, and asymptotic optimality within the reformulation class. The Gaussian supremum validator requires less regularity conditions and is applicable more generally, whereas the univariate Gaussian validator provides tighter guarantees when applicable. Our numerical results support these findings and show that our framework and validators consistently provide better solutions compared to several benchmarks in terms of better objective values and tighter feasibility confidence. Our study provides a first rigorous validation-based framework to tackle over-conservativeness in data-driven constrained optimization, and is foreseen to open up follow-up investigations on more powerful validation strategies and refined statistical guarantees regarding joint feasibility and optimality.

Bibliography

Abadie A, Imbens GW (2008) On the failure of the bootstrap for matching estimators. *Econometrica* 76(6):1537–1557.

Andrews DW, Guggenberger P (2009) Validity of subsampling and "plug-in asymptotic" inference for parameters defined by moment inequalities. *Econometric Theory* 25(3):669–709.

Andrews DW, Guggenberger P (2010) Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory* 26(2):426–468.

Asmussen S, Glynn PW (2007) *Stochastic Simulation: Algorithms and Analysis*, volume 57 (Springer Science & Business Media).

Atar R, Chowdhary K, Dupuis P (2015) Robust bounds on risk-sensitive functionals via renyi divergence. *SIAM/ASA Journal on Uncertainty Quantification* 3(1):18–33.

Atlason J, Epelman MA, Henderson SG (2004) Call center staffing with simulation and cutting plane methods. *Annals of Operations Research* 127(1-4):333–358.

Banks J, Carson II JS, Nelson BL, Nicol DM (2005) *Discrete-Event System Simulation* (Pearson), fourth edition.

Barton RR (2012) Tutorial: Input uncertainty in outout analysis. Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM, eds., *Proceedings of the 2012 Winter Simulation Conference (WSC)*, 1–12 (IEEE).

Barton RR, Chick SE, Cheng RC, Henderson SG, Law AM, Schmeiser BW, Leemis LM, Schruben LW, Wilson JR (2002) Panel discussion on current issues in input modeling. Yücesan E, Chen CH, Snowdon JL, Charnes JM, eds., *Proceedings of the 2002 Winter Simulation Conference (WSC)*, 353–369 (Piscataway, New Jersey: IEEE).

Barton RR, Lam H, Song E (2018) Revisiting direct bootstrap resampling for input model uncertainty. Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B, eds., *Proceedings of the 2018 Winter Simulation Conference (WSC)*, 1635–1645 (IEEE).

Barton RR, Nelson BL, Xie W (2013) Quantifying input uncertainty via simulation confidence intervals. *INFORMS Journal on Computing* 26(1):74–87.

Barton RR, Schruben LW (1993) Uniform and bootstrap resampling of empirical distributions. Evans GW, Mollaghasemi M, Russell EC, Biles WE, eds., *Proceedings of the 1993 Winter Simulation Conference (WSC)*, 503–508 (IEEE).

Barton RR, Schruben LW (2001) Resampling methods for input modeling. Peters BA, Smith JS, Medeiros DJ, Rohrer MW, eds., *Proceedings of the 2001 Winter Simulation Conference (WSC)*, volume 1, 372–378 (Piscataway, New Jersey: IEEE).

Barton RR, et al. (2007) Presenting a more complete characterization of uncertainty: Can it be done. *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*, 26–60 (INFORMS Simulation Society).

Bayraksan G, Love DK (2015) Data-driven stochastic programming using phi-divergences. *INFORMS TutORials in Operations Research* 1–19.

Bayraksan G, Morton DP (2006) Assessing solution quality in stochastic programs. *Mathematical Programming* 108(2-3):495–514.

Bayraksan G, Morton DP (2011) A sequential sampling procedure for stochastic programming. *Operations Research* 59(4):898–913.

Ben-Tal A, Den Hertog D, De Waegenaere A, Melenberg B, Rennen G (2013) Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59(2):341–357.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press).

Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming* 88(3):411–424.

Ben-Tal A, Nemirovski A (2002) Robust optimization–methodology and applications. *Mathematical Programming* 92(3):453–480.

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Review* 53(3):464–501.

Bertsimas D, Gupta V, Kallus N (2018a) Data-driven robust optimization. *Mathematical Programming* 167(2):235–292.

Bertsimas D, Gupta V, Kallus N (2018b) Robust sample average approximation. *Mathematical Programming* 171(1-2):217–282.

Bickel PJ, Götze F, van Zwet WR (1997) Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica* 7(1):1–31.

Bickel PJ, Sakov A (2008) On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica* 18(3):967–985.

Birge JR, Louveaux F (2011) *Introduction to Stochastic Programming* (Springer Science & Business Media).

Blanchet J, He F, Murthy K (2020) On distributionally robust extreme value analysis. *Extremes* 1–31.

Blanchet J, Kang Y (2020) Sample out-of-sample inference based on Wasserstein distance. *Operations Research* To appear.

Blanchet J, Kang Y, Murthy K (2019) Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* 56(3):830–857.

Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research* 44(2):565–600.

Blom G (1976) Some properties of incomplete U-statistics. *Biometrika* 63(3):573–580.

Bonnans JF, Shapiro A (2013) *Perturbation Analysis of Optimization Problems* (Springer Science & Business Media).

Borkar VS (2009) *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48 (Springer).

Boucheron S, Lugosi G, Massart P (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press).

Breiman L (1996) Bagging predictors. *Machine Learning* 24(2):123–140.

Büchlmann P, Yu B (2002) Analyzing bagging. *Annals of Statistics* 927–961.

Calafiore G, Campi MC (2005) Uncertain convex programs: Randomized solutions and confidence levels. *Mathematical Programming* 102(1):25–46.

Calafiore GC (2017) Repetitive scenario design. *IEEE Transactions on Automatic Control* 62(3):1125–1137.

Campi MC, Carè A (2013) Random convex programs with l_1-regularization: Sparsity and generalization. *SIAM Journal on Control and Optimization* 51(5):3532–3557.

Campi MC, Garatti S (2008) The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization* 19(3):1211–1230.

Campi MC, Garatti S (2011) A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality. *Journal of Optimization Theory and Applications* 148(2):257–280.

Campi MC, Garatti S (2018) Wait-and-judge scenario optimization. *Mathematical Programming* 167(1):155–189.

Carè A, Garatti S, Campi MC (2014) FAST – Fast algorithm for the scenario technique. *Operations Research* 62(3):662–671.

Cheng RC, Holland W (1997) Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation* 57(1-4):219–241.

Cheng RC, Holland W (1998) Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation Simulation* 60(3):183–205.

Cheng RC, Holland W (2004) Calculation of confidence intervals for simulation output. *ACM Transactions on Modeling and Computer Simulation* 14(4):344–362.

Chernozhukov V, Chetverikov D, Kato K, et al. (2017) Central limit theorems and bootstrap in high dimensions. *The Annals of Probability* 45(4):2309–2352.

Chick SE (2001) Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research* 49(5):744–758.

Chick SE (2006) Bayesian ideas and discrete event simulation: Why, what and how. Perrone LF, Wieland FP, Liu J, Lawson BG, Nicol DM, Fujimoto RM, eds., *Proceedings of the 2006 Winter Simulation Conference (WSC)*, 96–106 (Piscataway, New Jersey: IEEE).

Chu EM, Chang KY, Lu CY, Hsu CH, Liu CH, Hsiao YC (2014) A new approach to determine the critical path in stochastic activity network. *2014 International Symposium on Computer, Consumer and Control*, 1123–1128 (IEEE).

Datta S, McCormick WP (1995) Bootstrap inference for a first-order autoregression with positive innovations. *Journal of the American Statistical Association* 90(432):1289–1300.

De la Pena V, Giné E (2012) *Decoupling: From Dependence to Independence* (Springer Science & Business Media).

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* 58(3):595–612.

Dentcheva D, Penev S, Ruszczyński A (2017) Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics* 69(4):737–760.

Dey S, Juneja S (2012) Incorporating fat tails in financial models using entropic divergence measures. *arXiv preprint arXiv:1203.0643* .

Ding J, Eldan R, Zhai A (2015) On multiple peaks and moderate deviations for the supremum of a gaussian field. *The Annals of Probability* 43(6):3468–3493.

Doan XV, Li X, Natarajan K (2015) Robustness to dependency in portfolio optimization using overlapping marginals. *Operations Research* 63(6):1468–1488.

Duchi J, Glynn P, Namkoong H (2016) Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425* .

Dupuis P, Katsoulakis MA, Pantazis Y, Plechác P (2016) Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA Journal on Uncertainty Quantification* 4(1):80–111.

Dür M, Jargalsaikhan B, Still G (2016) Genericity results in linear conic programming—a tour d'horizon. *Mathematics of Operations Research* 42(1):77–94.

Durrett R (2010) *Probability: Theory and Examples* (Cambridge university press).

Efron B (2014) Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109(507):991–1007.

Efron B, Stein C (1981) The jackknife estimate of variance. *The Annals of Statistics* 9(3):586–596.

Eichhorn A, Römisch W (2007) Stochastic integer programming: Limit theorems and confidence intervals. *Mathematics of Operations Research* 32(1):118–135.

Esfahani PM, Kuhn D (2018) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming* 171(1-2):115–166.

Fournier N, Guillin A (2015) On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields* 162(3-4):707–738.

Frees EW (1989) Infinite order U-statistics. *Scandinavian Journal of Statistics* 16(1):29–45.

Freund RM (1985) Postoptimal analysis of a linear program under simultaneous changes in matrix coefficients. *Mathematical Programming Essays in Honor of George B. Dantzig Part I*, 1–13 (Springer).

Friedman J, Hastie T, Tibshirani R (2001) *The Elements of Statistical Learning*, volume 1 (Springer series in statistics New York).

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199* .

Ghadimi S, Lan G (2013) Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization* 23(4):2341–2368.

Ghaoui LE, Oks M, Oustry F (2003) Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations research* 51(4):543–556.

Ghosh S, Lam H (2015) Mirror descent stochastic approximation for computing worst-case stochastic input models. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds., *Proceedings of the 2015 Winter Simulation Conference (WSC)*, 425–436 (Piscataway, New Jersey: IEEE).

Ghosh S, Lam H (2019) Robust analysis in stochastic simulation: Computation and performance guarantees. *Operations Research* 67(1):232–249.

Glasserman P (2013) *Monte Carlo Methods in Financial Engineering*, volume 53 (Springer Science & Business Media).

Glasserman P, Xu X (2013) Robust portfolio control with stochastic factor dynamics. *Operations Research* 61(4):874–893.

Glasserman P, Xu X (2014) Robust risk measurement and model risk. *Quantitative Finance* 14(1):29–58.

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Operations Research* 58(4-part-1):902–917.

Goldfarb D, Iyengar G (2003) Robust portfolio selection problems. *Mathematics of Operations Research* 28(1):1–38.

Gotoh Jy, Kim MJ, Lim AE (2018) Robust empirical optimization is almost the same as mean–variance optimization. *Operations Research Letters* 46(4):448–452.

Gupta V (2019) Near-optimal Bayesian ambiguity sets for distributionally robust optimization. *Management Science* 65(9):4242–4260.

Hall P, Horowitz JL, Jing BY (1995) On blocking rules for the bootstrap with dependent data. *Biometrika* 82(3):561–574.

Hampel FR (1974) The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69(346):383–393.

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (2011) *Robust Statistics: The Approach Based on Influence Functions*, volume 114 (John Wiley & Sons).

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2015) A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming* 151(1):35–62.

Hansen LP, Sargent TJ (2008) *Robustness* (Princeton University Press).

Henderson SG (2003) Input model uncertainty: Why do we care and what should we do about it? Chick S, Sánchez PJ, Ferrin D, Morrice DJ, eds., *Proceedings of the 2003 Winter Simulation Conference (WSC)*, 90–100 (Piscataway, New Jersey: IEEE).

Hjort NL, McKeague IW, Van Keilegom I (2009) Extending the scope of empirical likelihood. *The Annals of Statistics* 1079–1111.

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* 58(301):13–30.

Hong LJ, Huang Z, Lam H (2017) Learning-based robust optimization: Procedures and statistical guarantees. *arXiv preprint arXiv:1704.04342* .

Hu Z, Cao J, Hong LJ (2012) Robust simulation of global warming policies using the dice model. *Management Science* 58(12):2190–2206.

Hu Z, Hong LJ (2013) Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online* .

Huber PJ, Ronchetti EM (2009) Robust statistics. *International Encyclopedia of Statistical Science*, 1248–1251 (Springer), second edition.

Iyengar GN (2005) Robust dynamic programming. *Mathematics of Operations Research* 30(2):257–280.

Jain A, Lim AE, Shanthikumar JG (2010) On the optimality of threshold control in queues with model uncertainty. *Queueing Systems* 65(2):157–174.

Janson S (1984) The asymptotic distributions of incomplete U-statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 66(4):495–505.

Jiang R, Guan Y (2016) Data-driven chance constrained stochastic program. *Mathematical Programming* 158(1-2):291–327.

Kleywegt AJ, Shapiro A, Homem-de-Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* 12(2):479–502.

Kosorok MR (2008) *Introduction to Empirical Processes and Semiparametric Inference* (Springer).

Krokhmal P, Palmquist J, Uryasev S (2002) Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk* 4:43–68.

Kushner H, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 (Springer Science & Business Media).

Lam H (2016a) Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation. Roeder TMK, Frazier PI, Szechtman R, Zhou E, Huschka T, Chick SE, eds., *Proceedings of the 2016 Winter Simulation Conference (WSC)*, 178–192 (IEEE).

Lam H (2016b) Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research* 41(4):1248–1275.

Lam H (2018) Sensitivity to serial dependency of input processes: A robust approach. *Management Science* 64(3):1311–1327.

Lam H (2019) Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research* 67(4):1090–1105.

Lam H, Qian H (2016) The empirical likelihood approach to simulation input uncertainty. Roeder TMK, Frazier PI, Szechtman R, Zhou E, Huschka T, Chick SE, eds., *Proceedings of the 2016 Winter Simulation Conference (WSC)*, 791–802 (IEEE).

Lam H, Qian H (2017) Optimization-based quantification of simulation input uncertainty via empirical likelihood. *arXiv preprint arXiv:1707.05917* .

Lam H, Qian H (2018a) Assessing solution quality in stochastic optimization via bootstrap aggregating. Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B, eds., *Proceedings of the 2018 Winter Simulation Conference (WSC)*, 2061–2071 (IEEE).

Lam H, Qian H (2018b) Bounding optimality gap in stochastic optimization via bagging: Statistical efficiency and stability. *arXiv preprint arXiv:1810.02905* .

Lam H, Qian H (2018c) Subsampling to enhance efficiency in input uncertainty quantification. *arXiv preprint arXiv:1811.04500* .

Lam H, Qian H (2018d) Subsampling variance for input uncertainty quantification. Rabe M, Juan AA, Mustafee N, Skoogh A, Jain S, Johansson B, eds., *Proceedings of the 2018 Winter Simulation Conference (WSC)*, 1611–1622 (IEEE).

Lam H, Qian H (2019a) Combating conservativeness in data-driven optimization under uncertainty: A solution path approach. *arXiv preprint arXiv:1909.06477* .

Lam H, Qian H (2019b) Validating optimization with uncertain constraints. Mustafee N, Bae KHG, Lazarova-Molnar S, Rabe M, Szabo C, Haas P, Son YJ, eds., *Proceedings of the 2019 Winter Simulation Conference (WSC)*, 3621–3632 (IEEE).

Lam H, Zhou E (2017) The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters* 45(4):301–307.

Lan G, Nemirovski A, Shapiro A (2012) Validation analysis of mirror descent stochastic approximation method. *Mathematical programming* 134(2):425–458.

Law AM, Kelton WD, Kelton WD (2000) *Simulation Modeling and Analysis*, volume 3 (McGraw-Hill New York).

Lee AJ (2019) *U-Statistics: Theory and Practice* (Routledge).

Lifshits MA (2013) *Gaussian Random Functions*, volume 322 (Springer Science & Business Media).

Lim AE, Shanthikumar JG (2007) Relative entropy, exponential utility, and robust dynamic pricing. *Operations Research* 55(2):198–214.

Lin Y, Song E, Nelson B (2015) Single-experiment input uncertainty. *Journal of Simulation* 9(3):249–259.

Liu Y, Zou C, Zhang R (2008) Empirical likelihood for the two-sample mean problem. *Statistics & Probability Letters* 78(5):548–556.

Love D, Bayraksan G (2011) Overlapping batches for the assessment of solution quality in stochastic programs. Jain S, Creasey RR, Himmelspach J, White KP, Fu M, eds., *Proceedings of the 2011 Winter Simulation Conference (WSC)*, 4179–4190 (IEEE).

Love D, Bayraksan G (2015) Overlapping batches for the assessment of solution quality in stochastic programs. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 25(3):1–20.

Luedtke J, Ahmed S (2008) A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization* 19(2):674–699.

Mak WK, Morton DP, Wood RK (1999) Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters* 24(1-2):47–56.

Marandi A, Ben-Tal A, Hertog Dd, Melenberg B (2019) Extending the scope of robust quadratic optimization. *arXiv preprint arXiv:1909.01762* .

Maurer A, Pontil M (2009) Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740* .

Mentch L, Hooker G (2016) Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research* 17(1):841–881.

Natarajan K, Pachamanova D, Sim M (2008) Incorporating asymmetric distributional information in robust value-at-risk optimization. *Management Science* 54(3):573–585.

Nelson B (2013) *Foundations and Methods of Stochastic Simulation: a First Course* (Springer Science & Business Media).

Nemirovski A (2003) On tractable approximations of randomly perturbed convex constraints. *42nd IEEE International Conference on Decision and Control*, volume 3, 2419–2422 (IEEE).

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* 19(4):1574–1609.

Nemirovski A, Shapiro A (2006) Convex approximations of chance constrained programs. *SIAM Journal on Optimization* 17(4):969–996.

Nilim A, El Ghaoui L (2005) Robust control of markov decision processes with uncertain transition matrices. *Operations Research* 53(5):780–798.

Owen A (1990) Empirical likelihood ratio confidence regions. *The Annals of Statistics* 90–120.

Owen A (1991) Empirical likelihood for linear models. *The Annals of Statistics* 1725–1747.

Owen AB (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75(2):237–249.

Owen AB (2001) *Empirical Likelihood* (CRC press).

Pagnoncelli B, Ahmed S, Shapiro A (2009) Sample average approximation method for chance constrained programming: Theory and applications. *Journal of Optimization Theory and Applications* 142(2):399–416.

Pardo L (2005) *Statistical Inference Based on Divergence Measures* (CRC Press).

Partani A (2007) *Adaptive Jacknife Estimators for Stochastic Programming*. Ph.D. thesis.

Partani A, Morton DP, Popova I (2006) Jackknife estimators for reducing bias in asset allocation. Perrone LF, Wieland FP, Liu J, Lawson BG, Nicol DM, Fujimoto RM, eds., *Proceedings of the 2006 Winter Simulation Conference (WSC)*, 783–791 (IEEE).

Peel T, Anthoine S, Ralaivola L (2010) Empirical bernstein inequalities for U-statistics. *Advances in Neural Information Processing Systems*, 1903–1911.

Petersen IR, James MR, Dupuis P (2000) Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control* 45(3):398–412.

Politis DN, Romano JP (1994) Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics* 2031–2050.

Politis DN, Romano JP, Wolf M (1999) *Subsampling* (Springer).

Prékopa A (2003) Probabilistic programming. *Handbooks in Operations Research and Management Science* 10:267–351.

Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. *The Annals of Statistics* 300–325.

Rao RR (1962) Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics* 659–680.

Rennie BC, Dobson AJ (1969) On Stirling numbers of the second kind. *Journal of Combinatorial Theory* 7(2):116–121.

Rockafellar RT (1970) *Convex Analysis* (Princeton University Press).

Schildbach G, Fagiano L, Morari M (2013) Randomized solutions to convex programs with multiple chance constraints. *SIAM Journal on Optimization* 23(4):2479–2501.

Searle SR, Casella G, McCulloch CE (2009) *Variance Components*, volume 391 (John Wiley & Sons).

Sen B, Banerjee M, Woodroofe M, et al. (2010) Inconsistency of bootstrap: The grenander estimator. *The Annals of Statistics* 38(4):1953–1977.

Serfling RJ (2009) *Approximation Theorems of Mathematical Statistics*, volume 162 (John Wiley & Sons).

Shapiro A (2003) Monte Carlo sampling methods. *Handbooks in Operations Research and Management Science* 10:353–425.

Shapiro A, Dentcheva D, Ruszczyński A (2014) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM).

Shapiro A, Homem-de-Mello T (2000) On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. *SIAM Journal on Optimization* 11(1):70–86.

Shapiro A, Nemirovski A (2005) On complexity of stochastic programming problems. *Continuous Optimization*, 111–146 (Springer).

Song E, Nelson BL (2015) Quickly assessing contributions to input uncertainty. *IIE Transactions* 47(9):893–909.

Song E, Nelson BL (2019) Input–output uncertainty comparisons for discrete optimization via simulation. *Operations Research* 67(2):562–576.

Song E, Nelson BL, Pegden CD (2014) Advanced tutorial: Input uncertainty quantification. Tolk A, Diallo S, Ryzhov I, Yilmaz L, Buckley S, Miller J, eds., *Proceedings of the 2014 Winter Simulation Conference (WSC)*, 162–176 (Piscataway, New Jersey: IEEE).

Stockbridge R, Bayraksan G (2013) A probability metrics approach for reducing the bias of optimality gap estimators in two-stage stochastic linear programming. *Mathematical Programming* 142(1-2):107–131.

Sun Y, Apley DW, Staum J (2011) Efficient nested simulation for estimating the variance of a conditional expectation. *Operations research* 59(4):998–1007.

Sun Y, Phillips PC, Jin S (2008) Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing. *Econometrica* 76(1):175–194.

Tulabandhula T, Rudin C (2014) Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097* .

Van der Vaart AW (2000) *Asymptotic Statistics*, volume 3 (Cambridge University Press).

Van der Vaart AW, Wellner JA (1996) *Weak Convergence and Empirical Processes with Applications to Statistics* (Springer).

Van Parys BP, Goulart PJ, Kuhn D (2016) Generalized gauss inequalities via semidefinite programming. *Mathematical Programming* 156(1-2):271–302.

Vershynin R (2010) Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027* .

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research* 15(1):1625–1651.

Wand MP, Jones MC (1994) *Kernel Smoothing* (Chapman and Hall/CRC).

Wang W, Ahmed S (2008) Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters* 36(5):515–519.

Wieland JR, Schmeiser BW (2006) Stochastic gradient estimation using a single design point. Perrone LF, Wieland FP, Liu J, Lawson BG, Nicol DM, Fujimoto RM, eds., *Proceedings of the 2006 Winter Simulation Conference (WSC)*, 390–397 (Piscataway, New Jersey: IEEE).

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Operations Research* 62(6):1358–1376.

Wilks SS (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* 9(1):60–62.

Wu C, Yan Y (2012) Empirical likelihood inference for two-sample problems. *Statistics and its Interface* 5(3):345–354.

Xie W (2019) On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming* URL `http://dx.doi.org/10.1007/s10107-019-01445-5`.

Xie W, Li C, Zhang P (2018) A Bayesian nonparametric hierarchical framework for uncertainty quantification in simulation. *Submitted to Operations Research* .

Xie W, Nelson BL, Barton RR (2014) A Bayesian framework for quantifying uncertainty in stochastic simulation. *Operations Research* 62(6):1439–1452.

Xie W, Nelson BL, Barton RR (2016) Multivariate input uncertainty in output analysis for stochastic simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 27(1):1–22.

Xu H, Caramanis C, Mannor S (2012) Optimization under probabilistic envelope constraints. *Operations Research* 60(3):682–699.

Xu H, Mannor S (2012) Distributionally robust markov decision processes. *Mathematics of Operations Research* 37(2):288–300.

Yi Y, Xie W (2017) An efficient budget allocation approach for quantifying the impact of input uncertainty in stochastic simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 27(4):1–23.

Zhu H, Liu T, Zhou E (2020) Risk quantification in stochastic simulation under input uncertainty. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 30(1):1–24.

Zouaoui F, Wilson JR (2003) Accounting for parameter uncertainty in simulation input modeling. *IIE Transactions* 35(9):781–792.

Zouaoui F, Wilson JR (2004) Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions* 36(11):1135–1151.

Appendix A

Technical Proofs for Chapter 2

This chapter presents proofs for results in Chapter 2. We first verify the proposed assumptions for the special case of finite-horizon performance measures in Section A.1. Section A.2 then proves results on the validity of the input variance decomposition (2.2). Section A.3 proves the consistency of the proposed input variance estimate and analyzes its Monte Carlo error in relation to the parameters B, R. Lastly, Section A.4 further analyzes the statistical error to obtain the overall error of the input variance estimate, and derives the optimal choices for θ, B, R that minimizes the overall error. In all the proofs, we write $a \approx b$ to mean $a/b \xrightarrow{p} 1$.

A.1 Finite-Horizon Performance Measures

In this section, we show that Assumptions 2.4.2-2.4.7 and 2.4.10-2.4.12 hold for the finite-horizon performance measure (2.16), thereby proving Theorems 2.4.2 and 2.4.7. We first prove Assumptions 2.4.2 and 2.4.10, then present the useful Lemma A.1.1 which will later be used to prove all other assumptions.

Proof of Assumptions 2.4.2 and 2.4.10. The finite horizon structure allows the following ex-

pansion of the performance measure $\psi(P_1^{\nu_1}, \ldots, P_m^{\nu_m})$ around the input models $P_1, \ldots, P_m$

$$
\begin{aligned}
&\psi(P_1^{\nu_1}, \ldots, P_m^{\nu_m}) \\
=&\int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i=1}^{m} \prod_{t=1}^{T_i} d(\nu_i(Q_i - P_i) + P_i)(x_{i,t}) \\
=&\psi(P_1, \ldots, P_m) + \sum_{d=1}^{T} \sum_{\sum_{i=1}^{m} |\mathcal{T}_i| = d} \prod_{i=1}^{m} \nu_i^{|\mathcal{T}_i|} \int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i=1}^{m} \prod_{t \notin \mathcal{T}_i} dP_i(x_{i,t}) \prod_{i=1}^{m} \prod_{t \in \mathcal{T}_i} d(Q_i - P_i)(x_{i,t}) \\
=&\psi(P_1, \ldots, P_m) + \sum_{d=1}^{T} \sum_{\sum_{i=1}^{m} |\mathcal{T}_i| = d} \prod_{i=1}^{m} \nu_i^{|\mathcal{T}_i|} \int h_{\mathcal{T}_1, \ldots, \mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1}, \ldots, \mathbf{x}_{m,\mathcal{T}_m}) \prod_{i=1}^{m} \prod_{t \in \mathcal{T}_i} d(Q_i - P_i)(x_{i,t})
\end{aligned}
$$

where $T = \sum_{i=1}^{m} T_i$ is the total run length, each $\mathcal{T}_i = \{\mathcal{T}_i(1), \ldots, \mathcal{T}_i(|\mathcal{T}_i|)\}$ is an ordered subset of $\{1, 2, \ldots, T_i\}$, and

$$
h_{\mathcal{T}_1, \ldots, \mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1}, \ldots, \mathbf{x}_{m,\mathcal{T}_m}) = \mathbb{E}_{P_1, \ldots, P_m}[h(\mathbf{X}_1, \ldots, \mathbf{X}_m) | X_i(t) = x_{i,t} \text{ for } i, t \in \mathcal{T}_i]. \tag{A.1}
$$

Here each $\mathbf{x}_{i,\mathcal{T}_i} := (x_{i,t})_{t \in \mathcal{T}_i}$. Expressing terms with $d = 1, 2, 3$ in a more explicit form gives

$$
\begin{aligned}
&\psi(P_1^{\nu_1}, \ldots, P_m^{\nu_m}) \\
=&\psi(P_1, \ldots, P_m) + \sum_{i=1}^{m} \nu_i \int \tilde{g}_i(x) d(Q_i - P_i)(x) \\
&+ \sum_{i_1 \le i_2} \nu_{i_1} \nu_{i_2} \int \tilde{g}_{i_1 i_2}(x, y) d(Q_{i_1} - P_{i_1})(x) d(Q_{i_2} - P_{i_2})(y) \\
&+ \sum_{i_1 \le i_2 \le i_3} \nu_{i_1} \nu_{i_2} \nu_{i_3} \int \tilde{g}_{i_1 i_2 i_3}(x, y, z) d(Q_{i_1} - P_{i_1})(x) d(Q_{i_2} - P_{i_2})(y) d(Q_{i_3} - P_{i_3})(z) \\
&+ \sum_{d=4}^{T} \sum_{\sum_{i=1}^{m} |\mathcal{T}_i| = d} \prod_{i=1}^{m} \nu_i^{|\mathcal{T}_i|} \int h_{\mathcal{T}_1, \ldots, \mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1}, \ldots, \mathbf{x}_{m,\mathcal{T}_m}) \prod_{i=1}^{m} \prod_{t \in \mathcal{T}_i} d(Q_i - P_i)(x_{i,t}).
\end{aligned} \tag{A.2}
$$

where

$$\tilde{g}_i(x) = \sum_{1\le t\le T_i} \mathbb{E}_{P_1,\ldots,P_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=x]$$

$$\tilde{g}_{i_1i_2}(x,y) = \begin{cases} \sum_{1\le t_1<t_2\le T_i} \mathbb{E}_{P_1,\ldots,P_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t_1)=x, X_i(t_2)=y], & \text{if } i_1=i_2=i \\ \sum_{t_1=1}^{T_{i_1}} \sum_{t_2=1}^{T_{i_2}} \mathbb{E}_{P_1,\ldots,P_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_{i_1}(t_1)=x, X_{i_2}(t_2)=y], & \text{if } i_1<i_2 \end{cases}$$

$$\tilde{g}_{i_1i_2i_3}(x,y,z)$$

$$= \begin{cases} \sum_{1\le t_1<t_2<t_3\le T_i} \mathbb{E}_{P_1,\ldots,P_m}[h|X_i(t_1)=x, X_i(t_2)=y, X_i(t_3)=z], & \text{if } i_1=i_2=i_3=i \\ \sum_{1\le t_1<t_2\le T_i} \sum_{t_3=1}^{T_{i_3}} \mathbb{E}_{P_1,\ldots,P_m}[h|X_i(t_1)=x, X_i(t_2)=y, X_{i_3}(t_3)=z], & \text{if } i_1=i_2=i<i_3 \\ \sum_{t_1=1}^{T_{i_1}} \sum_{1\le t_2<t_3\le T_i} \mathbb{E}_{P_1,\ldots,P_m}[h|X_{i_1}(t_1)=x, X_i(t_2)=y, X_i(t_3)=z], & \text{if } i_1<i_2=i_3=i \\ \sum_{t_1=1}^{T_{i_1}} \sum_{t_2=1}^{T_{i_2}} \sum_{t_3=1}^{T_{i_3}} \mathbb{E}_{P_1,\ldots,P_m}[h|X_{i_1}(t_1)=x, X_{i_2}(t_2)=y, X_{i_3}(t_3)=z], & \text{if } i_1<i_2<i_3 \end{cases}.$$

Since each signed measure $Q_i - P_i$ in the product measure in (A.2) has zero total measure, adding to the integrand a function that is independent of at least one of the integration variables does not change the integral value. Hence one can replace $\tilde{g}$'s by the following centered versions for $i_1 \le i_2 \le i_3$

$$\begin{aligned} \tilde{g}^c_i(x) &= \tilde{g}_i(x) - \mathbb{E}[\tilde{g}_i(X_i)] \\ \tilde{g}^c_{i_1i_2}(x,y) &= \tilde{g}_{i_1i_2}(x,y) - \mathbb{E}[\tilde{g}_{i_1i_2}(X_{i_1},y)] - \mathbb{E}[\tilde{g}_{i_1i_2}(x,X_{i_2})] + \mathbb{E}[\tilde{g}_{i_1i_2}(X_{i_1},X'_{i_2})] \\ \tilde{g}^c_{i_1i_2i_3}(x,y,z) &= \tilde{g}_{i_1i_2i_3}(x,y,z) - \mathbb{E}[\tilde{g}_{i_1i_2i_3}(X_{i_1},y,z)] - \mathbb{E}[\tilde{g}_{i_1i_2i_3}(x,X_{i_2},z)] - \mathbb{E}[\tilde{g}_{i_1i_2i_3}(x,y,X_{i_3})] \\ &\quad + \mathbb{E}[\tilde{g}_{i_1i_2i_3}(X_{i_1},X'_{i_2},z)] + \mathbb{E}[\tilde{g}_{i_1i_2i_3}(X_{i_1},y,X'_{i_3})] + \mathbb{E}[\tilde{g}_{i_1i_2i_3}(x,X_{i_2},X'_{i_3})] \\ &\quad - \mathbb{E}[\tilde{g}_{i_1i_2i_3}(X_{i_1},X'_{i_2},X''_{i_3})] \end{aligned}$$

where X_i, X'_i, X''_i denote independent variables distributed under F_i, and replace the function

$h_{\mathcal{T}_1,\ldots,\mathcal{T}_m}$ by

$$\begin{aligned}
&h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{m,\mathcal{T}_m})\\
=&h_{\mathcal{T}_1,\ldots,\mathcal{T}_m} - \sum_{i,t\in\mathcal{T}_i}\int h_{\mathcal{T}_1,\ldots,\mathcal{T}_m}dP_i(x_{i,t}) + \sum_{(i_1,t_1)<(i_2,t_2),t_1\in\mathcal{T}_{i_1},t_2\in\mathcal{T}_{i_2}}\int h_{\mathcal{T}_1,\ldots,\mathcal{T}_m}dP_{i_1}(x_{i_1,t_1})dP_{i_2}(x_{i_2,t_2})+\\
&\cdots+(-1)^{\sum_{i=1}^m|\mathcal{T}_i|}\int h_{\mathcal{T}_1,\ldots,\mathcal{T}_m}\prod_{i=1}^m\prod_{t\in\mathcal{T}_i}dP_i(x_{i,t}) \qquad \text{(A.3)}
\end{aligned}$$

where the order $(i_1,t_1)<(i_2,t_2)$ is defined as either $i_1<i_2$, or $i_1=i_2$ but $t_1<t_2$. This leads to the new Taylor expansion

$$\begin{aligned}
&\psi(P_1^{\nu_1},\ldots,P_m^{\nu_m})\\
=&\psi(P_1,\ldots,P_m)+\sum_{i=1}^m\nu_i\int\tilde{g}^c_i(x)d(Q_i-P_i)(x)\\
&+\sum_{i_1\le i_2}\nu_{i_1}\nu_{i_2}\int\tilde{g}^c_{i_1i_2}(x,y)d(Q_{i_1}-P_{i_1})(x)d(Q_{i_2}-P_{i_2})(y)\\
&+\sum_{i_1\le i_2\le i_3}\nu_{i_1}\nu_{i_2}\nu_{i_3}\int\tilde{g}^c_{i_1i_2i_3}(x,y,z)d(Q_{i_1}-P_{i_1})(x)d(Q_{i_2}-P_{i_2})(y)d(Q_{i_3}-P_{i_3})(z)\\
&+\sum_{d=4}^T\sum_{\sum_{i=1}^m|\mathcal{T}_i|=d}\prod_{i=1}^m\nu_i^{|\mathcal{T}_i|}\int h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{m,\mathcal{T}_m})\prod_{i=1}^m\prod_{t\in\mathcal{T}_i}d(Q_i-P_i)(x_{i,t}). \qquad \text{(A.4)}
\end{aligned}$$

Note that now all the integrands above have zero marginal means due to centering, e.g.

$$\int h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{m,\mathcal{T}_m})dP_i(x_{i,t})=0 \text{ for all } i \text{ and } t\in\mathcal{T}_i. \qquad \text{(A.5)}$$

However, the functions $\tilde{g}^c_i,\tilde{g}^c_{i_1i_2},\tilde{g}^c_{i_1i_2i_3}$ are not necessarily symmetric under permutations as required

in Assumption 2.4.10, so we perform the following symmetrization to find the influence functions

$$
\begin{aligned}
g_i(x) &:= \tilde{g}_i^c(x),\\
g_{ii}(x_1,x_2) &:= \tilde{g}_{ii}^c(x_1,x_2)+\tilde{g}_{ii}^c(x_2,x_1),\\
g_{i_1i_2}(x_1,x_2)=g_{i_2i_1}(x_2,x_1) &:= \tilde{g}_{i_1i_2}^c(x_1,x_2) \text{ for } i_1<i_2,\\
g_{iii}(x_1,x_2,x_3) &:= \sum_{\pi}\tilde{g}_{iii}^c(x_{\pi(1)},x_{\pi(2)},x_{\pi(3)}),\\
g_{i_1i_1i_2}(x_1,x_2,x_3)=g_{i_1i_2i_1}(x_1,x_3,x_2)&=g_{i_2i_1i_1}(x_3,x_1,x_2)\\
&:= \tilde{g}_{i_1i_1i_2}^c(x_1,x_2,x_3)+\tilde{g}_{i_1i_1i_2}^c(x_2,x_1,x_3) \text{ for } i_1<i_2,\\
g_{i_1i_2i_2}(x_1,x_2,x_3)=g_{i_2i_1i_2}(x_2,x_1,x_3)&=g_{i_2i_2i_1}(x_2,x_3,x_1)\\
&:= \tilde{g}_{i_1i_2i_2}^c(x_1,x_2,x_3)+\tilde{g}_{i_1i_2i_2}^c(x_1,x_3,x_2) \text{ for } i_1<i_2,\\
\text{for all } \pi \text{ let } g_{i_{\pi(1)}i_{\pi(2)}i_{\pi(3)}}(x_{\pi(1)},x_{\pi(2)},x_{\pi(3)}) &:= \tilde{g}_{i_1i_2i_3}^c(x_1,x_2,x_3) \text{ for } i_1<i_2<i_3,
\end{aligned}
$$

where the dependence on $P_1,\ldots,P_m$ is suppressed and π denotes any permutation of $(1,2,3)$. Then one can check that $g_{i_1i_2}$ and $g_{i_1i_2i_3}$ not only retain the property of zero marginal means, but also satisfy the symmetry condition in Assumption 2.4.10. Permutation symmetry implies that

$$
\begin{aligned}
&\psi(P_1^{\nu_1},\ldots,P_m^{\nu_m})\\
=&\psi(P_1,\ldots,P_m)+\sum_{i=1}^m\nu_i\int g_i(x)d(Q_i-P_i)(x)\\
&+\frac{1}{2}\sum_{i_1,i_2}\nu_{i_1}\nu_{i_2}\int g_{i_1i_2}(x,y)d(Q_{i_1}-P_{i_1})(x)d(Q_{i_2}-P_{i_2})(y)\\
&+\frac{1}{6}\sum_{i_1,i_2,i_3}\nu_{i_1}\nu_{i_2}\nu_{i_3}\int g_{i_1i_2i_3}(x,y,z)d(Q_{i_1}-P_{i_1})(x)d(Q_{i_2}-P_{i_2})(y)d(Q_{i_3}-P_{i_3})(z)\\
&+\sum_{d=4}^T\sum_{\sum_{i=1}^m|\mathcal{T}_i|=d}\prod_{i=1}^m\nu_i^{|\mathcal{T}_i|}\int h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{m,\mathcal{T}_m})\prod_{i=1}^m\prod_{t\in\mathcal{T}_i}d(Q_i-P_i)(x_{i,t}). \qquad \text{(A.6)}
\end{aligned}
$$

Since the integrals are all finite under Assumption 2.4.9, the first-order and third-order remainders of the above expansion are respectively of order $O\big(\sum_{i=1}^m\nu_i^2\big)$ and $O\big(\big(\sum_{i=1}^m\nu_i^2\big)^2\big)$, leading to Assumptions 2.4.2 and 2.4.10. □

We continue to verify other assumptions, for which we use the following lemma.

Lemma A.1.1 *Suppose Assumption 2.4.9 holds with positive and even k. For each i let $\widetilde{F}_i \in \{F_i, \widehat{F}_i\}$ be either the i-th true or empirical input model. Then the following bounds hold uniformly for every $(\widetilde{F}_1, \dots, \widetilde{F}_m) \in \prod_{i=1}^m \{F_i, \widehat{F}_i\}$ and arbitrary input data size n_i*

$$\max_{I_1,\dots,I_m} \mathbb{E}_{\widehat{F}_1,\dots,\widehat{F}_m}[h^k(\mathbf{X}_{1,I_1}, \dots, \mathbf{X}_{m,I_m})] = O_p(1) \tag{A.7}$$

$$\mathbb{E}_{F_1,\dots,F_m}\big[\big(\psi(\widetilde{F}_1, \dots, \widetilde{F}_m) - \psi(F_1, \dots, F_m)\big)^k\big] \le C_1\mathcal{M}\big(\sum_{i=1}^m \frac{1}{\sqrt{n_i}}\big)^k \tag{A.8}$$

$$\mathbb{E}_{F_1,\dots,F_m}\big[\big(\psi(\widetilde{F}_1, \dots, \widetilde{F}_m) - \psi(F_1, \dots, F_m) - \sum_{i=1}^m \int g_i(x) d(\widetilde{F}_i - F_i)(x)\big)^k\big] \le C_2\mathcal{M}\big(\sum_{i=1}^m \frac{1}{\sqrt{n_i}}\big)^{2k} \tag{A.9}$$

where the influence functions g_i's are now under the true input models $F_1, \dots, F_m$. Each empirical influence function $\hat{g}_i$ satisfies

$$\mathbb{E}_{F_1,\dots,F_m}[(g_i(X_{i,1}) - \hat{g}_i(X_{i,1}))^k] \le C_3\mathcal{M}\big(\sum_{i=1}^m \frac{1}{\sqrt{n_i}}\big)^k \tag{A.10}$$

$$\begin{aligned}&\mathbb{E}_{F_1,\dots,F_m}[(\hat{g}_i(X_{i,1}) - g_i(X_{i,1}) - \sum_{i'=1}^m \int g_{ii'}(X_{i,1}, x) d(\widehat{F}_{i'} - F_{i'})(x) + \int g_i(x) d(\widehat{F}_i - F_i)(x))^k] \\ \le& C_4\mathcal{M}\big(\sum_{i=1}^m \frac{1}{\sqrt{n_i}}\big)^{2k}\end{aligned} \tag{A.11}$$

Here C_1, C_2, C_3, C_4 are constants that only depend on k, m and $T := \sum_{i=1}^m T_i$, and

$$\mathcal{M} := \max_{I_1,\dots,I_m} \mathbb{E}_{F_1,\dots,F_m}[h^k(\mathbf{X}_{1,I_1}, \dots, \mathbf{X}_{m,I_m})] < \infty.$$

Proof. The first bound is the most straightforward. By rewriting the expectation $\mathbb{E}_{\widehat{F}_1,\dots,\widehat{F}_m}[\cdot]$ as a sum, one can see that for a particular choice of $I_1, \dots, I_m$

$$\mathbb{E}_{F_1,\dots,F_m}\Big[\mathbb{E}_{\widehat{F}_1,\dots,\widehat{F}_m}[h^k(\mathbf{X}_{1,I_1}, \dots, \mathbf{X}_{m,I_m})]\Big] \le \mathcal{M}.$$

Therefore $\mathbb{E}_{\widehat{F}_1,\dots,\widehat{F}_m}[h^k(\mathbf{X}_{1,I_1}, \dots, \mathbf{X}_{m,I_m})] = O_p(1)$ for each $I_1, \dots, I_m$. Since there are finitely many

of them, the maximum is also bounded in probability. This proves the first bound.

To explain the other bounds, we put $\psi(\widetilde{F}_1, \ldots, \widetilde{F}_m)$ in the form of the expansion (A.4) with $\nu_i = 1, P_i = F_i, Q_i = \widetilde{F}_i$ to get

$$
\begin{aligned}
&\psi(\widetilde{F}_1, \ldots, \widetilde{F}_m) \\
=&\psi(F_1, \ldots, F_m) + \sum_{d=1}^{T} \sum_{\sum_{i=1}^m |\mathcal{T}_i| = d} \int h^c_{\mathcal{T}_1, \ldots, \mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1}, \ldots, \mathbf{x}_{m,\mathcal{T}_m}) \prod_{i=1}^m \prod_{t \in \mathcal{T}_i} d\widetilde{F}_i(x_{i,t}) \\
=&\psi(F_1, \ldots, F_m) + \sum_{i=1}^{m} \int \Big(\sum_{t=1}^{T_i} \mathbb{E}_{F_1, \ldots, F_m}[h(\mathbf{X}_1, \ldots, \mathbf{X}_m) | X_i(t) = x] - T_i \psi(F_1, \ldots, F_m) \Big) d(\widetilde{F}_i - F_i)(x) \\
&+ \sum_{d=2}^{T} \sum_{\sum_{i=1}^m |\mathcal{T}_i| = d} \int h^c_{\mathcal{T}_1, \ldots, \mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1}, \ldots, \mathbf{x}_{m,\mathcal{T}_m}) \prod_{i=1}^m \prod_{t \in \mathcal{T}_i} d\widetilde{F}_i(x_{i,t})
\end{aligned}
$$

where $\int h^c_{\mathcal{T}_1, \ldots, \mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1}, \ldots, \mathbf{x}_{m,\mathcal{T}_m}) dF_i(x_{i,t}) = 0$ for all i and $t \in \mathcal{T}_i$, according to the property of zero marginal means (A.5). To obtain a moment bound for $h^c_{\mathcal{T}_1, \ldots, \mathcal{T}_m}$, observe that by Assumption 2.4.9 and Jensen's inequality any conditional expectation of the performance function h has a k-th moment at most $\mathcal{M}$. Since $h^c_{\mathcal{T}_1, \ldots, \mathcal{T}_m}$ is the sum of several conditional expectations of h, one can apply Minkowski inequality to establish that for any $I_i = (I_i(1), \ldots, I_i(|\mathcal{T}_i|)) \in \{1, 2, \ldots, |\mathcal{T}_i|\}^{|\mathcal{T}_i|}$, $i = 1, \ldots, m$

$$
\mathbb{E}_{F_1, \ldots, F_m}[(h^c_{\mathcal{T}_1, \ldots, \mathcal{T}_m}(\mathbf{X}_{1,\mathcal{T}_1(I_1)}, \ldots, \mathbf{X}_{m,\mathcal{T}_m(I_m)}))^k] \leq 2^{k \sum_{i=1}^m |\mathcal{T}_i|} \mathcal{M}. \tag{A.12}
$$

Again by Minkowski inequality

$$
\begin{aligned}
&\mathbb{E}_{F_1,\ldots,F_m}\big[\big(\psi(\widetilde{F}_1,\ldots,\widetilde{F}_m)-\psi(F_1,\ldots,F_m)\big)^k\big]\\
\le&\Big(\sum_{d=1}^{T}\sum_{\sum_{i=1}^m|\mathcal{T}_i|=d}\Big(\mathbb{E}_{F_1,\ldots,F_m}\Big[\Big(\int h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{m,\mathcal{T}_m})\prod_{i=1}^m\prod_{t\in\mathcal{T}_i}d\widetilde{F}_i(x_{i,t})\Big)^k\Big]\Big)^{\frac{1}{k}}\Big)^k\\
=&\Big(\sum_{d=1}^{T}\sum_{\sum_{i=1}^m|\mathcal{T}_i|=d}\prod_{i=1}^m\prod_{t\in\mathcal{T}_i}\mathbf{1}(\widetilde{F}_{i,t}=\widehat{F}_i)\cdot\\
&\qquad\Big(\mathbb{E}_{F_1,\ldots,F_m}\Big[\Big(\frac{1}{\prod_{i=1}^m n_i^{|\mathcal{T}_i|}}\sum_{J_1,\ldots,J_m}h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{X}_{1,J_1},\ldots,\mathbf{X}_{m,J_m})\Big)^k\Big]\Big)^{\frac{1}{k}}\Big)^k\\
\le&\Big(\sum_{d=1}^{T}\sum_{\sum_{i=1}^m|\mathcal{T}_i|=d}\Big(\mathbb{E}_{F_1,\ldots,F_m}\Big[\Big(\frac{1}{\prod_{i=1}^m n_i^{|\mathcal{T}_i|}}\sum_{J_1,\ldots,J_m}h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{X}_{1,J_1},\ldots,\mathbf{X}_{m,J_m})\Big)^k\Big]\Big)^{\frac{1}{k}}\Big)^k
\end{aligned}
$$

where each $J_i=(J_i(1),\ldots,J_i(|\mathcal{T}_i|))\in\{1,2,\ldots,n_i\}^{|\mathcal{T}_i|}$ and $\mathbf{X}_{i,J_i}=(X_{i,J_i(1)},\ldots,X_{i,J_i(|\mathcal{T}_i|)})$. Note that

$$
\begin{aligned}
&\mathbb{E}_{F_1,\ldots,F_m}\Big[\Big(\frac{1}{\prod_{i=1}^m n_i^{|\mathcal{T}_i|}}\sum_{J_1,\ldots,J_m}h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{X}_{1,J_1},\ldots,\mathbf{X}_{m,J_m})\Big)^k\Big]\\
=&\frac{1}{\prod_{i=1}^m n_i^{k|\mathcal{T}_i|}}\sum_{J_1^1,\ldots,J_m^1}\cdots\sum_{J_1^k,\ldots,J_m^k}\mathbb{E}_{F_1,\ldots,F_m}[h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{X}_{1,J_1^1},\ldots,\mathbf{X}_{m,J_m^1})\cdots h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{X}_{1,J_1^k},\ldots,\mathbf{X}_{m,J_m^k})].
\end{aligned}
$$

By (A.5) the expectation on the right hand side is zero if some data point $X_{i,j}$ appears only once. Therefore the number of non-zero expectations is bounded above by $C(k,m,\sum_{i=1}^m|\mathcal{T}_i|)\prod_{i=1}^m n_i^{k|\mathcal{T}_i|/2}$, where $C(k,m,\sum_{i=1}^m|\mathcal{T}_i|)$ is some constant that only depends on $k,m,\sum_{i=1}^m|\mathcal{T}_i|$. Moreover, from (A.12) each expectation satisfies the following by generalized Hölder's inequality

$$
|\mathbb{E}_{F_1,\ldots,F_m}[h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{X}_{1,J_1^1},\ldots,\mathbf{X}_{m,J_m^1})\cdots h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{X}_{1,J_1^k},\ldots,\mathbf{X}_{m,J_m^k})]|\le 2^{k\sum_{i=1}^m|\mathcal{T}_i|}\mathcal{M}.
$$

Hence

$$\begin{aligned}
&\mathbb{E}_{F_1,\ldots,F_m}\big[\big(\psi(\widehat{F}_1,\ldots,\widehat{F}_m)-\psi(F_1,\ldots,F_m)\big)^k\big]\\
\leq&\Big(\sum_{d=1}^{T}\sum_{\sum_{i=1}^m|\mathcal{T}_i|=d}\Big(C(k,m,\sum_{i=1}^m|\mathcal{T}_i|)\prod_{i=1}^m n_i^{-k|\mathcal{T}_i|/2}2^{k\sum_{i=1}^m|\mathcal{T}_i|}\mathcal{M}\Big)^{\frac{1}{k}}\Big)^k\\
=&\Big(\sum_{d=1}^{T}\sum_{\sum_{i=1}^m|\mathcal{T}_i|=d}C'(k,m,d)\prod_{i=1}^m n_i^{-|\mathcal{T}_i|/2}\mathcal{M}^{\frac{1}{k}}\Big)^k\\
\leq&\Big(\sum_{d=1}^{T}C'(k,m,d)\big(\sum_{i=1}^m\frac{T_i}{\sqrt{n_i}}\big)^d\Big)^k\mathcal{M}\leq C_1(k,m,T)\mathcal{M}\Big(\sum_{i=1}^m\frac{1}{\sqrt{n_i}}\Big)^k.
\end{aligned}$$

This gives the second bound.

The third bound can be established by the same argument, but considering only the remainders for which $d\geq 2$.

We then prove the bounds on influence functions. According to the expression of $g_i(P_1,\ldots,P_m;\cdot)$ in the proof of Assumptions 2.4.2 and 2.4.10, the empirical influence function $\hat{g}_i$ is

$$\hat{g}_i(x)=\sum_{t=1}^{T_i}\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=x]-T_i\psi(\widehat{F}_1,\ldots,\widehat{F}_m).$$

First we derive the following Taylor expansion for each conditional expectation

$$
\begin{aligned}
&\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=X_{i,1}] \\
=&\int h(\mathbf{x}_1,\ldots,\mathbf{x}_m)\prod_{t'\neq t} d\widehat{F}_i(x_{i,t'})\prod_{i'\neq i}\prod_{t'=1}^{T_{i'}} d\widehat{F}_{i'}(x_{i',t'})\Big|_{x_{i,t}=X_{i,1}} \\
=&\mathbb{E}_{F_1,\ldots,F_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=X_{i,1}] \\
&+\sum_{d=1}^{T-1}\sum_{\sum_{i'=1}^m|\mathcal{T}_{i'}|=d,t\notin\mathcal{T}_i}\int h(\mathbf{x}_1,\ldots,\mathbf{x}_m)\prod_{t'\notin\mathcal{T}_i,t'\neq t} dF_i(x_{i,t'})\prod_{i'\neq i}\prod_{t'\notin\mathcal{T}_{i'}} dF_{i'}(x_{i',t'}) \\
&\qquad\prod_{i'=1}^m\prod_{t'\in\mathcal{T}_{i'}} d(\widehat{F}_{i'}-F_{i'})(x_{i',t'})\Big|_{x_{i,t}=X_{i,1}} \\
=&\mathbb{E}_{F_1,\ldots,F_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=X_{i,1}] \\
&+\sum_{d=1}^{T-1}\sum_{\sum_{i'=1}^m|\mathcal{T}_{i'}|=d,t\notin\mathcal{T}_i}\int h_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{i,\mathcal{T}_i\cup\{t\}},\ldots,\mathbf{x}_{m,\mathcal{T}_m}) \\
&\qquad\prod_{i'=1}^m\prod_{t'\in\mathcal{T}_{i'}} d(\widehat{F}_{i'}-F_{i'})(x_{i',t'})\Big|_{x_{i,t}=X_{i,1}} \\
=&\mathbb{E}_{F_1,\ldots,F_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=X_{i,1}] \\
&+\sum_{t'=1,t'\neq t}^{T_i}\int \mathbb{E}_{F_1,\ldots,F_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=X_{i,1},X_i(t')=x_{i,t'}]d(\widehat{F}_i-F_i)(x_{i,t'}) \qquad \text{(A.13)} \\
&+\sum_{i'\neq i}\sum_{t'=1}^{T_{i'}}\int \mathbb{E}_{F_1,\ldots,F_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=X_{i,1},X_{i'}(t')=x_{i',t'}]d(\widehat{F}_{i'}-F_{i'})(x_{i',t'}) \qquad \text{(A.14)} \\
&+\sum_{d=2}^{T-1}\sum_{\sum_{i'=1}^m|\mathcal{T}_{i'}|=d,t\notin\mathcal{T}_i}\int h_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{i,\mathcal{T}_i\cup\{t\}},\ldots,\mathbf{x}_{m,\mathcal{T}_m}) \\
&\qquad\prod_{i'=1}^m\prod_{t'\in\mathcal{T}_{i'}} d(\widehat{F}_{i'}-F_{i'})(x_{i',t'})\Big|_{x_{i,t}=X_{i,1}}
\end{aligned}
$$

where each $\mathcal{T}_{i'}=\{\mathcal{T}_{i'}(1),\ldots,\mathcal{T}_{i'}(|\mathcal{T}_{i'}|)\}$ is still an ordered subset of $\{1,2,\ldots,T_{i'}\}$ but $t\notin\mathcal{T}_i$, and the function $h_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m}$ resembles (A.1) except that the expectation is now further conditioned

on $X_i(t) = x_{i,t}$. Introduce the counterpart of (A.3)

$$\begin{aligned}
&h^c_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{i,\mathcal{T}_i\cup\{t\}},\ldots,\mathbf{x}_{m,\mathcal{T}_m})\\
=&h_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m} - \sum_{i',t'\in\mathcal{T}_{i'}}\int h_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m} dF_{i'}(x_{i',t'})\\
&+\sum_{(i'_1,t'_1)<(i'_2,t'_2),t'_1\in\mathcal{T}_{i'_1},t'_2\in\mathcal{T}_{i'_2}}\int h_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m} dF_{i'_1}(x_{i'_1,t'_1})dF_{i'_2}(x_{i'_2,t'_2})\\
&+\cdots+(-1)^{\sum_{i'=1}^m|\mathcal{T}_{i'}|}\int h_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m}\prod_{i'=1}^m\prod_{t'\in\mathcal{T}_{i'}} dF_{i'}(x_{i',t'})
\end{aligned}$$

then we have the following parallel property of (A.5)

$$\int h^c_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{i,\mathcal{T}_i\cup\{t\}},\ldots,\mathbf{x}_{m,\mathcal{T}_m})dF_{i'}(x_{i',t'}) = 0 \text{ for all } i' \text{ and } t'\in\mathcal{T}_{i'}$$

and by comparing the first order remainders (A.13) and (A.14) of $\hat{g}_i$ with the second order influence functions $g_{i_1 i_2}$ it is easy to establish that

$$\begin{aligned}
&\hat{g}_i(X_{i,1}) - g_i(X_{i,1})\\
=&\sum_{t=1}^{T_i}\big(\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=X_{i,1}] - \mathbb{E}_{F_1,\ldots,F_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)=X_{i,1}]\big)\\
&-T_i(\psi(\widehat{F}_1,\ldots,\widehat{F}_m)-\psi(F_1,\ldots,F_m))\\
=&\sum_{i'=1}^m\int g_{ii'}(X_{i,1},x)d(\widehat{F}_{i'}-F_{i'})(x) - \int g_i(x)d(\widehat{F}_i-F_i)(x) \qquad \text{(A.15)}\\
&+\sum_{t=1}^{T_i}\sum_{d=2}^{T-1}\sum_{\sum_{i'=1}^m|\mathcal{T}_{i'}|=d}\int h^c_{(i,t),\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{i,\mathcal{T}_i\cup\{t\}},\ldots,\mathbf{x}_{m,\mathcal{T}_m})\prod_{i'=1}^m\prod_{t'\in\mathcal{T}_{i'}} d\widehat{F}_{i'}(x_{i',t'})\Big|_{x_{i,t}=X_{i,1}}\\
&-T_i\sum_{d=2}^{T}\sum_{\sum_{i'=1}^m|\mathcal{T}_{i'}|=d}\int h^c_{\mathcal{T}_1,\ldots,\mathcal{T}_m}(\mathbf{x}_{1,\mathcal{T}_1},\ldots,\mathbf{x}_{m,\mathcal{T}_m})\prod_{i'=1}^m\prod_{t'\in\mathcal{T}_{i'}} d\widehat{F}_{i'}(x_{i',t'}) \qquad \text{(A.16)}
\end{aligned}$$

By a similar technique used to bound the remainder of $\psi(\widetilde{F}_1,\ldots,\widetilde{F}_m)$, we can establish that the remainder (A.16) has a k-th moment of order $O\big(\mathcal{M}\big(\sum_{i=1}^m\frac{1}{\sqrt{n_i}}\big)^{2k}\big)$, and the first order term (A.15) has a k-th moment of order $O\big(\mathcal{M}\big(\sum_{i=1}^m\frac{1}{\sqrt{n_i}}\big)^{k}\big)$. This completes the proof. □

With Lemma A.1.1 we now prove the other assumptions:

Proof of Assumption 2.4.3. The moment bound on the remainder, i.e. $\mathbb{E}[\epsilon^2] = o(n^{-1})$, comes from the bound (A.9) in Lemma A.1.1 with $\widetilde{F}_i = \widehat{F}_i$ for all i and $k = 2$. The non-degeneracy condition on the influence functions is exactly Assumption 2.4.8, whereas the finiteness of fourth order moments of g_i easily follows because g_i is simply a sum of T_i conditional expectations of the performance function h and each of the conditional expectations has finite fourth order moment by Assumption 2.4.9 and Jensen's inequality. □

Proof of Assumption 2.4.4. The convergence of $\hat{g}_i$ to g_i in fourth order moment is a direct consequence of the bound (A.10) in Lemma A.1.1 with $k = 4$. The moment condition on the remainder ϵ^* can be argued as follows. We treat the empirical distributions $\widehat{F}_1, \ldots, \widehat{F}_m$ as the truth, and the resampled distributions $\widehat{F}^*_{s_1,1}, \ldots, \widehat{F}^*_{s_m,m}$ as the input data, then apply the third bound (A.9) in Lemma A.1.1 with $k = 4$ to get $\mathbb{E}_*[(\epsilon^*)^4] \leq C_2 \widehat{\mathcal{M}} \big(\sum_{i=1}^m \frac{1}{\sqrt{s_i}} \big)^8$, where $\widehat{\mathcal{M}} = \max_{I_1,\ldots,I_m} \mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h^4(\mathbf{X}_{1,I_1}, \ldots, \mathbf{X}_{m,I_m})]$ is $O_p(1)$ by the first bound (A.7) in Lemma A.1.1 with $k = 4$. Therefore $\mathbb{E}_*[(\epsilon^*)^4] = O_p((\sum_{i=1}^m \frac{1}{s_i})^4) = o_p(s^{-2})$. □

Proof of Assumption 2.4.5. It suffices to show that $\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h^2] \xrightarrow{p} \mathbb{E}_{F_1,\ldots,F_m}[h^2]$ and that $\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h] \xrightarrow{p} \mathbb{E}_{F_1,\ldots,F_m}[h]$. The latter convergence follows from the second bound (A.8) of Lemma A.1.1 with $k = 2$ and $\widetilde{F}_i = \widehat{F}_i$ for all i. Since Assumption 2.4.9 holds with $k = 4$ for the function h, it also holds with $k = 2$ for the squared function h^2. One can apply the same bound from Lemma A.1.1 with $k = 2$ to h^2 and then conclude the former convergence. □

Proof of Assumption 2.4.6. We write $\bar{\tau}^2 = \tau^2(\overline{F}_1, \ldots, \overline{F}_m)$ for short. First rewrite

$$
\begin{aligned}
(\bar{\tau}^2 - \hat{\tau}^2)^2 &= \big(\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h^2] - \mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h^2] - \big((\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h])^2 - (\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h])^2 \big) \big)^2 \\
&\leq 2 \big(\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h^2] - \mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h^2] \big)^2 + 2 \big((\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h])^2 - (\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h])^2 \big)^2 \\
&\leq 2 \big(\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h^2] - \mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h^2] \big)^2 + 4 \big(\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h] - \mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h] \big)^4 \\
&\quad + 16 (\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h])^2 \big(\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h] - \mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h] \big)^2.
\end{aligned}
$$

Applying Lemma A.1.1 to h^2 ($k=2$) with the true distributions being $\widehat{F}_1,\ldots,\widehat{F}_m$ we get

$$\mathbb{E}_*[\big(\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h^2]-\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h^2]\big)^2]\leq C_1\widehat{\mathcal{M}}\big(\sum_{i=1}^m\frac{1}{\sqrt{s_i}}\big)^2=O_p\big(\sum_{i=1}^m\frac{1}{s_i}\big)$$

where $\widehat{\mathcal{M}}=\max_{I_1,\ldots,I_m}\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h^4(\mathbf{X}_{1,I_1},\ldots,\mathbf{X}_{m,I_m})]=O_p(1)$. Another application of Lemma A.1.1 to h with $k=4$ gives

$$\mathbb{E}_*[\big(\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h]-\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h]\big)^4]\leq C_1\widehat{\mathcal{M}}\big(\sum_{i=1}^m\frac{1}{\sqrt{s_i}}\big)^4=O_p\big(\sum_{i=1}^m\frac{1}{s_i^2}\big)$$

which implies that $\mathbb{E}_*[\big(\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h]-\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h]\big)^2]=O_p\big(\sum_{i=1}^m\frac{1}{s_i}\big)$ as a consequence of Cauchy Schwartz inequality. Therefore in sum $\mathbb{E}_*[(\bar{\tau}^2-\hat{\tau}^2)^2]=O_p\big(\sum_{i=1}^m\frac{1}{s_i}\big)=o_p(1)$. □

Proof of Assumption 2.4.7. Note that $\mu_4(\overline{F}_1,\ldots,\overline{F}_m)\leq C\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h^4]$ for some absolute constant $C>0$, therefore

$$\mathbb{E}_*[\mu_4(\overline{F}_1,\ldots,\overline{F}_m)]\leq C\mathbb{E}_*[\mathbb{E}_{\overline{F}_1,\ldots,\overline{F}_m}[h^4]]\leq C\max_{I_1,\ldots,I_m}\mathbb{E}_{\widehat{F}_1,\ldots,\widehat{F}_m}[h^4(\mathbf{X}_{1,I_1},\ldots,\mathbf{X}_{m,I_m})]=O_p(1)$$

where the last equality is due to the first bound (A.7) in Lemma A.1.1. □

Proof of Assumption 2.4.11. The third order remainder ϵ_3, or equivalently the sum over $d\geq 4$ in (A.6) with each $\nu_i=1$, consists of integrals under the product of at least four signed measures of the form $\widehat{F}_i-F_i$. Therefore, by employing the technique used in proving the second and third bounds (A.8)(A.9) in Lemma A.1.1, one can show that $\mathbb{E}[\epsilon_3^2]=O(n^{-4})$. The details are omitted since they highly resemble those of Lemma A.1.1. The fourth moments of $g_{i_1i_2}$ and $g_{i_1i_2i_3}$ are finite, because each of them is a finite sum of conditional expectations of h which have finite fourth order moments due to Assumption 2.4.9 with $k=4$ and Jensen's inequality. □

Proof of Assumption 2.4.12. For the third order remainder of the resampled performance measure, one can derive the bound $\mathbb{E}_*[(\epsilon_3^*)^2]=O_p(s^{-4})$ in a similar way as in showing the bound (A.9) in Lemma A.1.1. The details are omitted to avoid repetition. Moreover, some straightforward modifications of the proof for the bound (A.10) in Lemma A.1.1 lead to $O(n^{-1})$ upper bounds for the the mean squared errors of second and third order influence functions. The remainder in the

Taylor expansion of the first order empirical influence function satisfies $\mathbb{E}[\epsilon_g^2] = O(n^{-2})$ due to the bound (A.11) in Lemma A.1.1 with $k = 2$. □

A.2 Proofs of Propositions 2.4.1 and 2.4.6

This section proves results concerning the validity of the additive decomposition 2.2 of the input variance. We first prove Proposition 2.4.1, and then provide the key Lemma A.2.1 that will be used in the proof of Proposition 2.4.6 as well as many results in Section A.4. *Proof of Proposition 2.4.1.* Following the expansion (2.14) we can write

$$\begin{aligned}
\mathrm{Var}[\psi(\widehat{F}_1, \ldots, \widehat{F}_m)] &= \mathrm{Var}[\sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} g_i(X_{i,j})] + \mathrm{Var}[\epsilon] + 2\mathrm{Cov}(\sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} g_i(X_{i,j}), \epsilon) \\
&= \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} + o(n^{-1}) + O\Big(\sqrt{\mathrm{Var}[\sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} g_i(X_{i,j})]\mathrm{Var}[\epsilon]}\Big) \\
&= \sum_{i=1}^{m} \frac{\sigma_i^2}{n_i} + o(n^{-1}).
\end{aligned}$$

This completes the proof. □

The following important lemma on variance decomposition plays a crucial role in our analysis.

Lemma A.2.1 (Adapted from Efron and Stein (1981)) *Let $Y_i, i = 1, \ldots, n$ be independent but not necessarily identically distributed random variables, and $\phi(y_1, \ldots, y_n)$ be a function such that $\mathbb{E}[\phi^2(Y_1, \ldots, Y_n)] < \infty$, then there exist functions $\phi_{i_1,\ldots,i_k}$ for $1 \le i_1 < \cdots < i_k \le n$ and $k \le n$ such that*

$$\begin{aligned}
&\phi(Y_1, \ldots, Y_n) \\
=&\mu + \sum_{i=1}^{n} \phi_i(Y_i) + \sum_{i_1<i_2} \phi_{i_1,i_2}(Y_{i_1}, Y_{i_2}) + \cdots + \sum_{i_1<\cdots<i_k} \phi_{i_1,\ldots,i_k}(Y_{i_1}, \ldots, Y_{i_k}) + \cdots + \phi_{1,\ldots,n}(Y_1, \ldots, Y_n)
\end{aligned}$$

where

$$\begin{aligned}
\mu &= \mathbb{E}[\phi(Y_1, \ldots, Y_n)] \\
\phi_i(y) &= \mathbb{E}[\phi(Y_1, \ldots, Y_n)|Y_i = y] - \mu \\
\phi_{i_1,i_2}(y_1, y_2) &= \mathbb{E}[\phi(Y_1, \ldots, Y_n)|Y_{i_1} = y_1, Y_{i_2} = y_2] - \phi_{i_1}(y_1) - \phi_{i_2}(y_2) - \mu \\
&\vdots
\end{aligned}$$

Moreover, the $2^n - 1$ *random variables in the decomposition have mean zero and are mutually uncorrelated.*

With this lemma, we can prove Proposition 2.4.6: *Proof of Proposition 2.4.6.* The proof of Proposition 2.1 derives the following expression for input variance

$$\mathrm{Var}[\psi(\widehat{F}_1, \ldots, \widehat{F}_m)] = \mathrm{Var}[\sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} g_i(X_{i,j})] + \mathrm{Var}[\epsilon] + 2\mathrm{Cov}(\sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} g_i(X_{i,j}), \epsilon)$$

where the covariances can be simplified to

$$\begin{aligned}
\mathrm{Cov}(\sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} g_i(X_{i,j}), \epsilon) &= \sum_{i=1}^{m} \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbb{E}[g_i(X_{i,j})(\epsilon - \mathbb{E}[\epsilon])] \\
&= \sum_{i=1}^{m} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbb{E}[g_i(X_{i,j})(\mathbb{E}[\epsilon|X_{i,j}] - \mathbb{E}[\epsilon])] \\
&= \sum_{i=1}^{m} \mathbb{E}[g_i(X_{i,1})(\mathbb{E}[\epsilon|X_{i,1}] - \mathbb{E}[\epsilon])].
\end{aligned}$$

Using the cubic expansion in Assumption 2.4.11 and the vanishing marginal expectations of influence functions we have

$$\begin{aligned}
\mathbb{E}[\epsilon|X_{i,1}] - \mathbb{E}[\epsilon] =& \frac{1}{2n_i^2}(g_{ii}(X_{i,1}, X_{i,1}) - \mathbb{E}[g_{ii}(X_i, X_i)]) + \frac{1}{6n_i^3}(g_{iii}(X_{i,1}, X_{i,1}, X_{i,1}) - \mathbb{E}[g_{iii}(X_i, X_i, X_i)]) \\
&+ \frac{n_i - 1}{2n_i^3}\mathbb{E}_{X_i}[g_{iii}(X_{i,1}, X_i, X_i)] + \sum_{i' \neq i} \frac{1}{2n_i n_{i'}}\mathbb{E}_{X_{i'}}[g_{ii'i'}(X_{i,1}, X_{i'}, X_{i'})] \\
&+ \mathbb{E}[\epsilon_3|X_{i,1}] - \mathbb{E}[\epsilon_3]. \qquad \text{(A.17)}
\end{aligned}$$

Each term except the last in (A.17) has a second moment of order $O(n^{-4})$. To argue the last term $\mathbb{E}[\epsilon_3|X_{i,1}] - \mathbb{E}[\epsilon_3]$ also has a second moment of order at most $O(n^{-4})$, note that ϵ_3 is a symmetric statistic hence by Lemma A.2.1 $\text{Var}[\mathbb{E}[\epsilon_3|X_{i,1}]] \leq \text{Var}[\epsilon_3]/n_i$ and $\text{Var}[\epsilon_3] = o(n^{-3})$ by assumption, hence $\text{Var}[\mathbb{E}[\epsilon_3|X_{i,1}]] = o(n^{-4})$. This leads to

$$\text{Var}[\mathbb{E}[\epsilon|X_{i,1}]] = O(n^{-4}).$$

Using Cauchy Schwartz inequality we conclude $\text{Cov}(\sum_{i=1}^m \frac{1}{n_i}\sum_{j=1}^{n_i} g_i(X_{i,j}), \epsilon) = O(n^{-2})$. On the other hand, one can easily show $\text{Var}[\epsilon] = O(n^{-2})$ by using the same technique in the proof of Lemma A.1.1 to bound each term in the cubic expansion. This leads to the desired conclusion. □

A.3 Proofs for Results in Section 2.4.2 and Section 2.3.2

We now prove the consistency of our proportionate subsampled bootstrap variance σ^2_{SVB} (Theorem 2.4.3), and derive the mean square error of the Monte Carlo estimate $\hat{\sigma}^2_{SVB}$ relative to σ^2_{SVB} (Lemma 2.4.4). These results will then be used to prove Theorems 2.3.1 and 2.4.5. Theorem 2.3.4, Corollaries 2.3.2-2.3.3 are consequences of Theorem 2.3.1. Theorem 2.3.5 is a consequence of Theorem 2.4.5.

Recall that $\sigma_i^2 = \text{Var}_{F_i}[g_i(X_i)]$ is the variance of the i-th influence function. For its empirical counterpart $\hat{g}_i$ we denote by $\hat{\sigma}_i^2 := \text{Var}_{\widehat{F}_i}[\hat{g}_i(X_i)]$ its variance under the empirical input models. Under the convergence condition $\mathbb{E}[(\hat{g}_i - g_i)^4(X_{i,1})] \to 0$ in Assumption 2.4.4, the convergence of $\hat{\sigma}_i^2$ to σ_i^2 follows from

$$\begin{aligned}\left|\hat{\sigma}_i^2 - \frac{1}{n_i}\sum_{j=1}^{n_i} g_i^2(X_{i,j})\right| &= \left|\frac{1}{n_i}\sum_{j=1}^{n_i}\hat{g}_i^2(X_{i,j}) - \frac{1}{n_i}\sum_{j=1}^{n_i} g_i^2(X_{i,j})\right| \\ &\leq \frac{2}{n_i}\sqrt{\sum_{j=1}^{n_i} g_i^2(X_{i,j})\sum_{j=1}^{n_i}(\hat{g}_i - g_i)^2(X_{i,j})} + \frac{1}{n_i}\sum_{j=1}^{n_i}(\hat{g}_i - g_i)^2(X_{i,j}) = o_p(1)\end{aligned}$$

and that $\sum_{j=1}^{n_i} g_i^2(X_{i,j})/n_i \xrightarrow{p} \sigma_i^2$. For convenience we denote by

$$\psi^* = \psi(\widehat{F}^*_{s_1,1}, \ldots, \widehat{F}^*_{s_m,m}),\ \hat{\psi}^* = \hat{\psi}(\widehat{F}^*_{s_1,1}, \ldots, \widehat{F}^*_{s_m,m})$$

the expected value and a single simulation replication, respectively, of the performance measure under the resampled input models, and by

$$\hat{\tau}_*^2 = \tau^2(\widehat{F}^*_{s_1,1},\ldots,\widehat{F}^*_{s_m,m}),\ \hat{\mu}_4^* = \mu_4(\widehat{F}^*_{s_1,1},\ldots,\widehat{F}^*_{s_m,m})$$

the variance and central fourth moment of a single Monte Carlo replication $\hat{\psi}^*$ conditioned on the resampled input models.

Proof of Theorem 2.4.3. Let $s_i = \lfloor \theta n_i \rfloor$. Following the expansion (2.15) with each $\overline{F}_i = \widehat{F}^*_{s_i,i}$ we have

$$\begin{aligned}
\mathrm{Var}_*[\psi^*] &= \mathrm{Var}_*[\sum_{i=1}^m \frac{1}{s_i}\sum_{k=1}^{s_i}\hat{g}_i(X^*_{i,k}) + \epsilon^*] \\
&= \mathrm{Var}_*[\sum_{i=1}^m \frac{1}{s_i}\sum_{k=1}^{s_i}\hat{g}_i(X^*_{i,k})] + \mathrm{Var}_*[\epsilon^*] + 2\mathrm{Cov}_*(\sum_{i=1}^m \frac{1}{s_i}\sum_{k=1}^{s_i}\hat{g}_i(X^*_{i,k}), \epsilon^*) \\
&= \sum_{i=1}^m \frac{\hat{\sigma}_i^2}{s_i} + \mathrm{Var}_*[\epsilon^*] + O\Big(\sqrt{\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{s_i}\mathrm{Var}_*[\epsilon^*]}\Big) \\
&= \sum_{i=1}^m \frac{\hat{\sigma}_i^2}{\lceil \theta n_i \rceil} + \mathrm{Var}_*[\epsilon^*] + O\Big(\sqrt{\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{\lceil \theta n_i \rceil}\mathrm{Var}_*[\epsilon^*]}\Big).
\end{aligned}$$

Hence

$$\begin{aligned}
\sigma^2_{SVB} = \theta\mathrm{Var}_*[\psi^*] &= \sum_{i=1}^m \frac{\hat{\sigma}_i^2}{\lceil \theta n_i \rceil/\theta} + \theta\mathrm{Var}_*[\epsilon^*] + O\Big(\sqrt{\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{\lceil \theta n_i \rceil/\theta}\theta\mathrm{Var}_*[\epsilon^*]}\Big) \\
&= \sum_{i=1}^m (\frac{\hat{\sigma}_i^2}{n_i} + O(\frac{\hat{\sigma}_i^2}{n_i^2\theta})) + \theta\mathrm{Var}_*[\epsilon^*] + O\Big(\sqrt{\sum_{i=1}^m (\frac{\hat{\sigma}_i^2}{n_i} + O(\frac{\hat{\sigma}_i^2}{n_i^2\theta}))\theta\mathrm{Var}_*[\epsilon^*]}\Big). \qquad \text{(A.18)}
\end{aligned}$$

The convergence $\hat{\sigma}_i^2 \xrightarrow{p} \sigma_i^2$ and that $\theta = \omega(1/n)$ allow us to conclude

$$\frac{1}{\theta}\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{n_i^2} = o_p(\sum_{i=1}^m \frac{\sigma_i^2}{n_i}),\ \theta\mathrm{Var}_*[\epsilon^*] = \theta o_p(\sum_{i=1}^m \frac{1}{\lceil \theta n_i \rceil}) = o_p(\sum_{i=1}^m \frac{1}{n_i})$$

therefore $\sigma^2_{SVB} = \sum_{i=1}^m \frac{\sigma_i^2}{n_i} + o_p(\sum_{i=1}^m \frac{\sigma_i^2}{n_i})$. □

Proof of Lemma 2.4.4: Define $w := \hat{\psi}^* - \psi^*$ and $\delta := \psi^* - \mathbb{E}_*[\psi^*]$. Unbiasedness is well known, see e.g. Searle et al. (2009). The variance of $\hat{\sigma}^2_{SVB}/\theta$ has been derived in Sun et al. (2011) as

$$\begin{aligned}\frac{1}{\theta^2}\mathrm{Var}_*[\hat{\sigma}^2_{SVB}] =& \frac{1}{B}(\mathbb{E}_*[\delta^4] - (\mathbb{E}_*[\delta^2])^2) + \frac{2}{B(B-1)}(\mathbb{E}_*[\delta^2])^2 + \frac{2}{B^2R^2(B-1)}(\mathbb{E}_*[w^2])^2 \\ &+ \frac{2(B+1)}{B^2R(B-1)}\mathbb{E}_*[\delta^2]\mathbb{E}_*[w^2] + \frac{2}{B^2R^3}\mathbb{E}_*[w^4] + \frac{4B+2}{B^2R}\mathbb{E}_*[\delta^2 w^2] \\ &+ \frac{2(BR^2+R^2-4R+3)}{B^2R^3(R-1)}\mathbb{E}_*[(\mathbb{E}[w^2|\widehat{F}^*_{s_1,1},\ldots,\widehat{F}^*_{s_m,m}])^2] + \frac{4}{B^2R^2}\mathbb{E}_*[\delta w^3].\end{aligned}$$

Applying Jensen's inequality (or generalized Holder's inequality) gives

$$\begin{aligned}&\mathbb{E}_*[(\mathbb{E}[w^2|\widehat{F}^*_1,\ldots,\widehat{F}^*_m])^2] \le \mathbb{E}_*[w^4] \\ &\mathbb{E}_*[\delta^2 w^2] \le (\mathbb{E}_*[\delta^4]\mathbb{E}_*[w^4])^{1/2},\ |\mathbb{E}_*[\delta w^3]| \le (\mathbb{E}_*[\delta^4](\mathbb{E}_*[w^4])^3)^{1/4}\end{aligned}$$

The convergence condition $\mathbb{E}[(\hat{g}_i - g_i)^4(X_{i,1})] \to 0$ implies that $\frac{1}{n_i}\sum_{j=1}^{n_i}\hat{g}_i^4(X_{i,j}) = \frac{1}{n_i}\sum_{j=1}^{n_i} g_i^4(X_{i,j}) + o_p(1) = O_p(1)$. Together with the moment condition $\mathbb{E}_*[(\epsilon^* - \mathbb{E}_*[\epsilon^*])^4] = o_p(s^{-2})$, we get

$$\mathbb{E}_*[\delta^4] = 3\big(\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{s_i}\big)^2 + o_p\big(\big(\sum_{i=1}^m \frac{1}{s_i}\big)^2\big), \mathbb{E}_*[\delta^2] = \sum_{i=1}^m \frac{\hat{\sigma}_i^2}{s_i} + o_p\big(\sum_{i=1}^m \frac{1}{s_i}\big), \mathbb{E}_*[w^4] = \mathbb{E}_*[\mu_4^*] = O_p(1).$$

Hence the leading terms of the mean squared error can be identified as

$$\begin{aligned}\frac{1}{\theta^2}\mathrm{Var}_*[\hat{\sigma}^2_{SVB}] &\approx \frac{1}{B}(\mathbb{E}_*[\delta^4] - (\mathbb{E}_*[\delta^2])^2) + \frac{4}{BR}\mathbb{E}_*[\delta^2 w^2] + \frac{2}{BR^2}\mathbb{E}_*[(\mathbb{E}[w^2|\widehat{F}^*_{s_1,1},\ldots,\widehat{F}^*_{s_m,m}])^2] \\ &\approx \frac{2}{B}\big(\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{s_i}\big)^2 + \frac{4\hat{\tau}^2}{BR}\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{s_i} + \frac{2\hat{\tau}^4}{BR^2} = \frac{2}{B}\big(\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{s_i} + \frac{\hat{\tau}^2}{R}\big)^2.\end{aligned}$$

Here $a \approx b$ means $a/b \xrightarrow{p} 1$ as aforementioned. Therefore the variance can be expressed as

$$\mathrm{Var}_*[\hat{\sigma}^2_{SVB}] = \frac{2}{B}\big(\sum_{i=1}^m \frac{\hat{\sigma}_i^2}{n_i} + \frac{\hat{\tau}^2\theta}{R}\big)^2(1+o_p(1)) = \frac{2}{B}\big(\sum_{i=1}^m \frac{\sigma_i^2}{n_i} + \frac{\tau^2\theta}{R}\big)^2(1+o_p(1))$$

where the second equality holds because of the convergence of $\hat{\sigma}_i^2, \hat{\tau}^2$ to σ_i^2, τ^2. □

Proof of Theorem 2.3.1. Under the choice of B, R, θ stated in the theorem, we have that

$\mathrm{Var}_*[\hat{\sigma}^2_{SVB}] = o_p(1/n^2)$ hence $\hat{\sigma}^2_{SVB} - \sigma^2_{SVB} = o_p(1/n)$ on one hand. On the other hand we know the subsampling bootstrap variance estimate σ^2_{SVB} is consistent for σ^2_I and $\sigma^2_I = \Theta(1/n)$ hence $\sigma^2_{SVB} - \sigma^2_I = o_p(1/n)$. Now $\hat{\sigma}^2_{SVB} - \sigma^2_I = \hat{\sigma}^2_{SVB} - \sigma^2_{SVB} + \sigma^2_{SVB} - \sigma^2_I = o_p(1/n)$ from which consistency immediately follows. □

Proof of Theorem 2.4.5. One can easily verify that such B^* and R^* minimize the mean squared error (2.18) under the constraint that $BR = N$ and $B = \omega(1)$. The mean square error (2.19) then follows from evaluating (2.18) at B^*, R^*. □

Proof of Corollary 2.3.2. It is obvious that when $B = \omega(1)$ and $R = \Omega(\theta n)$ the configuration (2.7) is satisfied hence the estimate $\hat{\sigma}^2_{SVB}$ is relatively consistent under such allocation. To show that a simulation budget $N = \omega(\theta n)$ is necessary for (2.7) to hold, note that multiplying the first two requirements in (2.7) gives that $B^2R^2 = \omega((\theta n)^2)$, hence $BR = \omega(\theta n)$ must hold true. □

Proof of Corollary 2.3.3. This follows from letting $\theta = \omega(1/n)$ in Corollary 2.3.2 so that the require simulation budget $N = \omega(\theta n) = \omega(\omega(1)) = \omega(1)$. □

Proof of Theorem 2.3.4. The requirement $\omega(1/n) \le \theta$ is stipulated by (2.7). If $\theta \le o(N/n) \wedge 1$, then we have $\theta n = o(N)$, or equivalently $N/(\theta n) = \omega(1)$, so that we can afford a $B = \omega(1)$ when $R = \Omega(\theta n)$ to satisfy the first two requirements of (2.7). Theorem 2.3.1 then guarantees consistent variance estimation. □

Proof of Theorem 2.3.5. It follows from Theorem 2.4.5 by observing that $\tau^2 = \Theta(1)$ and $\sum_{i=1}^m \sigma_i^2/n_i = \Theta(1/n)$. □

A.4 Proofs for Results in Section 2.4.3 and Theorem 2.3.6

In this section we analyze the statistical error of σ^2_{SVB} relative to the true input variance σ^2_I, therefore, combined with the Monte Carlo error $\hat{\sigma}^2_{SVB} - \sigma^2_{SVB}$ given in Lemma 2.4.4, provide the overall error of the estimate $\hat{\sigma}^2_{SVB}$, and then minimize the overall error to obtain the optimal choices for the parameters θ, B, R. We first prove Lemma 2.4.8 using Lemma A.1.1 and Proposition 2.4.6 which have been presented in Section A.2, then use Lemma 2.4.8 to conclude Theorem 2.4.9. Lastly, Theorem 2.3.6 is derived from Theorem 2.4.9.

Proof of Lemma 2.4.8. The proof of Theorem 2.4.3 derives the following expression for the proportionate subsampled bootstrap variance

$$\frac{\sigma^2_{SVB}}{\theta} = \sum_{i=1}^{m} \frac{\hat{\sigma}_i^2}{s_i} + \mathrm{Var}_*[\epsilon^*] + 2\mathrm{Cov}_*(\sum_{i=1}^{m} \frac{1}{s_i} \sum_{k=1}^{s_i} \hat{g}_i(X^*_{i,k}), \epsilon^*).$$

As is the case in the proof of Proposition 2.4.6, the covariances can be simplified to

$$\mathrm{Cov}_*(\sum_{i=1}^{m} \frac{1}{s_i} \sum_{k=1}^{s_i} \hat{g}_i(X^*_{i,k}), \epsilon^*) = \sum_{i=1}^{m} \mathbb{E}_*[\hat{g}_i(X^*_{i,1})(\mathbb{E}_*[\epsilon^*|X^*_{i,1}] - \mathbb{E}_*[\epsilon^*])].$$

This leads to

$$\sigma^2_{SVB} = \sum_{i=1}^{m} \frac{\theta\hat{\sigma}_i^2}{\lfloor \theta n_i \rfloor} + \theta\mathbb{E}_*[(\epsilon^* - \mathbb{E}_*[\epsilon^*])^2] + 2\theta \sum_{i=1}^{m} \mathbb{E}_*[\hat{g}_i(X^*_{i,1})(\mathbb{E}_*[\epsilon^*|X^*_{i,1}] - \mathbb{E}_*[\epsilon^*])].$$

From the above expression of the variance estimator one can verify that it suffices to show the following three results

$$\sum_{i=1}^{m} \frac{\hat{\sigma}_i^2}{n_i} = \sigma_I^2 + \mathcal{Z} + o_p(\frac{1}{n^{3/2}}) \tag{A.19}$$

$$\mathbb{E}_*[(\epsilon^* - \mathbb{E}_*[\epsilon^*])^2] = \sum_{i,i'=1}^{m} \frac{1}{4s_i s_{i'}} \mathrm{Var}[g_{ii'}(X_i, X'_{i'})] + o_p(\frac{1}{s^2}) \tag{A.20}$$

$$\begin{aligned} &\mathbb{E}_*[\hat{g}_i(X^*_{i,1})(\mathbb{E}_*[\epsilon^*|X^*_{i,1}] - \mathbb{E}_*[\epsilon^*])] \qquad (A.21)\\ =&\frac{1}{2s_i^2}\mathrm{Cov}(g_i(X_i), g_{ii}(X_i, X_i)) + \sum_{i'=1}^{m} \frac{1}{2s_i s_{i'}} \mathrm{Cov}(g_i(X_i), \mathbb{E}_{X'_{i'}}[g_{ii'i'}(X_i, X'_{i'}, X'_{i'})]) + o_p(\frac{1}{s^2}). \end{aligned}$$

To see this, if the three equations hold then

$$\begin{aligned}
\sigma^2_{SVB} &= \sum_{i=1}^{m} \frac{\theta\hat{\sigma}_i^2}{\theta n_i - \mathrm{frac}(\theta n_i)} + \sum_{i,i'=1}^{m} \frac{1}{4n_i s_{i'}} \mathrm{Var}[g_{ii'}(X_i, X'_{i'})] + o_p(\frac{\theta}{s^2}) \\
&\quad + \sum_{i=1}^{m} \frac{1}{n_i s_i} \mathrm{Cov}(g_i(X_i), g_{ii}(X_i, X_i)) + \sum_{i,i'=1}^{m} \frac{1}{n_i s_{i'}} \mathrm{Cov}(g_i(X_i), \mathbb{E}_{X'_{i'}}[g_{ii'i'}(X_i, X'_{i'}, X'_{i'})]) \\
&\quad + o_p(\frac{\theta}{s^2}) \\
&= \sum_{i=1}^{m} \frac{\hat{\sigma}_i^2}{n_i} + \sum_{i=1}^{m} \frac{\mathrm{frac}(\theta n_i)\sigma_i^2}{n_i s_i} + o_p(\frac{1}{ns}) + \sum_{i,i'=1}^{m} \frac{1}{4n_i s_{i'}} \mathrm{Var}[g_{ii'}(X_i, X'_{i'})] \\
&\quad + \sum_{i=1}^{m} \frac{1}{n_i s_i} \mathrm{Cov}(g_i(X_i), g_{ii}(X_i, X_i)) + \sum_{i,i'=1}^{m} \frac{1}{n_i s_{i'}} \mathrm{Cov}(g_i(X_i), \mathbb{E}_{X'_{i'}}[g_{ii'i'}(X_i, X'_{i'}, X'_{i'})]) \\
&\quad + o_p(\frac{1}{ns}) \\
&= \sum_{i=1}^{m} \frac{\hat{\sigma}_i^2}{n_i} + \mathcal{R} + o_p(\frac{1}{ns}) \\
&= \sigma_I^2 + \mathcal{Z} + \mathcal{R} + o_p(\frac{1}{ns}) + o_p(\frac{1}{n^{3/2}})
\end{aligned}$$

where (A.20) and (A.21) are used in the first equality and (A.19) used in the last equality.

Now we prove the above three equations (A.19)-(A.21). By the expansion of $\hat{g}_i$ from Assumption 2.4.12 and the vanishing moment condition on the remainder ϵ_g, we write

$$\begin{aligned}
\hat{\sigma}_i^2 &= \frac{1}{n_i}\sum_{j=1}^{n_i} g_i^2(X_{i,j}) + \frac{2}{n_i}\sum_{j=1}^{n_i} g_i(X_{i,j})\Big(\sum_{i'=1}^{m}\frac{1}{n_{i'}}\sum_{j'=1}^{n_{i'}} g_{ii'}(X_{i,j}, X_{i',j'}) + \frac{1}{n_i}\sum_{j'=1}^{n_i} g_i(X_{i,j'})\Big) \\
&\qquad + \frac{1}{n_i}\sum_{j=1}^{n_i}\Big(\sum_{i'=1}^{m}\frac{1}{n_{i'}}\sum_{j'=1}^{n_{i'}} g_{ii'}(X_{i,j}, X_{i',j'}) + \frac{1}{n_i}\sum_{j'=1}^{n_i} g_i(X_{i,j'})\Big)^2 + o_p(\frac{1}{\sqrt{n}}) \quad \text{(A.22)} \\
&= \frac{1}{n_i}\sum_{j=1}^{n_i} g_i^2(X_{i,j}) + \frac{2}{n_i}\sum_{j=1}^{n_i} g_i(X_{i,j})\sum_{i'=1}^{m}\frac{1}{n_{i'}}\sum_{j'=1}^{n_{i'}} g_{ii'}(X_{i,j}, X_{i',j'}) + 2\Big(\frac{1}{n_i}\sum_{j=1}^{n_i} g_i(X_{i,j})\Big)^2 + o_p(\frac{1}{\sqrt{n}}) \\
&= \frac{1}{n_i}\sum_{j=1}^{n_i} g_i^2(X_{i,j}) + 2\sum_{i'=1}^{m}\frac{1}{n_i n_{i'}}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_{i'}} g_i(X_{i,j}) g_{ii'}(X_{i,j}, X_{i',j'}) + o_p(\frac{1}{\sqrt{n}}). \quad \text{(A.23)}
\end{aligned}$$

Note that the first term in line (A.22) has an expectation of order $O(1/n)$ hence can be absorbed into the $o_p(1/\sqrt{n})$ term. Similarly the fourth line (A.23) holds because $(\sum_{j=1}^{n_i} g_i(X_{i,j})/n_i)^2$ has an

expectation of order $O(1/n)$. The second term in (A.23) is a sum of m V-statistics, each of which by standard results is well approximated by the Hajek projection

$$\frac{1}{n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} g_i(X_{i,j}) g_{ii'}(X_{i,j}, X_{i',j'}) = \frac{1}{n_{i'}} \sum_{j'=1}^{n_{i'}} \mathbb{E}_{X_i}[g_i(X_i) g_{ii'}(X_i, X_{i',j'})] + O_p(\frac{1}{n}).$$

The finite fourth moment condition of g_i and $g_{i_1 i_2}$ are used to ensure that $g_i(X_{i,j}) g_{ii'}(X_{i,j}, X_{i',j'})$ has a finite second moment so that the above approximation holds. Denoting

$$\mu_1^i = \frac{1}{n_i} \sum_{j=1}^{n_i} g_i^2(X_{i,j}),\ \mu_2^{ii'} = \frac{1}{n_{i'}} \sum_{j'=1}^{n_{i'}} \mathbb{E}_{X_i}[g_i(X_i) g_{ii'}(X_i, X_{i',j'})]$$

we have

$$\sum_{i=1}^{m} \frac{\hat{\sigma}_i^2}{n_i} = \sum_{i=1}^{m} \frac{\mu_1^i}{n_i} + 2 \sum_{i=1}^{m} \sum_{i'=1}^{m} \frac{\mu_2^{ii'}}{n_i} + o_p(\frac{1}{n^{3/2}}).$$

Because of independence among input models the variance of the leading term takes the additive form $\sum_{i=1}^m \lambda_i^T \Sigma_i \lambda_i / n_i$ as described in the theorem. By Proposition 2.4.6 $\sigma_I^2 = \sum_{i=1}^m \sigma_i^2 / n_i + O(n^{-2})$ hence equation (A.19) follows. To show (A.20), we note that in the cubic expansion of Assumption 2.4.12 the cubic term and the remainder ϵ_3^* both have a second moment of order $O_p(s^{-3})$. Therefore it suffices to consider the quadratic term. Since the second order influence function $\hat{g}_{i_1 i_2}$ has vanishing marginal expected value, one can verify that

$$\begin{aligned}
\mathrm{Var}_*\Big[\sum_{i,i'=1}^{m} \frac{1}{s_i s_{i'}} \sum_{j=1}^{s_i} \sum_{j'=1}^{s_{i'}} \hat{g}_{ii'}(X_{i,j}^*, X_{i',j'}^*) \Big] &= \sum_{i,i'=1}^{m} \frac{1}{s_i s_{i'} n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} \hat{g}_{ii'}^2(X_{i,j}, X_{i',j'}) + O_p(\frac{1}{s^3}) \\
&= \sum_{i,i'=1}^{m} \frac{1}{s_i s_{i'} n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} g_{ii'}^2(X_{i,j}, X_{i',j'}) + o_p(\frac{1}{s^2})
\end{aligned}$$

where the second equality follows from the convergence of $\hat{g}_{ii'}$ to $g_{ii'}$ as imposed in Assumption 2.4.12. Equation (A.20) then follows from the fact that $\frac{1}{n_i n_{i'}} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} g_{ii'}^2(X_{i,j}, X_{i',j'})$ is consistent as a V-statistic.

Let's continue to prove equation (A.21). Denote by X_i^* a generic resampled data point from

the i-th input data set. Then one can check that

$$\begin{aligned}
&\mathbb{E}_*[\epsilon^*|X^*_{i,1} = X_{i,j}] - \mathbb{E}_*[\epsilon^*] \\
=&\frac{1}{2s_i^2}(\hat{g}_{ii}(X_{i,j}, X_{i,j}) - \mathbb{E}_*[\hat{g}_{ii}(X^*_i, X^*_i)]) + \frac{1}{6s_i^3}(\hat{g}_{iii}(X_{i,j}, X_{i,j}, X_{i,j}) - \mathbb{E}_*[\hat{g}_{iii}(X^*_i, X^*_i, X^*_i)]) \\
&+ \frac{s_i - 1}{2s_i^3}\mathbb{E}_*[\hat{g}_{iii}(X_{i,j}, X^*_i, X^*_i)] + \sum_{i' \neq i} \frac{1}{2s_i s_{i'}}\mathbb{E}_*[\hat{g}_{ii'i'}(X_{i,j}, X^*_{i'}, X^*_{i'})] + \mathbb{E}_*[\epsilon^*_3|X^*_{i,1} = X_{i,j}] - \mathbb{E}_*[\epsilon^*_3].
\end{aligned}$$

Note that $\mathrm{Var}_*[\epsilon^*_3|X^*_{i,1}] = o_p(s^{-4})$ because of Assumption 2.4.12 and Lemma A.2.1. Hence

$$\begin{aligned}
&\mathbb{E}_*[\hat{g}_i(X^*_{i,1})(\mathbb{E}_*[\epsilon^*|X^*_{i,1}] - \mathbb{E}_*[\epsilon^*])] \\
=&\frac{1}{n_i}\sum_{j=1}^{n_i} \hat{g}_i(X_{i,j})(\frac{1}{2s_i^2}(\hat{g}_{ii}(X_{i,j}, X_{i,j}) - \mathbb{E}_*[\hat{g}_{ii}(X^*_i, X^*_i)]) + \sum_{i'=1}^{m} \frac{1}{2s_i s_{i'}}\mathbb{E}_*[\hat{g}_{ii'i'}(X_{i,j}, X^*_{i'}, X^*_{i'})]) + o_p(\frac{1}{s^2}) \\
=&\frac{1}{2s_i^2}\mathrm{Cov}_*(\hat{g}_i(X^*_i), \hat{g}_{ii}(X^*_i, X^*_i)) + \sum_{i'=1}^{m} \frac{1}{2s_i s_{i'}}\mathrm{Cov}_*(\hat{g}_i(X^*_i), \mathbb{E}_{X^{*\prime}_{i'}}[\hat{g}_{ii'i'}(X^*_i, X^{*\prime}_{i'}, X^{*\prime}_{i'})]) + o_p(\frac{1}{s^2}) \\
=&\frac{1}{2s_i^2}\mathrm{Cov}(g_i(X_i), g_{ii}(X_i, X_i)) + \sum_{i'=1}^{m} \frac{1}{2s_i s_{i'}}\mathrm{Cov}(g_i(X_i), \mathbb{E}_{X'_{i'}}[g_{ii'i'}(X_i, X'_{i'}, X'_{i'})]) + o_p(\frac{1}{s^2})
\end{aligned}$$

where the $o_p(1/s^2)$ term in the first equality comes from applying Cauchy Schwartz inequality, and the last equality holds since convergence of $\hat{g}_i, \hat{g}_{i_1 i_2}, \hat{g}_{i_1 i_2 i_3}$ to $g_i, g_{i_1 i_2}, g_{i_1 i_2 i_3}$ in mean squared error implies

$$\begin{aligned}
&\mathrm{Cov}_*(\hat{g}_i(X^*_i), \hat{g}_{ii}(X^*_i, X^*_i)) \xrightarrow{p} \mathrm{Cov}(g_i(X_i), g_{ii}(X_i, X_i)) \\
&\mathrm{Cov}_*(\hat{g}_i(X^*_i), \mathbb{E}_{X^{*\prime}_{i'}}[\hat{g}_{ii'i'}(X^*_i, X^{*\prime}_{i'}, X^{*\prime}_{i'})]) \xrightarrow{p} \mathrm{Cov}(g_i(X_i), \mathbb{E}_{X'_{i'}}[g_{ii'i'}(X_i, X'_{i'}, X'_{i'})]).
\end{aligned}$$

This gives rise to the equation (A.21). □

Proof of Theorems 2.3.6 and 2.4.9. We first show Theorem 2.4.9. Under a given subsampling ratio θ, we know from Lemma 2.4.8 and Theorem 2.3.5 that under the optimal allocation $B^* =$

N/R^* and $R^* = \Theta(\theta n)$

$$\hat{\sigma}^2_{SVB} - \sigma^2_{SVB} = \mathcal{E}_1 + o_p\big(\sqrt{\frac{\theta}{Nn}}\big)$$
$$\sigma^2_{SVB} - \sigma^2_I = \mathcal{E}_2 + o_p\big(\frac{1}{n^{3/2}} + \frac{1}{\theta n^2}\big)$$

where the errors $\mathcal{E}_1, \mathcal{E}_2$ satisfy $\mathbb{E}_*[\mathcal{E}_1] = 0, \mathbb{E}[\mathcal{E}_1^2] = \Theta(\theta/(Nn))$ and $\mathbb{E}[\mathcal{E}_2^2] = \mathcal{R}^2 + \sum_{i=1}^m \lambda_i^T \Sigma_i \lambda_i / n_i$. Letting $\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2$, we have $\mathbb{E}[\mathcal{E}^2] = \mathbb{E}[\mathcal{E}_1^2] + \mathbb{E}[\mathcal{E}_2^2]$ because

$$\mathbb{E}[\mathcal{E}_1\mathcal{E}_2] = \mathbb{E}_{\text{data}}[\mathbb{E}_*[\mathcal{E}_1\mathcal{E}_2]] = \mathbb{E}_{\text{data}}[\mathcal{E}_2\mathbb{E}_*[\mathcal{E}_1]] = 0.$$

This gives Theorem 2.4.9.

To prove Theorem 2.3.6, note that if $\mathcal{R} = \Theta((ns)^{-1})$, and at least one of the Σ_i's are positive definite, then $\sum_{i=1}^m \lambda_i^T \Sigma_i \lambda_i / n_i = \Theta(1/n^3)$ hence $\mathbb{E}[\mathcal{E}_2^2] = \Theta(1/n^3 + 1/(\theta^2 n^4))$. We have

$$\hat{\sigma}^2_{SVB} - \sigma^2_I = \mathcal{E} + o_p\big(\sqrt{\frac{\theta}{Nn}} + \frac{1}{n^{3/2}} + \frac{1}{\theta n^2}\big)$$

where $\mathbb{E}[\mathcal{E}^2] = \Theta(\theta/(Nn) + 1/n^3 + 1/(\theta^2 n^4))$. To minimize the leading term $\mathcal{E}$, just note that $\theta/(Nn) + 1/(\theta^2 n^4)$ is minimized at $\theta^* = (2N)^{1/3}/n$ resulting in $\mathbb{E}[\mathcal{E}^2] = \Theta(1/(N^{2/3}n^2) + 1/n^3)$. When $N > n^{3/2}$, we have $1/(N^{2/3}n^2) < 1/n^3$, hence as long as θ^* is chosen such that $\theta^*/(Nn) \leq 1/n^3$ and $1/(\theta^{*2}n^4) \leq 1/n^3$, or equivalently $1/\sqrt{n} \leq \theta^* \leq N/n^2 \wedge 1$, then the error $\mathbb{E}[\mathcal{E}^2] = \Theta(1/n^3)$. This leads to the optimal subsample size (2.10). If the depicted conditions do not hold, we have $\mathbb{E}[\mathcal{E}_2^2] \leq \Theta(1/n^3 + 1/(\theta^2 n^4))$ in general, hence all upper bounds we just obtained for $\mathbb{E}[\mathcal{E}^2]$ could be loose in order, leading to (2.12). □

Appendix B

Technical Proofs for Chapter 3

This chapter contains technical proofs of results in Chapter 3.

B.1 Notation and Outline

We introduce some notations to be used throughout this chapter. Quantities that appear in our analysis can involve up to three sources of randomness: The input data $\{X_{i,j} : i = 1, \ldots, m, j = 1, \ldots, n_i\}$, the simulation runs in Step 1, and the simulation runs in Step 3 of our algorithms. We denote the three sources of randomness by D, ξ_1 and ξ_2 respectively.

In order to highlight the dependence of each quantity on different types of randomness, expectation operators in this chapter are accompanied by subscripts indicating relevant sources of randomness. For example, the simulation-based performance estimate $\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)$ from Step 1 of Algorithm 4 depends on the input data and the simulation runs in Step 1 but not on Step 3 so its (total) expectation is written as $\mathbb{E}_{D,\xi_1}[\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)]$, whereas the empirical performance measure $Z(\hat{P}_1, \ldots, \hat{P}_m)$ depends on input data only hence its expectation is denoted by $\mathbb{E}_D[Z(\hat{P}_1, \ldots, \hat{P}_m)]$. However, if a quantity depends on all the three sources, we write its expectation in the plain format $\mathbb{E}[\cdot]$ in place of $\mathbb{E}_{D,\xi_1,\xi_2}[\cdot]$ for simplicity.

If conditioning on one or two sources of randomness is needed, we use the notation $\mathbb{E}_{\cdot|\cdot}$ for conditional expectation and $\mathrm{Var}_{\cdot|\cdot}$ for conditional variance where in the subscript the conditioned

randomness is placed after the vertical bar and the randomness to be integrated out placed before. For instance, it holds for the performance estimate $\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)$ from Step 1 of Algorithm 4 that $\mathbb{E}_{\xi_1|D}[\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)] = Z(\hat{P}_1, \ldots, \hat{P}_m)$ and that $\mathrm{Var}_{\xi_1|D}(\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)) = \frac{1}{R_1} \cdot \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)$ where $\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)$ is the variance of a single simulation replication for h driven by the empirical input models (see the detailed definition in the next paragraph).

We also need to deal with various simulation outputs where the underlying input models are weighted distributions supported on the input data. For any performance function $g(\mathbf{X}_1, \ldots, \mathbf{X}_m) : \mathfrak{X}_1^{T_1} \times \cdots \times \mathfrak{X}_m^{T_m} \to \mathbb{R}$ and any probability weight $(\mathbf{w}_1, \ldots, \mathbf{w}_m) \in \mathcal{U}_\alpha$ we define

$$\mathbb{E}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}[g(\mathbf{X}_1, \ldots, \mathbf{X}_m)] := \mathbb{E}_{\sum_{j=1}^{n_1} w_{1,j}\delta_{X_{1,j}},\ldots,\sum_{j=1}^{n_m} w_{m,j}\delta_{X_{m,j}}}[g(\mathbf{X}_1, \ldots, \mathbf{X}_m)]$$

where each $\delta_{X_{i,j}}$ is the delta measure at $X_{i,j}$ and the expectation on the right hand side shall be interpreted as under the product measure $\left(\sum_{j=1}^{n_1} w_{1,j}\delta_{X_{1,j}}\right)^{T_1} \times \cdots \times \left(\sum_{j=1}^{n_m} w_{m,j}\delta_{X_{m,j}}\right)^{T_m}$ like in (3.1). In particular, if $g = h$ then $\mathbb{E}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}[h(\mathbf{X}_1, \ldots, \mathbf{X}_m)]$ is the same as $Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)$, and if furthermore the weight $w_{i,j} = 1/n_i$ for each i, j it is essentially $Z(\hat{P}_1, \ldots, \hat{P}_m)$. Similarly

$$\mathrm{Var}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}(g(\mathbf{X}_1, \ldots, \mathbf{X}_m)) := \mathrm{Var}_{\sum_{j=1}^{n_1} w_{1,j}\delta_{X_{1,j}},\ldots,\sum_{j=1}^{n_m} w_{m,j}\delta_{X_{m,j}}}(g(\mathbf{X}_1, \ldots, \mathbf{X}_m))$$

represents the variance of g when each input variate $X_i(t)$ is generated from the weighted input distribution $\sum_{j=1}^{n_i} w_{i,j}\delta_{X_{i,j}}$. Particularly, $\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)$ is the output variance under the uniformly weighted input distributions. When the probability weights are $(\mathbf{w}_1^{\min}, \ldots, \mathbf{w}_m^{\min})$ or $(\mathbf{w}_1^{\max}, \ldots, \mathbf{w}_m^{\max})$ from Step 2 of our algorithms, we write the corresponding output variances as

$$\sigma^2_{\min} := \mathrm{Var}_{\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min}}(h(\mathbf{X}_1, \ldots, \mathbf{X}_m))$$
$$\sigma^2_{\max} := \mathrm{Var}_{\mathbf{w}_1^{\max},\ldots,\mathbf{w}_m^{\max}}(h(\mathbf{X}_1, \ldots, \mathbf{X}_m))$$

where each $\mathbf{w}_i^{\min}$ should be understood as the probability measure $\sum_{j=1}^{n_i} w_{i,j}^{\min}\delta_{X_{i,j}}$. Note that $\mathbb{E}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}[\cdot]$ and $\mathrm{Var}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}[\cdot]$ are conditional expectations/variances of one simulation replication given both the input data and the weights, but to make the notations compact we choose to suppress

the data dependence.

X_i is used as a generic random variable distributed under the true input model P_i for each i. Given a positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, $\mathcal{N}(\mathbf{0}, \Sigma)$ denotes the multivariate normal distribution on $\mathbb{R}^d$ with mean zero and covariance matrix Σ. In particular, $\mathcal{N}(0, 1)$ denotes the univariate standard normal. $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0, 1)$. We use $\Rightarrow$ to denote weak convergence of probability measures.

We present our proofs as follows. We first prove all the results in Section 3.4, organized via the subsections. Given these developments, we then prove the main results in Section 3.3 including Theorems 3.3.2, 3.3.3, 3.3.4, and also Proposition 3.3.1.

B.2 Proofs of Results in Section 3.4.2

Proof of Proposition 3.4.1. Let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,T_i})$. First we rewrite the performance measure as an integral

$$\begin{aligned}
&Z((1-\epsilon)Q_1^1 + \epsilon Q_1^2, \ldots, (1-\epsilon)Q_m^1 + \epsilon Q_m^2) \\
=&\int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i=1}^{m}\prod_{t=1}^{T_i} d(Q_i^1 + \epsilon(Q_i^2 - Q_i^1))(x_{i,t}) \\
=&Z(Q_1^1, \ldots, Q_m^1) + \sum_{i=1}^{m}\sum_{t=1}^{T_i} \epsilon \int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{r\neq i \text{ or } s\neq t} dQ_r^1(x_{r,s}) \cdot d(Q_i^2 - Q_i^1)(x_{i,t}) + \mathcal{R} \qquad \text{(B.1)}
\end{aligned}$$

by expanding out all the Q_i^1 and $\epsilon(Q_i^2 - Q_i^1)$ in the product measure, and the remainder $\mathcal{R}$ includes all the terms that have an ϵ^k with $k \geq 2$. The integrability condition guarantees that all the integral

terms above, including those in $\mathcal{R}$, are finite. Note that

$$
\begin{aligned}
&\sum_{i=1}^{m}\sum_{t=1}^{T_i}\int h(\mathbf{x}_1,\ldots,\mathbf{x}_m)\prod_{r\neq i \text{ or } s\neq t} dQ_r^1(x_{r,s})\cdot d(Q_i^2-Q_i^1)(x_{i,t})\\
=&\sum_{i=1}^{m}\sum_{t=1}^{T_i}\left(\int h(\mathbf{x}_1,\ldots,\mathbf{x}_m)\prod_{r\neq i \text{ or } s\neq t} dQ_r^1(x_{r,s})\cdot dQ_i^2(x_{i,t}) - Z(Q_1^1,\ldots,Q_m^1)\right)\\
=&\sum_{i=1}^{m}\sum_{t=1}^{T_i}\int\left(\int h(\mathbf{x}_1,\ldots,\mathbf{x}_m)\prod_{r\neq i \text{ or } s\neq t} dQ_r^1(x_{r,s}) - Z(Q_1^1,\ldots,Q_m^1)\right) dQ_i^2(x_{i,t})\\
=&\sum_{i=1}^{m}\sum_{t=1}^{T_i}\int\left(\int h(\mathbf{x}_1,\ldots,\mathbf{x}_i^{(t)},\ldots,\mathbf{x}_m)\prod_{r\neq i \text{ or } s\neq t} dQ_r^1(x_{r,s}) - Z(Q_1^1,\ldots,Q_m^1)\right) dQ_i^2(x_i)\\
=&\sum_{i=1}^{m}\int\sum_{t=1}^{T_i}\left(\int h(\mathbf{x}_1,\ldots,\mathbf{x}_i^{(t)},\ldots,\mathbf{x}_m)\prod_{r\neq i \text{ or } s\neq t} dQ_r^1(x_{r,s}) - Z(Q_1^1,\ldots,Q_m^1)\right) dQ_i^2(x_i)\\
=&\sum_{i=1}^{m}\int G_i^{Q_1^1,\ldots,Q_m^1}(x_i)dQ_i^2(x_i) = \sum_{i=1}^{m}\mathbb{E}_{Q_i^2}[G_i^{Q_1^1,\ldots,Q_m^1}(X)],
\end{aligned}
$$

where the second equality holds because dQ_i^2 is a probability measure, and the third equality is a notational replacement of $x_{i,t}$ by x_i, with $\mathbf{x}_i^{(t)}$ defined as $\mathbf{x}_i$ but with $x_{i,t}$ replaced by x_i. This and (B.1) together show the derivative expression (3.11). The mean zero property of $G_i^{Q_1^1,\ldots,Q_m^1}$ follows from the tower property

$$\mathbb{E}_{Q_i^1}\left[\mathbb{E}_{Q_1^1,\ldots,Q_m^1}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_i(t)]\right] = Z(Q_1^1,\ldots,Q_m^1)$$

for all $t=1,\ldots,T_i$. □

Proof of Proposition 3.4.2. We first provide two lemmas.

Lemma B.2.1 *Every feasible solution* $(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha$ *satisfies*

$$\frac{l(\alpha)}{n_i}\leq w_{i,j}\leq\frac{u(\alpha)}{n_i},\forall\, i=1,\ldots,m, j=1,\ldots,n_i$$

where $0<l(\alpha)<1<u(\alpha)<+\infty$ *are the two solutions of the equation* $xe^{1+\frac{\chi^2_{1,1-\alpha}}{2}-x}=1$.

Proof of Lemma B.2.1. Consider $(\mathbf{w}_1, \ldots, \mathbf{w}_m) \in \mathcal{U}_\alpha$. By Jensen's inequality, for each i we have

$$-\sum_{j=1}^{n_i} \log(n_i w_{i,j}) \geq -n_i \log \sum_{j=1}^{n_i} w_{i,j} = 0,$$

and thus

$$-2\sum_{j=1}^{n_i} \log(n_i w_{i,j}) \leq -2\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log(n_i w_{i,j}) \leq \mathcal{X}^2_{1,1-\alpha}.$$

This implies for each $i = 1, \ldots, m$

$$\prod_{j=1}^{n_i} n_i w_{i,j} \geq e^{-\frac{\mathcal{X}^2_{1,1-\alpha}}{2}}. \tag{B.2}$$

For any $s = 1, \ldots, n_i$, we shall show that $l(\alpha) \leq n_i w_{i,s} \leq u(\alpha)$. Taking $n_i w_{i,s}$ out of the product in (B.2) and noticing the inequality $\prod_{j\neq s} n_i w_{i,j} \leq \left(\frac{n_i}{n_i-1}\sum_{j\neq s} w_{i,j}\right)^{n_i-1} = \left(\frac{n_i(1-w_{i,s})}{n_i-1}\right)^{n_i-1}$ gives

$$n_i w_{i,s}\left(1 + \frac{1 - n_i w_{i,s}}{n_i - 1}\right)^{n_i-1} \geq n_i w_{i,s} \prod_{j\neq s} n_i w_{i,j} \geq e^{-\frac{\mathcal{X}^2_{1,1-\alpha}}{2}}.$$

Applying $e^x \geq 1 + x$ to $1 + \frac{1-n_i w_{i,s}}{n_i-1}$ gives

$$n_i w_{i,s} e^{1-n_i w_{i,s}} \geq e^{-\frac{\mathcal{X}^2_{1,1-\alpha}}{2}}. \tag{B.3}$$

Simple calculations show that the function xe^{1-x} strictly increases from 0 to 1 for $x \in (0, 1)$ and decreases from 1 to 0 for $x \in (1, +\infty)$. So it follows from (B.3) that $n_i w_{i,s}$ must fall between the two solutions of $xe^{1-x} = e^{-\frac{\mathcal{X}^2_{1,1-\alpha}}{2}}$. □

Lemma B.2.2 *Let $u(\alpha)$ be the constant from Lemma B.2.1. Every feasible solution $(\mathbf{w}_1, \ldots, \mathbf{w}_m) \in \mathcal{U}_\alpha$ satisfies*

$$\sum_{i=1}^{m} n_i^2 \sum_{j=1}^{n_i} (w_{i,j} - \frac{1}{n_i})^2 \leq u(\alpha)^2 \mathcal{X}^2_{1,1-\alpha}.$$

Proof of Lemma B.2.2. Taylor expand each summand in the left hand side of the first constraint

in $\mathcal{U}_\alpha$, around the uniform weights, and use the mean value theorem to get

$$\begin{aligned}-2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}) &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(0-2n_i(w_{i,j}-\frac{1}{n_i})+(\theta_{i,j}w_{i,j}+(1-\theta_{i,j})\frac{1}{n_i})^{-2}(w_{i,j}-\frac{1}{n_i})^2\right)\\ &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\theta_{i,j}w_{i,j}+(1-\theta_{i,j})\frac{1}{n_i})^{-2}(w_{i,j}-\frac{1}{n_i})^2\end{aligned}$$

where $\theta_{i,j}$ is some constant such that $0 \le \theta_{i,j} \le 1$, for each i, j. Lemma B.2.1 implies $\theta_{i,j}w_{i,j} + (1-\theta_{i,j})\frac{1}{n_i} \le \frac{u(\alpha)}{n_i}$. Hence

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{n_i^2}{u(\alpha)^2}(w_{i,j}-\frac{1}{n_i})^2 \le -2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}) \le \mathcal{X}^2_{1,1-\alpha}.$$

Multiplying $u(\alpha)^2$ on both sides completes the proof. □

Now we are ready to prove Proposition 3.4.2. Let $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,T_i})$. We will first show the uniform error bound of the linear approximation Z_L, and then $\widehat{Z_L}$. We start the analysis by expressing $Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ as

$$Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) = \int h(\mathbf{x}_1, \ldots, \mathbf{x}_m)\prod_{i=1}^{m}\prod_{t=1}^{T_i} d\mathbf{w}_i(x_{i,t}) \tag{B.4}$$

where we abuse notation to write $\mathbf{w}_i$ as a probability measure over the observations $\{X_{i,j}\}_{j=1,\ldots,n_i}$. Rewrite $d\mathbf{w}_i$ as $d(\mathbf{w}_i - \hat{P}_i + \hat{P}_i - P_i + P_i)$, where $\hat{P}_i$ is the empirical distribution of the i-th sample, and expand out $\mathbf{w}_i - \hat{P}_i$, $\hat{P}_i - P_i$ and P_i in the product measure in (B.4) to get

$$\begin{aligned}&Z(\mathbf{w}_1, \ldots, \mathbf{w}_m)\\ =&\sum_{\mathcal{T}_i^1, \mathcal{T}_i^2}\int h(\mathbf{X}_1, \ldots, \mathbf{X}_m)\prod_{i=1}^{m}\prod_{t\notin\mathcal{T}_i^1\cup\mathcal{T}_i^2} dP_i(x_{i,t})\prod_{i=1}^{m}\prod_{t\in\mathcal{T}_i^1} d(\hat{P}_i - P_i)(x_{i,t})\prod_{i=1}^{m}\prod_{t\in\mathcal{T}_i^2} d(\mathbf{w}_i - \hat{P}_i)(x_{i,t})\\ =&\sum_{d=0}^{T}\sum_{\sum_i(|\mathcal{T}_i^1|+|\mathcal{T}_i^2|)=d}\int h\prod_{i=1}^{m}\prod_{t\notin\mathcal{T}_i^1\cup\mathcal{T}_i^2} dP_i(x_{i,t})\prod_{i=1}^{m}\prod_{t\in\mathcal{T}_i^1} d(\hat{P}_i - P_i)(x_{i,t})\prod_{i=1}^{m}\prod_{t\in\mathcal{T}_i^2} d(\mathbf{w}_i - \hat{P}_i)(x_{i,t})\end{aligned} \tag{B.5}$$

where for each i, $\mathcal{T}_i^1, \mathcal{T}_i^2$ are two disjoint and ordered (possibly empty) subsets of $\{1, 2, \ldots, T_i\}$

that specifies the second subscript t of the argument $x_{i,t}$, $|\cdot|$ denotes the cardinality of a set, and $T = \sum_{i=1}^{m} T_i$.

The desired conclusion can be achieved upon completing the following two tasks: (1) show that the terms with $d = 0, 1$ above give the linear approximation; (2) each term with $d \geq 2$ is of order $O(1/n^d)$ in terms of its mean square.

Task one: $d = 0, 1$

The only summand with $d = 0$ is

$$\int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i=1}^{m} \prod_{t=1}^{T_i} dP_i(x_{i,t}) = Z(P_1, \ldots, P_m) = Z^*,$$

and each summand with $d = 1$ is one of the following two types

$$\int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i \neq r \text{ or } t \neq s} dP_i(x_{i,t}) d(\hat{P}_r - P_r)(x_{r,s}), \text{ for } r = 1, \ldots, m, s = 1, \ldots, T_i$$

or

$$\int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i \neq r \text{ or } t \neq s} dP_i(x_{i,t}) d(\mathbf{w}_r - \hat{P}_r)(x_{r,s}), \text{ for } r = 1, \ldots, m, s = 1, \ldots, T_i.$$

For each r and s the two types sum up to

$$\int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i \neq r \text{ or } t \neq s} dP_i(x_{i,t}) d(\mathbf{w}_r - P_r)(x_{r,s}).$$

Summing over all r, s gives

$$\begin{aligned}
&\sum_{r=1}^{m}\sum_{s=1}^{T_r}\int h(\mathbf{x}_1,\ldots,\mathbf{x}_m)\prod_{i\neq r \text{ or } t\neq s} dP_i(x_{i,t})d(\mathbf{w}_r - P_r)(x_{r,s}) \\
=&\sum_{r=1}^{m}\sum_{s=1}^{T_r}\int h(\mathbf{x}_1,\ldots,\mathbf{x}_r^{(s)},\ldots \mathbf{x}_m)\prod_{i\neq r \text{ or } t\neq s} dP_i(x_{i,t})d(\mathbf{w}_r - P_r)(x_r) \\
&\text{by replacing } x_{r,s} \text{ with } x_r, \text{ and denoting } \mathbf{x}_r^{(s)} \text{ as } \mathbf{x}_r \text{ but with } x_{r,s} \text{ replaced by } x_r \\
=&\sum_{r=1}^{m}\int \Big(\sum_{s=1}^{T_r}\int h(\mathbf{x}_1,\ldots,\mathbf{x}_r^{(s)},\ldots,\mathbf{x}_m)\prod_{i\neq r \text{ or } t\neq s} dP_i(x_{i,t})\Big)d(\mathbf{w}_r - P_r)(x_r) \\
=&\sum_{r=1}^{m}\int \sum_{s=1}^{T_r}\Big(\int h(\mathbf{x}_1,\ldots,\mathbf{x}_r^{(s)},\ldots,\mathbf{x}_m)\prod_{i\neq r \text{ or } t\neq s} dP_i(x_{i,t}) - Z(P_1,\ldots,P_m)\Big)d\mathbf{w}_r(x_r) \\
=&\sum_{r=1}^{m}\sum_{j=1}^{n_i} w_{r,j}G_r(X_{r,j}).
\end{aligned}$$

This concludes that the summands with $d = 0, 1$ sum up to the linear approximation $Z_L = Z^* + \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{i,j}G_i(X_{i,j})$.

Task two: $d \geq 2$

Now we deal with the terms in (B.5) with $d \geq 2$. Define

$$\mathcal{M} := \max_{I_1,\ldots,I_m} \mathbb{E}_{P_1,\ldots,P_m}[|h(\mathbf{X}_{1,I_1},\ldots,\mathbf{X}_{m,I_m})|^2], \tag{B.6}$$

where each $I_i \in \{1, 2, \ldots, T_i\}^{T_i}$. Note that $\mathcal{M}$ is finite under Assumption 3.3.3 due to Jensen's inequality. Consider a generic summand from (B.5)

$$R_d(\mathcal{T}^1, \mathcal{T}^2) = \int h(\mathbf{X}_1,\ldots,\mathbf{X}_m)\prod_{i=1}^{m}\prod_{t\notin\mathcal{T}_i^1\cup\mathcal{T}_i^2} dP_i(x_{i,t})\prod_{i=1}^{m}\prod_{t\in\mathcal{T}_i^1} d(\hat{P}_i - P_i)(x_{i,t})\prod_{i=1}^{m}\prod_{t\in\mathcal{T}_i^2} d(\mathbf{w}_i - \hat{P}_i)(x_{i,t})$$

where we denote $\mathcal{T}^1 = (\mathcal{T}_1^1, \ldots, \mathcal{T}_m^1)$, $\mathcal{T}^2 = (\mathcal{T}_1^2, \ldots, \mathcal{T}_m^2)$. Note that $\sum_{i=1}^{m}(|\mathcal{T}_i^1| + |\mathcal{T}_i^2|) = d$, and the subscript d in $R_d(\mathcal{T}^1, \mathcal{T}^2)$ is used to emphasize this dependence. Let $\mathcal{T}_i^1(t)$ (or $\mathcal{T}_i^2(t)$) be the

t-th element of $\mathcal{T}_i^1$ (or $\mathcal{T}_i^2$). Our goal is to show that

$$\mathbb{E}_D\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}|R_d(\mathcal{T}^1,\mathcal{T}^2)|^2\big]=O\big(\prod_{i=1}^m n_i^{-(|\mathcal{T}_i^1|+|\mathcal{T}_i^2|)}\big)=O(n^{-d}), \tag{B.7}$$

where $\mathbb{E}_D$ is used for the expectation because $\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}|R_d(\mathcal{T}^1,\mathcal{T}^2)|^2$ is a theoretical quantity that depends on the input data only and can not be exactly computed from simulation runs.

First, we rewrite $R_d(\mathcal{T}^1,\mathcal{T}^2)$ as a sum and from there derive an upper bound (B.10) of its supremum. Define the conditional expectation of h for given subscripts $\mathcal{T}^1=(\mathcal{T}_1^1,\ldots,\mathcal{T}_m^1),\mathcal{T}^2=(\mathcal{T}_1^2,\ldots,\mathcal{T}_m^2)$

$$h_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{x}_{1,\mathcal{T}_1^1},\mathbf{x}_{1,\mathcal{T}_1^2},\ldots,\mathbf{x}_{m,\mathcal{T}_m^1},\mathbf{x}_{m,\mathcal{T}_m^2})=\mathbb{E}_{P_1,\ldots,P_m}[h(\mathbf{X}_1,\ldots,\mathbf{X}_m)|X_{i,t}=x_{i,t},\forall i \text{ and } t\in\mathcal{T}_i^1\cup\mathcal{T}_i^2]$$

where each $\mathbf{x}_{i,\mathcal{T}_i^1}=(x_{i,t})_{t\in\mathcal{T}_i^1}$ and $\mathbf{x}_{i,\mathcal{T}_i^2}=(x_{i,t})_{t\in\mathcal{T}_i^2}$. Considering all possible subsets $\tilde{\mathcal{T}}_i^1$ of $\mathcal{T}_i^1$ for each i and denoting $\tilde{\mathcal{T}}^1=(\tilde{\mathcal{T}}_1^1,\ldots,\tilde{\mathcal{T}}_m^1)$, we define a centered conditional expectation (its property will be discussed momentarily)

$$\begin{aligned}&\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{x}_{1,\mathcal{T}_1^1},\mathbf{x}_{1,\mathcal{T}_1^2},\ldots,\mathbf{x}_{m,\mathcal{T}_m^1},\mathbf{x}_{m,\mathcal{T}_m^2})\\ =&\sum_{\tilde{\mathcal{T}}_i^1\subset\mathcal{T}_i^1,\forall i}(-1)^{\sum_i(|\mathcal{T}_i^1|-|\tilde{\mathcal{T}}_i^1|)}h_{\tilde{\mathcal{T}}^1,\mathcal{T}^2}(\mathbf{x}_{1,\tilde{\mathcal{T}}_1^1},\mathbf{x}_{1,\mathcal{T}_1^2},\ldots,\mathbf{x}_{m,\tilde{\mathcal{T}}_m^1},\mathbf{x}_{m,\mathcal{T}_m^2}).\end{aligned} \tag{B.8}$$

By expanding out the product measure $\prod_{i=1}^m\prod_{t\notin\mathcal{T}_i^1\cup\mathcal{T}_i^2}dP_i(x_{i,t})\prod_{i=1}^m\prod_{t\in\mathcal{T}_i^1}d(\hat{P}_i-P_i)(x_{i,t})$ and noticing that each $\hat{P}_i$ is a probability measure, $R_d(\mathcal{T}^1,\mathcal{T}^2)$ can be expressed as

$$\int\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{x}_{1,\mathcal{T}_1^1},\mathbf{x}_{1,\mathcal{T}_1^2},\ldots,\mathbf{x}_{m,\mathcal{T}_m^1},\mathbf{x}_{m,\mathcal{T}_m^2})\prod_{i=1}^m\prod_{t\in\mathcal{T}_i^1}d\hat{P}_i(x_{i,t})\prod_{i=1}^m\prod_{t\in\mathcal{T}_i^2}d(\mathbf{w}_i-\hat{P}_i)(x_{i,t}). \tag{B.9}$$

From now on, we denote $X_{i,j},i=1,\ldots,m,j=1,\ldots,n_i$as the observations, and for each i let

$$\begin{aligned}J_i^1&=(J_i^1(1),\ldots,J_i^1(|\mathcal{T}_i^1|))\in\{1,2,\ldots,n_i\}^{|\mathcal{T}_i^1|}\\ J_i^2&=(J_i^2(1),\ldots,J_i^2(|\mathcal{T}_i^2|))\in\{1,2,\ldots,n_i\}^{|\mathcal{T}_i^2|}\end{aligned}$$

be two sequences of indices (if $\mathcal{T}_i^1$ or $\mathcal{T}_i^2$ is empty, then J_i^1 or J_i^2 is empty accordingly) that specify the second subscript of data $X_{i,j}$. Then (B.9) can be written more explicitly as

$$R_d(\mathcal{T}^1,\mathcal{T}^2) = \sum_{J_1^2,\ldots,J_m^2} \Big[\prod_{i,t}\big(w_{i,J_i^2(t)} - \frac{1}{n_i}\big)\Big] \sum_{J_1^1,\ldots,J_m^1} \frac{1}{\prod_i n_i^{|\mathcal{T}_i^1|}} \tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,J_1^1},\mathbf{X}_{1,J_1^2},\ldots,\mathbf{X}_{m,J_m^1},\mathbf{X}_{m,J_m^2})$$

where each $\mathbf{X}_{i,J_i^1} = (X_{i,J_i^1(1)},\ldots,X_{i,J_i^1(|\mathcal{T}_i^1|)})$ contains the input data specified by J_i^1, and similarly $\mathbf{X}_{i,J_i^2} = (X_{i,J_i^2(1)},\ldots,X_{i,J_i^2(|\mathcal{T}_i^2|)})$. We bound the supremum as follows

$$\begin{aligned} |R_d(\mathcal{T}^1,\mathcal{T}^2)|^2 &\le \Big[\sum_{J_1^2,\ldots,J_m^2}\prod_{i,t}\big(w_{i,J_i^2(t)} - \frac{1}{n_i}\big)^2\Big]\Big[\sum_{J_1^2,\ldots,J_m^2}\Big(\frac{1}{\prod_i n_i^{|\mathcal{T}_i^1|}}\sum_{J_1^1,\ldots,J_m^1}\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}\Big)^2\Big] \\ &= \prod_{i=1}^m\left(\sum_{j=1}^{n_i}\left(w_{i,j}-\frac{1}{n_i}\right)^2\right)^{|\mathcal{T}_i^2|}\Big[\sum_{J_1^2,\ldots,J_m^2}\Big(\frac{1}{\prod_i n_i^{|\mathcal{T}_i^1|}}\sum_{J_1^1,\ldots,J_m^1}\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}\Big)^2\Big] \end{aligned}$$

where we suppress the arguments of $\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}$, and use the Cauchy-Schwartz inequality. The upper bound from Lemma B.2.2 then implies that $\sum_{j=1}^{n_i}\left(w_{i,j}-\frac{1}{n_i}\right)^2 \le u(\alpha)^2\mathcal{X}^2_{1,1-\alpha}/n_i^2$, and hence for some constant C_1 depending on α and d

$$\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}|R_d(\mathcal{T}^1,\mathcal{T}^2)|^2 \le C_1\prod_{i=1}^m n_i^{-2|\mathcal{T}_i^2|}\cdot\Big[\sum_{J_1^2,\ldots,J_m^2}\Big(\frac{1}{\prod_i n_i^{|\mathcal{T}_i^1|}}\sum_{J_1^1,\ldots,J_m^1}\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}\Big)^2\Big]. \tag{B.10}$$

From (B.10), the proof now boils down to bounding the expectation of

$$\Big(\frac{1}{\prod_i n_i^{|\mathcal{T}_i^1|}}\sum_{J_1^1,\ldots,J_m^1}\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,J_1^1},\mathbf{X}_{1,J_1^2},\ldots,\mathbf{X}_{m,J_m^1},\mathbf{X}_{m,J_m^2})\Big)^2$$

for each fixed $J_1^2,\ldots,J_m^2$. We need a few properties of $\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}$. The first property, which follows from its definition, is that, for any i and $t\in\mathcal{T}_i^1$, the marginal expectation under the true input distributions is zero, i.e.

$$\int \tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{x}_{1,\mathcal{T}_1^1},\mathbf{x}_{1,\mathcal{T}_1^2},\ldots,\mathbf{x}_{m,\mathcal{T}_m^1},\mathbf{x}_{m,\mathcal{T}_m^2})dP_i(x_{i,t}) = 0. \tag{B.11}$$

The second property is a bound of the second moment that is uniform in $\mathcal{T}^1, \mathcal{T}^2$. By Jensen's inequality, one can show that for any m sequences of indices $I_i = (I_i(1), \ldots, I_i(|\mathcal{T}_i^1| + |\mathcal{T}_i^2|)) \in \big\{1, 2, \ldots, |\mathcal{T}_i^1| + |\mathcal{T}_i^2|\big\}^{|\mathcal{T}_i^1|+|\mathcal{T}_i^2|}$ the conditional expectation $h_{\mathcal{T}^1,\mathcal{T}^2}$ satisfies

$$\mathbb{E}_{P_1,\ldots,P_m}[h^2_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,I_1}, \ldots, \mathbf{X}_{m,I_m})] \leq \mathcal{M}$$

where $\mathbf{X}_{i,I_i} = (X_i(I_i(1)), \ldots, X_i(I_i(|\mathcal{T}_i^1| + |\mathcal{T}_i^2|)))$ and $\mathcal{M}$ is the second moment bound defined in (B.6). (B.8) tells us that $\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}$ is the sum of $2^{|\mathcal{T}^1|}$ conditional expectations of such type. By the Minkowski inequality we have

$$\mathbb{E}_{P_1,\ldots,P_m}[\tilde{h}^2_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,I_1}, \ldots, \mathbf{X}_{m,I_m})] \leq 4^{|\mathcal{T}^1|}\mathcal{M}. \tag{B.12}$$

Now we are able to proceed with

$$\begin{aligned}
&\mathbb{E}_D\Big(\frac{1}{\prod_i n_i^{|\mathcal{T}_i^1|}} \sum_{J_1^1,\ldots,J_m^1} \tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,J_1^1}, \mathbf{X}_{1,J_1^2}, \ldots, \mathbf{X}_{m,J_m^1}, \mathbf{X}_{m,J_m^2})\Big)^2 \\
=&\frac{1}{\prod_i n_i^{2|\mathcal{T}_i^1|}} \sum_{J_1^1,\ldots,J_m^1} \sum_{\tilde{J}_1^1,\ldots,\tilde{J}_m^1} \mathbb{E}_D[\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,J_1^1}, \mathbf{X}_{1,J_1^2}, \ldots, \mathbf{X}_{m,J_m^1}, \mathbf{X}_{m,J_m^2}) \cdot \\
&\qquad\qquad \tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,\tilde{J}_1^1}, \mathbf{X}_{1,J_1^2}, \ldots, \mathbf{X}_{m,\tilde{J}_m^1}, \mathbf{X}_{m,J_m^2})].
\end{aligned} \tag{B.13}$$

Note that because of property (B.11), the expectation in (B.13) is zero if there is some index $i^* \in \{1, \ldots, m\}$ and $j^* \in \{1, \ldots, n_i\}$ such that X_{i^*,j^*} does not appear in $\mathbf{X}_{i^*,J_{i^*}^2}$ and shows up exactly once among $\mathbf{X}_{i^*,J_{i^*}^1}, \mathbf{X}_{i^*,\tilde{J}_{i^*}^1}$. Note that, for each fixed $i = 1, \ldots, m$, the number of choices of $J_i^1, \tilde{J}_i^1$ that avoid this occurrence is no more than $C_2 n_i^{|\mathcal{T}_i^1|}$, where C_2 is some constant depending on d only. So the total number of choices of $J_i^1, \tilde{J}_i^1, i = 1, \ldots, m$ that can possibly produce a nonzero expectation in (B.13) is at most

$$C_2^m\Big(\prod_{i=1}^m n_i^{|\mathcal{T}_i^1|}\Big). \tag{B.14}$$

On the other hand, applying the Cauchy-Schwartz inequality and the upper bound (B.12) to the

expectation in (B.13) gives

$$\Big|\mathbb{E}_D[\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,J_1^1},\mathbf{X}_{1,J_1^2},\ldots,\mathbf{X}_{m,J_m^1},\mathbf{X}_{m,J_m^2})\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,\tilde{J}_1^1},\mathbf{X}_{1,J_1^2},\ldots,\mathbf{X}_{m,\tilde{J}_m^1},\mathbf{X}_{m,J_m^2})]\Big| \leq 4^{|\mathcal{T}^1|}\mathcal{M}$$

for any $J_i^1, \tilde{J}_i^1, J_i^2, i = 1,\ldots,m$. We conclude from (B.13), (B.14) and the above bound that

$$\mathbb{E}_D\Big(\frac{1}{\prod_i n_i^{|\mathcal{T}_i^1|}}\sum_{J_1^1,\ldots,J_m^1}\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}(\mathbf{X}_{1,J_1^1},\mathbf{X}_{1,J_1^2},\ldots,\mathbf{X}_{m,J_m^1},\mathbf{X}_{m,J_m^2})\Big)^2 \leq \frac{4^{|\mathcal{T}^1|}C_2^m\mathcal{M}}{\prod_i n_i^{|\mathcal{T}_i^1|}} \tag{B.15}$$

uniformly for all choices of $J_i^2, i = 1,\ldots,m$.

Finally, we go back to the inequality (B.10) to arrive at

$$\begin{aligned}
\mathbb{E}_D\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}|R_d(\mathcal{T}^1,\mathcal{T}^2)|^2\big] &\leq C_1\prod_{i=1}^m n_i^{-2|\mathcal{T}_i^2|}\cdot\Big[\sum_{J_1^2,\ldots,J_m^2}\mathbb{E}_D\Big(\prod_i\frac{1}{n_i^{|\mathcal{T}_i^1|}}\sum_{J_1^1,\ldots,J_m^1}\tilde{h}_{\mathcal{T}^1,\mathcal{T}^2}\Big)^2\Big]\\
&\leq C_1\prod_{i=1}^m n_i^{-2|\mathcal{T}_i^2|}\cdot\Big[\sum_{J_1^2,\ldots,J_m^2}\frac{4^{|\mathcal{T}^1|}C_2^m\mathcal{M}}{\prod_i n_i^{|\mathcal{T}_i^1|}}\Big]\\
&\leq 4^{|\mathcal{T}^1|}C_1C_2^m\mathcal{M}\prod_{i=1}^m n_i^{-(|\mathcal{T}_i^1|+|\mathcal{T}_i^2|)}.
\end{aligned} \tag{B.16}$$

This proves (B.7). Note that, since T is fixed, from (B.5),

$$\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}\Big|\sum_{\mathcal{T}^1,\mathcal{T}^2,d\geq 2}R_d(\mathcal{T}^1,\mathcal{T}^2)\Big| \leq \sum_{\mathcal{T}^1,\mathcal{T}^2,d\geq 2}\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}|R_d(\mathcal{T}^1,\mathcal{T}^2)|$$

and the Minkowski inequality we conclude that $\mathbb{E}_D\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}|\sum_{\mathcal{T}^1,\mathcal{T}^2,d\geq 2}R_d(\mathcal{T}^1,\mathcal{T}^2)|^2\big] = O(n^{-2})$. This therefore shows that $\mathbb{E}_D\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}|Z - Z_L|^2\big] = O(n^{-2})$ as the data size $n\to\infty$.

Now we prove the uniform approximation error of $\widehat{Z_L}$. The approach is to expand the integral form of $Z(\mathbf{w}_1,\ldots,\mathbf{w}_m)$ in a similar way to (B.5), but around $\hat{P}_i$'s instead of P_i's

$$Z(\mathbf{w}_1,\ldots,\mathbf{w}_m) = \sum_{d=0}^T\sum_{\sum_i|\mathcal{T}_i^2|=d}\int h(\mathbf{x}_1,\ldots,\mathbf{x}_m)\prod_{i=1}^m\prod_{t\notin\mathcal{T}_i^2}d\hat{P}_i(x_{i,t})\prod_{i=1}^m\prod_{t\in\mathcal{T}_i^2}d(\mathbf{w}_i-\hat{P}_i)(x_{i,t}) \tag{B.17}$$

where each $\mathcal{T}_i^2$ is again an ordered subset of $\{1, 2, \ldots, T_i\}$ that contains the second subscript t of the argument $x_{i,t}$. Similar to above, summands with $d = 0, 1$ gives the linear approximation at the empirical distributions, i.e. $\widehat{Z_L}(\mathbf{w}_1, \ldots, \mathbf{w}_m)$, and all summands with $d \geq 2$ will be the associated approximation error. To bound each summand with $d \geq 2$, we rewrite $\hat{P}_i$ as $\hat{P}_i - P_i + P_i$, and suitably expand out the product measure $\prod_{i=1}^m \prod_{t \notin \mathcal{T}_i^2} d\hat{P}_i(x_{i,t})$ in (B.17) to get

$$
\begin{aligned}
& \int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i=1}^m \prod_{t \notin \mathcal{T}_i^2} d\hat{P}_i(x_{i,t}) \prod_{i=1}^m \prod_{t \in \mathcal{T}_i^2} d(\mathbf{w}_i - \hat{P}_i)(x_{i,t}) \\
= & \sum_{\mathcal{T}_i^1, \text{ s.t.} \mathcal{T}_i^1 \cap \mathcal{T}_i^2 = \emptyset} \int h(\mathbf{x}_1, \ldots, \mathbf{x}_m) \prod_{i=1}^m \prod_{t \notin \mathcal{T}_i^1 \cup \mathcal{T}_i^2} dP_i(x_{i,t}) \prod_{i=1}^m \prod_{t \in \mathcal{T}_i^1} d(\hat{P}_i - P_i)(x_{i,t}) \prod_{i=1}^m \prod_{t \in \mathcal{T}_i^2} d(\mathbf{w}_i - \hat{P}_i)(x_{i,t}) \\
= & \sum_{\mathcal{T}_i^1, \text{ s.t.} \mathcal{T}_i^1 \cap \mathcal{T}_i^2 = \emptyset} R_{|\mathcal{T}^1|+d}(\mathcal{T}^1, \mathcal{T}^2)
\end{aligned}
$$

where each $\mathcal{T}_i^1$ is the ordered set consisting of the second subscripts t of all $x_{i,t}$'s to which $\hat{P}_i - P_i$ is distributed, and $R_{|\mathcal{T}^1|+d}(\mathcal{T}^1, \mathcal{T}^2)$ is the remainder term defined before. The desired conclusion then follows from (B.7) and an argument analogous to the first part of the theorem. □

B.3 Proof of Results in Section 3.4.3

Proof of Theorem 3.4.3. To simplify the proof, we first argue that one can assume $\mathrm{Var}(Y_i) > 0$ and $\mathbb{E}Y_i = 0$ for all $i = 1, \ldots, m$ without loss of generality. Let $I = \{i : \mathrm{Var}(Y_i) > 0, i = 1, \ldots, m\}$ be the set of indices whose corresponding Y_i's have non-zero variances. Then for $i \notin I$ each $Y_{i,j} = \mathbb{E}Y_i$

almost surely, hence

$$
\begin{aligned}
&R(\mu_0)\\
=&\max\left\{\prod_{i=1}^{m}\prod_{j=1}^{n_i} n_i w_{i,j} \,\middle|\, \sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{i,j}w_{i,j} = \mu_0,\ \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for all } i,\ w_{i,j}\geq 0 \text{ for all } i,j\right\}\\
=&\max\left\{\prod_{i=1}^{m}\prod_{j=1}^{n_i} n_i w_{i,j} \,\middle|\, \sum_{i\in I}\sum_{j=1}^{n_i} Y_{i,j}w_{i,j} = \mu_0 - \sum_{i\notin I}\mathbb{E}Y_i,\ \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for all } i,\ w_{i,j}\geq 0 \text{ for all } i,j\right\}\\
=&\max\left\{\prod_{i\in I}\prod_{j=1}^{n_i} n_i w_{i,j} \,\middle|\, \sum_{i\in I}\sum_{j=1}^{n_i} Y_{i,j}w_{i,j} = \sum_{i\in I}\mathbb{E}Y_i,\ \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for } i\in I,\ w_{i,j}\geq 0 \text{ for all } i\in I,j\right\}\\
=&\max\left\{\prod_{i\in I}\prod_{j=1}^{n_i} n_i w_{i,j} \,\middle|\, \sum_{i\in I}\sum_{j=1}^{n_i} (Y_{i,j}-\mathbb{E}Y_i)w_{i,j} = 0,\ \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for } i\in I,\ w_{i,j}\geq 0 \text{ for all } i\in I,j\right\}\\
=&R_I(0)
\end{aligned}
$$

where $R_I(0)$ is the analog of $R(\mu_0)$ defined for the translated observations $\{Y_{i,1}-\mathbb{E}Y_i,\ldots,Y_{i,n_i}-\mathbb{E}Y_i\}, i\in I$, and in the third equality we put $w_{i,j}=1/n_i$ for $i\notin I$ into the objective, which can be easily seen to be the maximizing weights for $i\notin I$. Therefore, to prove the theorem for $R(\mu_0)$, one can work with $R_I(0)$ instead, and note that the change of m, the number of independent distributions, does not affect the limit chi-square distribution.

In view of the above, we shall assume $\mathrm{Var}(Y_i)>0$ and $\mathbb{E}Y_i=0$ for each i, hence $R(\mu_0)$ is just $R(0)$. Introducing a slack variable μ_i for each $\sum_{j=1}^{n_i} Y_{i,j}w_{i,j}$ and taking the negative logarithm of the objective convert the defining maximization of $R(0)$ to the following convex program

$$
\begin{aligned}
\min_{\mathbf{w}_1,\ldots,\mathbf{w}_m,\boldsymbol{\mu}}\quad & -\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j})\\
\text{subject to}\quad & \sum_{j=1}^{n_i} Y_{i,j}w_{i,j} = \mu_i,\ i=1,\ldots,m\\
& \sum_{j=1}^{n_i} w_{i,j} = 1,\ i=1,\ldots,m\\
& \sum_{i=1}^{m}\mu_i = 0
\end{aligned}
\tag{B.18}
$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$. The non-negativity constraints $w_{i,j} \geq 0$ are dropped since they are implicitly imposed in the objective function.

Step one: We prove that, with probability tending to one, Slater's condition holds for (B.18). In other words, consider the event

$$\mathcal{S} = \left\{ \begin{array}{l} \text{(B.18) has at least one feasible solution} \\ (\mathbf{w}_1, \ldots, \mathbf{w}_m, \boldsymbol{\mu}) \text{ such that } w_{i,j} > 0 \text{ for all } i, j \end{array} \right\}$$

and we prove $P(\mathcal{S}) \to 1$ as $n \to \infty$. To this end, consider the following events indexed by i

$$\tilde{\mathcal{S}}_i = \left\{ \min_{j=1,\ldots,n_i} Y_{i,j} < 0 < \max_{j=1,\ldots,n_i} Y_{i,j} \right\}.$$

We shall prove that $P(\tilde{\mathcal{S}}_i) \to 1$ for all i and that $\cap_{i=1}^m \tilde{\mathcal{S}}_i \subseteq \mathcal{S}$, which imply that $P(\mathcal{S}) \to 1$ because

$$P(\mathcal{S}^c) \leq P((\cap_{i=1}^m \tilde{\mathcal{S}}_i)^c) = P(\cup_{i=1}^m \tilde{\mathcal{S}}_i^c) \leq \sum_{i=1}^m P(\tilde{\mathcal{S}}_i^c) = \sum_{i=1}^m (1 - P(\tilde{\mathcal{S}}_i)) \to 0.$$

Note that $\mathrm{Var}(Y_i) > 0$ and $\mathbb{E}Y_i = 0$ imply $P(Y_i \geq 0) < 1, P(Y_i \leq 0) < 1$. Hence as $n \to \infty$

$$P\big(\min_{j=1,\ldots,n_i} Y_{i,j} \geq 0 \big) = \prod_{j=1}^{n_i} P(Y_{i,j} \geq 0) = (P(Y_i \geq 0))^{n_i} \to 0$$

which is equivalently $P\big(\min_j Y_{i,j} < 0 \big) \to 1$. Similarly, $P\big(\max_j Y_{i,j} > 0 \big) \to 1$ holds. Combining these two limits gives $P(\tilde{\mathcal{S}}_i) \to 1$. To show $\cap_{i=1}^m \tilde{\mathcal{S}}_i \subseteq \mathcal{S}$, note that if $\tilde{\mathcal{S}}_i$ happens then there must exist convex-combination weights $w_{i,j} > 0, \sum_{j=1}^{n_i} w_{i,j} = 1$ such that $\sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = 0$. When all $\tilde{\mathcal{S}}_i$'s happen, one can take such weights and $\mu_i = 0$ for each i to see that $\mathcal{S}$ also happens.

Step two: We derive the KKT conditions for (B.18), conditioned on Slater's condition $\mathcal{S}$. Notice that each $-\log(n_i w_{i,j})$ is bounded below by $-\log n_i$, and when $w_{i,j} \to 0$ for some i, j the corresponding $-\log(n_i w_{i,j}) \to +\infty$, hence the objective $-\sum_{i,j} \log(n_i w_{i,j}) \to +\infty$ as $\min_{i,j} w_{i,j} \to 0$. Therefore, the optimal solution, if it exists, must lie in the region where $\min_{i,j} w_{i,j} \geq \epsilon$ for some small $\epsilon > 0$ that depends on n_i's. Since the set $\{(\mathbf{w}_1, \ldots, \mathbf{w}_m) : \sum_{i=1}^m \sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = 0, \sum_{j=1}^{n_i} w_{i,j} = 1, w_{i,j} \geq \epsilon \text{ for all } i, j\}$ is compact, an optimal solution $(\mathbf{w}_1^*, \ldots, \mathbf{w}_m^*, \boldsymbol{\mu}^*)$ exists for (B.18). Moreover,

strict convexity of the objective forces the optimal solution to be unique. By Corollary 28.3.1 of Rockafellar (1970), there must exist Lagrange multipliers $(\boldsymbol{\lambda}_{\mathbf{1}}^*, \boldsymbol{\lambda}_{\mathbf{2}}^*, \lambda^*)$, where $\boldsymbol{\lambda}_{\mathbf{1}}^* = (\lambda_{1,1}, \ldots, \lambda_{1,m})$ is associated with the first m constraints, $\boldsymbol{\lambda}_{\mathbf{2}}^* = (\lambda_{2,1}, \ldots, \lambda_{2,m})$ with the second m constraints, and λ^* with the last constraint in (B.18), that together with the optimal solution $(\mathbf{w}_1^*, \ldots, \mathbf{w}_m^*, \boldsymbol{\mu}^*)$ satisfy the following KKT conditions

$$\begin{aligned}
\sum_{j=1}^{n_i} Y_{i,j} w_{i,j}^* &= \mu_i^*, \text{ for } i = 1, \ldots, m \\
\sum_{j=1}^{n_i} w_{i,j}^* &= 1, \text{ for } i = 1, \ldots, m \\
\sum_{i=1}^{m} \mu_i^* &= 0 \\
-\frac{1}{w_{i,j}^*} + Y_{i,j}\lambda_{1,i}^* + \lambda_{2,i}^* &= 0, \text{ for all } i, j \\
-\lambda_{1,i}^* + \lambda^* &= 0, \text{ for } i = 1, \ldots, m.
\end{aligned} \tag{B.19}$$

Some basic algebra shows $\lambda_{2,i}^* = n_i - \lambda_{1,i}^* \mu_i^*, \lambda^* = \lambda_{1,i}^*$ for all i, hence it follows from (B.19) that

$$w_{i,j}^* = \frac{1}{n_i + \lambda^*(Y_{i,j} - \mu_i^*)} \tag{B.20}$$

and λ^*, μ_i^* satisfy

$$\sum_{j=1}^{n_i} \frac{Y_{i,j} - \mu_i^*}{n_i + \lambda^*(Y_{i,j} - \mu_i^*)} = 0, \text{ for } i = 1, \ldots, m \tag{B.21}$$

$$\sum_{i=1}^{m} \mu_i^* = 0. \tag{B.22}$$

A note on Slater's condition: Note that $\lambda^*, \mu_i^*, i = 1, \ldots, m$ are guaranteed to exist and defined by (B.21) and (B.22) only when Slater's condition $\mathcal{S}$ holds. To make these variables well defined regardless of the Slater's condition, we simply let $\lambda^*, \mu_i^*, i = 1, \ldots, m$ all be 0 when $\mathcal{S}$ does not hold in the rest of the proof. Every intermediate inequality/equality below related to $\lambda^*, \mu_i^*, i = 1, \ldots, m$ is interpreted as restricted to the event of $\mathcal{S}$. For example, $a \leq b$ and $a = b$

should be interpreted as $a \cdot \mathbf{1}\{\mathcal{S}\} \leq b \cdot \mathbf{1}\{\mathcal{S}\}$ and $a \cdot \mathbf{1}\{\mathcal{S}\} = b \cdot \mathbf{1}\{\mathcal{S}\}$. All asymptotic statements or quantities that rely on stochastic orders o_p, O_p and convergence in distribution, remain valid via a decomposition of the involved probability into $\mathcal{S}$ and $\mathcal{S}^c$ and using $P(\mathcal{S}) \to 1$. To demonstrate this argument concretely, we will show as an example in (B.43) how it works. But to avoid adding overwhelming complexities to our proof, we will keep this aspect silent until then.

Step three: We show that the Lagrange multiplier λ^* has a magnitude of $O_p(n^{1/2})$. Write (B.20) as

$$\frac{1}{n_i + \lambda^*(Y_{i,j} - \mu_i^*)} = \frac{1}{n_i}\left(1 - \frac{\frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)}{1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)}\right) \tag{B.23}$$

and substituting (B.23) into (B.21) gives

$$\bar{Y}_i - \mu_i^* = \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{\frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)^2}{1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)}, \tag{B.24}$$

where $\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}$. Multiply both sides by $\text{sign}(\lambda^*)$ to make the right hand side positive

$$\text{sign}(\lambda^*)(\bar{Y}_i - \mu_i^*) = \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{\frac{|\lambda^*|}{n_i}(Y_{i,j} - \mu_i^*)^2}{1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)}. \tag{B.25}$$

This is because, since each $w_{i,j}^*$ is strictly positive, from (B.20) we must have $1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*) > 0$, $\forall i, j$. Also note that $|\mu_i^*| = \left|\sum_{j=1}^{n_i} Y_{i,j} w_{i,j}^*\right| \leq \sum_{j=1}^{n_i} w_{i,j}^* |Y_{i,j}| \leq \max_{j=1,\ldots,n_i} |Y_{i,j}|$. Let $Z_N = \max_{i=1,\ldots,m, j=1,\ldots,n_i} |Y_{i,j}|$, and $0 < \underline{c} \leq \bar{c} < \infty$ be constants such that $\underline{c} \leq n_i/n \leq \bar{c}$ for all i. The existence of $\underline{c}$ and $\bar{c}$ is guaranteed under Assumption 3.3.1. A lower bound of the right hand side

of (B.25) can be derived as follows

$$\begin{aligned}
\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\frac{|\lambda^*|}{n_i}(Y_{i,j}-\mu_i^*)^2}{1+\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*)} &\geq \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\frac{|\lambda^*|}{n_i}(Y_{i,j}-\mu_i^*)^2}{1+\frac{|\lambda^*|}{n_i}\left|Y_{i,j}-\mu_i^*\right|} \\
&\geq \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\frac{|\lambda^*|}{n_i}(Y_{i,j}-\mu_i^*)^2}{1+\frac{|\lambda^*|}{n_i}\cdot 2\max_{j'=1,\ldots,n_i}\left|Y_{i,j'}\right|} \\
&\geq \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{\frac{|\lambda^*|}{n\cdot\bar{c}/\underline{c}}(Y_{i,j}-\mu_i^*)^2}{1+\frac{|\lambda^*|}{n\cdot\underline{c}/\bar{c}}\cdot 2Z_N} \\
&= \frac{\frac{|\lambda^*|}{n\cdot\bar{c}/\underline{c}}}{1+\frac{|\lambda^*|}{n\cdot\underline{c}/\bar{c}}\cdot 2Z_N}\left(\hat{\sigma}_i^2-2\bar{Y}_i\mu_i^*+\mu_i^{*2}\right) \text{ where } \hat{\sigma}_i^2=\frac{1}{n_i}\sum_{j=1}^{n_i}Y_{i,j}^2 \\
&\geq \frac{\frac{|\lambda^*|}{n\cdot\bar{c}/\underline{c}}}{1+\frac{|\lambda^*|}{n\cdot\underline{c}/\bar{c}}\cdot 2Z_N}\left(\hat{\sigma}_i^2-2\bar{Y}_i\mu_i^*\right) \qquad \text{(B.26)}
\end{aligned}$$

Applying Lemma 11.2 in Owen (2001) to $\{Y_{i,1},\ldots,Y_{i,n_i}\}$ reveals that, almost surely, we have $\max_{j=1,\ldots,n_i}|Y_{i,j}|=o(n_i^{\frac{1}{2}})$ as $n_i\to\infty$ for each i, hence $Z_N=o(n^{\frac{1}{2}})$ and $\mu_i^*=o(n^{\frac{1}{2}})$ almost surely. By the central limit theorem, each $\bar{Y}_i=O_p(n_i^{-\frac{1}{2}})=O_p(n^{-\frac{1}{2}})$. Substituting the lower bound (B.26) into (B.25) and multiplying each side by $1+\frac{|\lambda^*|}{n\cdot\underline{c}/\bar{c}}\cdot 2Z_N$ give

$$\begin{aligned}
\left(1+\frac{|\lambda^*|}{n\cdot\underline{c}/\bar{c}}\cdot 2Z_N\right)\operatorname{sign}(\lambda^*)\left(\bar{Y}_i-\mu_i^*\right) &\geq \frac{|\lambda^*|}{n\cdot\bar{c}/\underline{c}}\left(\hat{\sigma}_i^2-2\bar{Y}_i\mu_i^*\right) \\
&\geq \frac{|\lambda^*|}{n\cdot\bar{c}/\underline{c}}(\hat{\sigma}_i^2+O_p(n^{-\frac{1}{2}})o(n^{\frac{1}{2}})) \qquad \text{(B.27)} \\
&= \frac{|\lambda^*|}{n\cdot\bar{c}/\underline{c}}(\hat{\sigma}_i^2+o_p(1)). \qquad \text{(B.28)}
\end{aligned}$$

Summing up both sides of (B.28) over $i=1,\ldots,m$, and using (B.22) and $Z_N=o(n^{\frac{1}{2}})$ we have

$$\left(1+\frac{|\lambda^*|}{n}o(n^{\frac{1}{2}})\right)\operatorname{sign}(\lambda^*)\sum_{i=1}^{m}\bar{Y}_i\geq\frac{|\lambda^*|}{n\cdot\bar{c}/\underline{c}}\left(\sum_{i=1}^{m}\hat{\sigma}_i^2+o_p(1)\right). \qquad \text{(B.29)}$$

Rearranging the terms gives

$$\frac{|\lambda^*|}{n}\left(\frac{\underline{c}}{\bar{c}}\sum_{i=1}^{m}\hat{\sigma}_i^2+o_p(1)+o(n^{\frac{1}{2}})\sum_{i=1}^{m}\bar{Y}_i\right)\leq\left|\sum_{i=1}^{m}\bar{Y}_i\right|. \qquad \text{(B.30)}$$

Note that $\hat{\sigma}_i^2 \to \sigma_i^2 := Var(Y_i)$ almost surely by the strong law of large numbers, and $\sum_{i=1}^m \bar{Y}_i = \sum_{i=1}^m O_p(n_i^{-\frac{1}{2}}) = O_p(n^{-\frac{1}{2}})$. By the assumption $\sum_{i=1}^m \sigma_i^2 > 0$, (B.30) implies

$$\frac{|\lambda^*|}{n} \leq \frac{O_p(n^{-\frac{1}{2}})}{\frac{c}{\bar{c}} \sum_{i=1}^m \sigma_i^2 + o_p(1)}.$$

That is, $\frac{|\lambda^*|}{n} = O_p(n^{-\frac{1}{2}})$.

Step four: We show the convergence of μ_i^* to the true mean 0, i.e., $\mu_i^* = o_p(1)$. From (B.23) it follows that

$$\begin{aligned} \bar{Y}_i - \mu_i^* &= \sum_{j=1}^{n_i} \left(\frac{1}{n_i} - w_{i,j}^* \right) Y_{i,j} \\ &= \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)}{1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)} Y_{i,j}. \end{aligned} \tag{B.31}$$

We have shown in the Step three that $Z_N = o(n^{\frac{1}{2}})$, $|\mu_i^*| \leq Z_N$ and $\frac{|\lambda^*|}{n} = O_p(n^{-\frac{1}{2}})$. Hence $\max_j \left| \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*) \right| = O\big(\frac{|2\lambda^*|}{n} Z_N\big) = O_p(n^{-\frac{1}{2}}) o(n^{\frac{1}{2}}) = o_p(1)$. Therefore

$$\begin{aligned} \left| \bar{Y}_i - \mu_i^* \right| &\leq \frac{1}{n_i} \sum_{j=1}^{n_i} \left| \frac{\frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)}{1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)} \right| |Y_{i,j}| \\ &\leq \frac{1}{n_i} \sum_{j=1}^{n_i} \left| \frac{\max_{j'} \left| \frac{\lambda^*}{n_i}(Y_{i,j'} - \mu_i^*) \right|}{1 - \max_{j'} \left| \frac{\lambda^*}{n_i}(Y_{i,j'} - \mu_i^*) \right|} \right| |Y_{i,j}| \, \mathbf{1} \left\{ \max_{j'} \left| \frac{\lambda^*}{n_i}(Y_{i,j'} - \mu_i^*) \right| < 1 \right\} \\ &\quad + \frac{1}{n_i} \sum_{j=1}^{n_i} \left| \frac{\frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)}{1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)} \right| |Y_{i,j}| \cdot \mathbf{1} \left\{ \max_{j'} \left| \frac{\lambda^*}{n_i}(Y_{i,j'} - \mu_i^*) \right| \geq 1 \right\} \\ &= \left| \frac{\max_{j'} \left| \frac{\lambda^*}{n_i}(Y_{i,j'} - \mu_i^*) \right|}{1 - \max_{j'} \left| \frac{\lambda^*}{n_i}(Y_{i,j'} - \mu_i^*) \right|} \right| \mathbf{1} \left\{ \max_{j'} \left| \frac{\lambda^*}{n_i}(Y_{i,j'} - \mu_i^*) \right| < 1 \right\} \frac{1}{n_i} \sum_{j=1}^{n_i} |Y_{i,j}| + o_p(1) \qquad \text{(B.32)} \\ &\leq \left| \frac{o_p(1)}{1 - o_p(1)} \right| \cdot \frac{1}{n_i} \sum_{j=1}^{n_i} |Y_{i,j}| + o_p(1) \\ &= o_p(1). \end{aligned}$$

We explain the $o_p(1)$ term in line (B.32) through a general statement:

Lemma B.3.1 *Let $\xi_{n_1,\ldots,n_m} \in [-\infty, +\infty]$ be a sequence of random variables indexed by $n_1, \ldots, n_m$,*

and $\mathcal{E}_{n_1,\ldots,n_m}$ *be a sequence of events such that* $P(\mathcal{E}_{n_1,\ldots,n_m}) \to 0$ *as all* $n_i \to \infty$*. Suppose further that* $\xi_{n_1,\ldots,n_m} = -\infty$ *or* $+\infty$ *only when* $\mathcal{E}_{n_1,\ldots,n_m}$ *occurs, so that* $\xi_{n_1,\ldots,n_m} \cdot \mathbf{1}\{\mathcal{E}_{n_1,\ldots,n_m}\}$ *does not involve the indeterminate operation* $\infty \cdot 0$*. Then for every sequence of constants* $a_{n_1,\ldots,n_m} > 0$*, we have* $\xi_{n_1,\ldots,n_m} \cdot \mathbf{1}\{\mathcal{E}_{n_1,\ldots,n_m}\} = o_p(a_{n_1,\ldots,n_m})$ *as all* $n_i \to \infty$*.*

Proof of Lemma B.3.1. For any $\epsilon > 0$ we write $P(|\xi_{n_1,\ldots,n_m}| \cdot \mathbf{1}\{\mathcal{E}_{n_1,\ldots,n_m}\} > \epsilon a_{n_1,\ldots,n_m}) \leq P(|\xi_{n_1,\ldots,n_m}| \cdot \mathbf{1}\{\mathcal{E}_{n_1,\ldots,n_m}\} \neq 0) \leq P(\mathcal{E}_{n_1,\ldots,n_m}) \to 0$. The desired conclusion follows from the definition of $o_p(\cdot)$. □

The $o_p(1)$ in (B.32) follows from Lemma B.3.1 by taking $\xi_{n_1,\ldots,n_m} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left|\frac{\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*)}{1+\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*)}\right||Y_{i,j}|$ and $a_{n_1,\ldots,n_m} = 1$, and noting that $P\big(\max_{j'}\left|\frac{\lambda^*}{n_i}(Y_{i,j'}-\mu_i^*)\right| \geq 1\big) \to 0$ because $\max_{j'}\left|\frac{\lambda^*}{n_i}(Y_{i,j'}-\mu_i^*)\right| = o_p(1)$.

On the other hand, $\bar{Y}_i = o_p(1)$ by the law of large numbers. Hence $\mu_i^* = o_p(1)$.

Step five: We derive formula (B.39) for the Lagrange multiplier λ^* in terms of the data. Rewrite (B.24) as

$$\begin{aligned}\bar{Y}_i - \mu_i^* &= \frac{1}{n_i}\sum_{j=1}^{n_i}\left[\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*)^2 - \frac{(\frac{\lambda^*}{n_i})^2(Y_{i,j}-\mu_i^*)^3}{1+\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*)}\right], \\ &= \frac{\lambda^*}{n_i}\left[\frac{1}{n_i}\sum_{j=1}^{n_i}(Y_{i,j}-\mu_i^*)^2\right] - \left|\frac{\lambda^*}{n_i}\right|^2\frac{1}{n_i}\sum_{j=1}^{n_i}\frac{(Y_{i,j}-\mu_i^*)^3}{1+\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*)}. \qquad \text{(B.33)}\end{aligned}$$

The second term in (B.33) can be bounded as

$$
\begin{aligned}
&\leq \left|\frac{\lambda^*}{n_i}\right|^2 \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{|Y_{i,j}-\mu_i^*|^3}{\left|1+\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*)\right|} \\
&\leq \left|\frac{\lambda^*}{n_i}\right|^2 \cdot \frac{2Z_N}{1-\max_{j'}\left|\frac{\lambda^*}{n_i}(Y_{i,j'}-\mu_i^*)\right|} \cdot \frac{1}{n_i}\sum_{j=1}^{n_i} |Y_{i,j}-\mu_i^*|^2 \mathbf{1}\left\{\max_{j'}\left|\frac{\lambda^*}{n_i}(Y_{i,j'}-\mu_i^*)\right| < 1\right\} \\
&\quad + \left|\frac{\lambda^*}{n_i}\right|^2 \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{|Y_{i,j}-\mu_i^*|^3}{\left|1+\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*)\right|} \cdot \mathbf{1}\left\{\max_{j'}\left|\frac{\lambda^*}{n_i}(Y_{i,j'}-\mu_i^*)\right| \geq 1\right\} \\
&\leq \left|\frac{\lambda^*}{n_i}\right|^2 \cdot \frac{2Z_N}{1-\max_{j'}\left|\frac{\lambda^*}{n_i}(Y_{i,j'}-\mu_i^*)\right|} \cdot \frac{1}{n_i}\sum_{j=1}^{n_i} |Y_{i,j}-\mu_i^*|^2 + o_p(n^{-\frac{1}{2}}) \qquad \text{(B.34)} \\
&= O_p(n^{-1})\frac{o(n^{1/2})}{1-o_p(1)}O_p(1) + o_p(n^{-\frac{1}{2}}) \qquad \text{(B.35)} \\
&= o_p(n^{-\frac{1}{2}})
\end{aligned}
$$

where in passing from line (B.34) to line (B.35), we use $\frac{1}{n_i}\sum_{j=1}^{n_i}|Y_{i,j}-\mu_i^*|^2 = O_p(1)$ which holds because

$$
\frac{1}{n_i}\sum_{j=1}^{n_i}|Y_{i,j}-\mu_i^*|^2 = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}^2 - 2\bar{Y}_i\mu_i^* + \mu_i^{*2} = \sigma_i^2 + O_p\big(n^{-\frac{1}{2}}\big)o_p(1) + o_p(1) = \sigma_i^2 + o_p(1), \quad \text{(B.36)}
$$

and the $o_p(n^{-1/2})$ term in (B.34) follows from Lemma B.3.1 with $a_{n_1,\ldots,n_m} = n^{-\frac{1}{2}}$. (B.36) also implies that the first term in (B.33) is $\frac{\lambda^*}{n_i}(\sigma_i^2 + o_p(1))$. Hence (B.33) can be written as

$$
\bar{Y}_i - \mu_i^* = \frac{\lambda^*}{n_i}\sigma_i^2 + o_p(n^{-\frac{1}{2}}). \qquad \text{(B.37)}
$$

Summing (B.37) over $i = 1, \ldots, m$ and using (B.22) give

$$
\sum_{i=1}^{m}\bar{Y}_i = \lambda^*\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i} + o_p(n^{-\frac{1}{2}}). \qquad \text{(B.38)}
$$

Therefore the expression for λ^* is

$$
\lambda^* = \frac{\sum_{i=1}^{m}\bar{Y}_i + o_p(n^{-\frac{1}{2}})}{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}}. \qquad \text{(B.39)}
$$

Step six: We substitute $\mu_i^* = o_p(1)$ and (B.39) into (B.20) to derive a formula for $w_{i,j}^*$, and from there we analyze the Taylor expansion of $-2\log R(0)$ to conclude the desired result. Each

$$-\log(n_i w_{i,j}^*) = \log(1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)) = \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*) - \frac{\lambda^{*2}}{2n_i^2}(Y_{i,j} - \mu_i^*)^2 + \eta_{i,j},$$

where $\eta_{i,j} = \frac{1}{3(1+\theta_{i,j}\frac{\lambda^*}{n_i}(Y_{i,j}-\mu_i^*))^3}\left(\frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)\right)^3$ for some $\theta_{i,j} \in (0,1)$, so the log profile likelihood ratio can be expressed as

$$\begin{aligned}
-2\log R(0) &= 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(1 + \frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*)) \\
&= 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\frac{\lambda^*}{n_i}(Y_{i,j} - \mu_i^*) - \frac{\lambda^{*2}}{2n_i^2}(Y_{i,j} - \mu_i^*)^2 + \eta_{i,j}\right) \\
&= 2\sum_{i=1}^{m}\lambda^*(\bar{Y}_i - \mu_i^*) - \sum_{i=1}^{m}\frac{\lambda^{*2}}{n_i}\cdot\frac{1}{n_i}\sum_{j=1}^{n_i}(Y_{i,j} - \mu_i^*)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}2\eta_{i,j} \qquad \text{(B.40)} \\
&= 2\lambda^*\sum_{i=1}^{m}\bar{Y}_i - \sum_{i=1}^{m}\frac{\lambda^{*2}}{n_i}(\sigma_i^2 + o_p(1)) + \sum_{i=1}^{m}\sum_{j=1}^{n_i}2\eta_{i,j} \qquad \text{(B.41)}
\end{aligned}$$

The equality between (B.40) and (B.41) follows from (B.22) and (B.36). To bound the last term

in (B.41) we write

$$
\begin{aligned}
& \left|\sum_{i,j} 2\eta_{ij}\right| \\
\leq\ & \frac{2}{3(1-\max_{i',j'}\left|\frac{\lambda^*}{n_{i'}}(Y_{i',j'}-\mu^*_{i'})\right|)^3}\left|\frac{\lambda^*}{\min_{i'} n_{i'}}\right|^3 \sum_{i,j}|Y_{i,j}-\mu^*_i|^3\cdot\mathbf{1}\left\{\max_{i',j'}\left|\frac{\lambda^*}{n_{i'}}(Y_{i',j'}-\mu^*_{i'})\right|<1\right\} \\
& +\left|\sum_{i,j}2\eta_{ij}\right|\cdot\mathbf{1}\left\{\max_{i',j'}\left|\frac{\lambda^*}{n_{i'}}(Y_{i',j'}-\mu^*_{i'})\right|\geq 1\right\} \\
=\ & \frac{2}{3(1-o_p(1))^3}O_p\big(n^{-\frac{3}{2}}\big)\sum_{i=1}^m 2n_iZ_N\sum_{j=1}^{n_i}\frac{1}{n_i}|Y_{i,j}-\mu^*_i|^2+o_p(1) \\
& \text{where the second } o_p(1) \text{ term follows from Lemma B.3.1 with } a_{n_1,\ldots,n_m}=1 \\
=\ & O_p\big(n^{-\frac{3}{2}}\big)\sum_{i=1}^m 2n_iZ_NO_p(1)+o_p(1) \\
=\ & O_p(n^{-\frac{3}{2}})no(n^{\frac{1}{2}})O_p(1)+o_p(1) \\
=\ & o_p(1).
\end{aligned}
$$

Hence using the above bound and (B.39), the log profile likelihood ratio (B.41) becomes

$$
\begin{aligned}
-2\log R(0) &= 2\lambda^*\sum_{i=1}^m\bar{Y}_i-\lambda^{*2}\sum_{i=1}^m\frac{\sigma_i^2}{n_i}+o_p(1) \\
&= \frac{\big(\sum_{i=1}^m\bar{Y}_i\big)^2}{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}+o_p(1). \qquad \text{(B.42)}
\end{aligned}
$$

To resolve the issue caused by the possible absence of Slater's condition (recall the note on Slater's condition at the end of Step two), note that the above result holds only in the event of $\mathcal{S}$, namely

$$
-2\log R(0)=\frac{\big(\sum_{i=1}^m\bar{Y}_i\big)^2}{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}+o_p(1)\ \text{ when } \mathbf{1}\{\mathcal{S}\}=1.
$$

Also note that by the definition of $\mathcal{S}$ the optimization problem in (B.18) is feasible when $\mathcal{S}$ occurs. Equivalently, $-2\log R(0)$ can take infinity values only when $\mathcal{S}$ does not occur. Thus Lemma B.3.1

is applicable to $-2\log R(0)\cdot\mathbf{1}\{\mathcal{S}^c\}$. Now, regardless of the occurrence of $\mathcal{S}$, we have

$$\begin{aligned}
-2\log R(0) &= \Big(\frac{\big(\sum_{i=1}^m \bar{Y}_i\big)^2}{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}} + o_p(1)\Big)\cdot\mathbf{1}\{\mathcal{S}\} - 2\log R(0)\cdot\mathbf{1}\{\mathcal{S}^c\} \\
&= \frac{\big(\sum_{i=1}^m \bar{Y}_i\big)^2}{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}} + o_p(1) - \Big(\frac{\big(\sum_{i=1}^m \bar{Y}_i\big)^2}{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}} + o_p(1)\Big)\cdot\mathbf{1}\{\mathcal{S}^c\} - 2\log R(0)\cdot\mathbf{1}\{\mathcal{S}^c\} \\
&= \frac{\big(\sum_{i=1}^m \bar{Y}_i\big)^2}{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}} + o_p(1). \qquad \text{(B.43)}
\end{aligned}$$

Here we apply Lemma B.3.1 with $a_{n_1,\ldots,n_m}=1$ to both $\big((\sum_{i=1}^m \bar{Y}_i)^2/\sum_{i=1}^m \frac{\sigma_i^2}{n_i}+o_p(1)\big)\cdot\mathbf{1}\{\mathcal{S}^c\}$ and $2\log R(0)\cdot\mathbf{1}\{\mathcal{S}^c\}$ to derive the last equality. This brings us back to (B.42).

By Slutsky's theorem, it remains to show that the leading term in (B.42) $\Rightarrow \mathcal{X}_1^2$. The leading term can be written as

$$\left[\sum_{i=1}^m\sum_{j=1}^{n_i}\frac{Y_{i,j}}{n_i\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}}\right]^2. \qquad \text{(B.44)}$$

By the continuous mapping theorem it suffices to show that the sum in (B.44) $\Rightarrow \mathcal{N}(0,1)$. We check the Lindeberg condition for the triangular array

$$(W_{N,1},\ldots,W_{N,N}) := \big(Y_{1,1},\ldots,Y_{1,n_1},\ldots,Y_{m,1},\ldots,Y_{m,n_m}\big)\Big/\Big(n_i\sqrt{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}\Big)$$

where $N=\sum_{i=1}^m n_i$. The independence and mean zero conditions are obviously met, and

$$\sum_{k=1}^N \mathbb{E}W_{N,k}^2 = \sum_{i=1}^m\sum_{j=1}^{n_i}\mathbb{E}\left[\frac{Y_{i,j}^2}{n_i^2\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}\right] = \sum_{i=1}^m\sum_{j=1}^{n_i}\frac{\sigma_i^2}{\sum_{i=1}^m n_i\sigma_i^2} = 1.$$

For any $\epsilon > 0$

$$\begin{aligned}
& \sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathbb{E}\left[\left(\frac{Y_{i,j}}{n_i\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}}\right)^2 \cdot \mathbf{1}\left\{\left|\frac{Y_{i,j}}{n_i\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}}\right| > \epsilon\right\}\right] \\
= & \sum_{i=1}^{m} n_i\mathbb{E}\left[\left(\frac{Y_{i,1}}{n_i\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}}\right)^2 \cdot \mathbf{1}\left\{\left|\frac{Y_{i,1}}{n_i\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}}\right| > \epsilon\right\}\right] \\
\leq & \sum_{i=1}^{m} C_1\mathbb{E}\left[Y_{i,1}^2 \cdot \mathbf{1}\left\{|Y_{i,1}| > \epsilon C_2\sqrt{n}\right\}\right] \text{ for some constants } C_1, C_2 \\
\to & \ 0 \ \text{ by the dominated convergence theorem.}
\end{aligned}$$

Therefore the Lindeberg condition holds for $W_{N,k}$. By the Lindeberg-Feller theorem (e.g., Theorem 3.4.5 in Durrett 2010), the sum in (B.44) $\Rightarrow \mathcal{N}(0,1)$ hence (B.44) itself $\Rightarrow \mathcal{X}_1^2$. □

B.4 Proofs of Results in Section 3.4.4

Proof of Theorem 3.4.5. From Theorem 3.4.3 we know $P(-2\log R(\mu_0) \leq \mathcal{X}^2_{1,1-\alpha}) \to 1-\alpha$ as $n \to \infty$. That is, the set $\{\mu \in \mathbb{R} | -2\log R(\mu) \leq \mathcal{X}^2_{1,1-\alpha}\}$ contains the true value μ_0 with probability $1-\alpha$ asymptotically. Note that this set can be identified as

$$\mathcal{V} = \left\{\left|\sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{i,j}w_{i,j}\right| - 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}) \leq \mathcal{X}^2_{1,1-\alpha},\ \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for all } i, w_{i,j} \geq 0 \text{ for all } i,j\right\}.$$

It is obvious that $\underline{\mu}/\overline{\mu} = \min / \max\{\mu : \mu \in \mathcal{V}\}$, and they are attained because the feasible set $\mathcal{U}_\alpha$ is compact and the objective is linear hence continuous in $w_{i,j}$'s. So if the set $\mathcal{V}$ is convex, then $\mathcal{V} = [\underline{\mu}, \overline{\mu}]$ which concludes the theorem. To show convexity, it is enough to notice that $\mathcal{U}_\alpha$ is convex, and the objective is linear in $w_{i,j}$. □

Proof of Proposition 3.4.6. We need the following corollary of Theorem 3.4.3:

Corollary B.4.1 *Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{i,j}/n_i$ be the sample mean of the i-th sample, $\sigma_i^2 = \mathrm{Var}(Y_i)$ be the true variance, and z be a fixed constant. Under the same conditions of Theorem 3.4.3, $-2\log R(\sum_{i=1}^m \bar{Y}_i + z\sqrt{\sum_{i=1}^m \sigma_i^2/n_i}) \to z^2$ in probability as $n \to \infty$.*

Proof of Corollary B.4.1. The proof closely follows that of Theorem 3.4.3, and we only point out how each step should be modified in order to prove this corollary. Assuming $\mathrm{Var}(Y_i) > 0$ and $\mathbb{E}Y_i = 0$ is still without loss of generality because, with $I = \{i : \mathrm{Var}(Y_i) > 0\}$,

$$
\begin{aligned}
& R\Big(\sum_{i=1}^{m} \bar{Y}_i + z\sqrt{\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}}\Big) \\
= \; & \max\Bigg\{ \prod_{i=1}^{m}\prod_{j=1}^{n_i} n_i w_{i,j} \Bigg| \sum_{i=1}^{m}\sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = \sum_{i=1}^{m} \bar{Y}_i + z\sqrt{\sum_{i=1}^{m} \frac{\sigma_i^2}{n_i}}, \\
& \qquad \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for all } i,\ w_{i,j} \geq 0 \text{ for all } i,j \Bigg\} \\
= \; & \max\Bigg\{ \prod_{i=1}^{m}\prod_{j=1}^{n_i} n_i w_{i,j} \Bigg| \sum_{i\in I}\sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = \sum_{i\in I} \bar{Y}_i + z\sqrt{\sum_{i\in I} \frac{\sigma_i^2}{n_i}}, \\
& \qquad \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for all } i,\ w_{i,j} \geq 0 \text{ for all } i,j \Bigg\} \\
= \; & \max\Bigg\{ \prod_{i\in I}\prod_{j=1}^{n_i} n_i w_{i,j} \Bigg| \sum_{i\in I}\sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = \sum_{i\in I} \bar{Y}_i + z\sqrt{\sum_{i\in I} \frac{\sigma_i^2}{n_i}}, \\
& \qquad \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for } i \in I,\ w_{i,j} \geq 0 \text{ for all } i \in I, j \Bigg\} \\
= \; & \max\Bigg\{ \prod_{i\in I}\prod_{j=1}^{n_i} n_i w_{i,j} \Bigg| \sum_{i\in I}\sum_{j=1}^{n_i} (Y_{i,j} - \mathbb{E}Y_i) w_{i,j} = \sum_{i\in I} (\bar{Y}_i - \mathbb{E}Y_i) + z\sqrt{\sum_{i\in I} \frac{\sigma_i^2}{n_i}}, \\
& \qquad \sum_{j=1}^{n_i} w_{i,j} = 1 \text{ for } i \in I,\ w_{i,j} \geq 0 \text{ for all } i \in I, j \Bigg\}
\end{aligned}
$$

and the limit distribution, i.e., the point mass at z^2, does not depend on the number of distributions

m. Next we consider the following counterpart of (B.18)

$$\begin{aligned} \min_{\mathbf{w}_1,\ldots,\mathbf{w}_m,\boldsymbol{\mu}} \quad & -\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}) \\ \text{subject to} \quad & \sum_{j=1}^{n_i} Y_{i,j} w_{i,j} = \mu_i,\ i=1,\ldots,m \\ & \sum_{j=1}^{n_i} w_{i,j} = 1,\ i=1,\ldots,m \\ & \sum_{i=1}^{m}\mu_i = \sum_{i=1}^{m}\bar{Y}_i + z\sqrt{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}}. \end{aligned} \tag{B.45}$$

Step one: We show Slater's condition holds for (B.45) with a probability tending to one. Instead of $\tilde{\mathcal{S}}_i$, consider the event indexed by i

$$\min_{j=1,\ldots,n_i} Y_{i,j} < \bar{Y}_i + \frac{z}{m}\sqrt{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}} < \max_{j=1,\ldots,n_i} Y_{i,j}. \tag{B.46}$$

We need to show the probability that (B.46) happens goes to one. Note that $\bar{Y}_i + \frac{z}{m}\cdot\sqrt{\sum_{i=1}^{m}\sigma_i^2/n_i} = o_p(1)$, and for a small enough $\epsilon > 0$ it holds $P(Y_i \geq -\epsilon) < 1, P(Y_i \leq \epsilon) < 1$. Hence

$$\begin{aligned} P\big(\min_{j=1,\ldots,n_i} Y_{i,j} \geq \bar{Y}_i + \frac{z}{m}\sqrt{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}}\big) &\leq P\big(\min_{j=1,\ldots,n_i} Y_{i,j} \geq -\epsilon\big) + P\big(\bar{Y}_i + \frac{z}{m}\sqrt{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}} < -\epsilon\big) \\ &= (P(Y_i \geq -\epsilon))^{n_i} + P(o_p(1) < -\epsilon) \to 0. \end{aligned}$$

This justifies the first inequality of (B.46), and the second inequality can be treated in the same way. Applying the union bound shows that the probability of (B.46) approaches one. The rest of this step remains the same.

Step two: The only change is that one of the KKT conditions, (B.22), is replaced by

$$\sum_{i=1}^{m}\mu_i^* = \sum_{i=1}^{m}\bar{Y}_i + z\sqrt{\sum_{i=1}^{m}\frac{\sigma_i^2}{n_i}}.$$

Step three: (B.29) is replaced by

$$-\left(1+\frac{|\lambda^*|}{n}o(n^{\frac{1}{2}})\right)\operatorname{sign}(\lambda^*)\cdot z\sqrt{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}\geq\frac{|\lambda^*|}{n\cdot\bar{c}/\underline{c}}\left(\sum_{i=1}^m\hat{\sigma}_i^2+o_p(1)\right)$$

and (B.30) becomes

$$\frac{|\lambda^*|}{n}\left(\frac{\underline{c}}{\bar{c}}\sum_{i=1}^m\hat{\sigma}_i^2+o_p(1)+o(n^{\frac{1}{2}})\cdot z\sqrt{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}\right)\leq|z|\sqrt{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}.$$

The final bound $\lambda^*=O_p(n^{1/2})$ still holds by observing that $z\sqrt{\sum_{i=1}^m\sigma_i^2/n_i}=O(n^{-1/2})$ just like $\sum_{i=1}^m\bar{Y}_i$.

Step four: No changes needed.

Step five: (B.38) needs to be replaced by

$$-z\sqrt{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}=\lambda^*\sum_{i=1}^m\frac{\sigma_i^2}{n_i}+o_p(n^{-\frac{1}{2}}).$$

Hence (B.39) becomes

$$\lambda^*=\frac{-z+o_p(1)}{\sqrt{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}}.$$

Step six: (B.41) and (B.42) are replaced by

$$\begin{aligned}-2\log R\Big(\sum_{i=1}^m\bar{Y}_i+z\sqrt{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}\Big)&=-2\lambda^*z\sqrt{\sum_{i=1}^m\frac{\sigma_i^2}{n_i}}-\sum_{i=1}^m\frac{\lambda^{*2}}{n_i}(\sigma_i^2+o_p(1))+\sum_{i=1}^m\sum_{j=1}^{n_i}2\eta_{i,j}\\&=z^2+o_p(1)\end{aligned}$$

and the desired conclusion follows. □

Now we are ready to prove Proposition 3.4.6. Recall the definition of profile likelihood ratio

$R(\mu)$ in (3.18). Since $z^2_{1-\alpha/2} = \mathcal{X}^2_{1,1-\alpha}$, Corollary B.4.1 entails that for any fixed small $\epsilon > 0$

$$P\Big(-2\log R\Big(\sum_{i=1}^m \bar{Y}_i - (z_{1-\alpha/2} - \epsilon)\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}\Big) < \mathcal{X}^2_{1,1-\alpha}\Big) \to 1, \tag{B.47}$$

$$P\Big(-2\log R\Big(\sum_{i=1}^m \bar{Y}_i - (z_{1-\alpha/2} + \epsilon)\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}\Big) > \mathcal{X}^2_{1,1-\alpha}\Big) \to 1. \tag{B.48}$$

In the proof of Theorem 3.4.5 it is shown that $\{\mu \in \mathbb{R} | -2\log R(\mu) \leq \mathcal{X}^2_{1,1-\alpha}\} = [\underline{\mu}, \overline{\mu}]$. Therefore conditioned on the event in (B.47) we must have $\sum_{i=1}^m \bar{Y}_i - (z_{1-\alpha/2} - \epsilon)\sqrt{\sum_{i=1}^m \sigma_i^2/n_i} \in [\underline{\mu}, \overline{\mu}]$. Conditioned on the event in (B.48) we have $\sum_{i=1}^m \bar{Y}_i - (z_{1-\alpha/2} + \epsilon)\sqrt{\sum_{i=1}^m \sigma_i^2/n_i} \notin [\underline{\mu}, \overline{\mu}]$. Moreover, since the sum of sample means $\sum_{i=1}^m \bar{Y}_i \in [\underline{\mu}, \overline{\mu}]$ almost surely and $\sum_{i=1}^m \bar{Y}_i - (z_{1-\alpha/2} + \epsilon)\sqrt{\sum_{i=1}^m \sigma_i^2/n_i} < \sum_{i=1}^m \bar{Y}_i$, it must be the case that $\sum_{i=1}^m \bar{Y}_i - (z_{1-\alpha/2} + \epsilon)\sqrt{\sum_{i=1}^m \sigma_i^2/n_i} < \underline{\mu}$. Applying the union bound we get

$$P\Big(\Big|\underline{\mu} - \Big(\sum_{i=1}^m \bar{Y}_i - z_{1-\alpha/2}\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}\Big)\Big| \leq \epsilon\sqrt{\sum_{i=1}^m \frac{\sigma_i^2}{n_i}}\Big) \to 1.$$

Sending ϵ to 0 gives the desired conclusion for $\underline{\mu}$. The proof for $\overline{\mu}$ is similar. □

Proof of Corollary 3.4.7. If we can show that $\mathrm{Var}(G_i(X_i)) < \infty$ for all $i = 1, \ldots, m$, then this is a direct consequence of Theorem 3.4.5 and Proposition 3.4.6 with $Y_{i,j} = \frac{Z^*}{m} + G_i(X_{i,j})$ and the fact that $\mathbb{E}[G_i(X_i)] = 0$. Since Assumption 3.3.3 implies $\mathbb{E}_{P_1,\ldots,P_m}[h^2(\mathbf{X}_1, \ldots, \mathbf{X}_m)] < \infty$, by Jensen's inequality every conditional expectation of $h(\mathbf{X}_1, \ldots, \mathbf{X}_m)$ also has a finite second moment. Note that $G_i(X_i)$ is the sum of T_i conditional expectations of $h(\mathbf{X}_1, \ldots, \mathbf{X}_m)$. Therefore it has a finite second moment, hence a finite variance, by the Minkowski inequality. □

Proof of Theorem 3.4.8. We have

$$\begin{aligned}
\mathscr{L} &= \inf_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) \\
&= \inf_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big[Z_L(\mathbf{w}_1, \ldots, \mathbf{w}_m) + \big(Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - Z_L(\mathbf{w}_1, \ldots, \mathbf{w}_m)\big)\big] \\
&\geq \inf_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} Z_L(\mathbf{w}_1, \ldots, \mathbf{w}_m) + \inf_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big(Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - Z_L(\mathbf{w}_1, \ldots, \mathbf{w}_m)\big) \\
&\geq \mathscr{L}_L - \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - Z_L(\mathbf{w}_1, \ldots, \mathbf{w}_m)\big|.
\end{aligned}$$

Similarly it can be shown that $\mathscr{L}_L \geq \mathscr{L} - \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|Z(\mathbf{w}_1,\ldots,\mathbf{w}_m) - Z_L(\mathbf{w}_1,\ldots,\mathbf{w}_m)\big|$. Therefore

$$|\mathscr{L} - \mathscr{L}_L| \leq \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|Z(\mathbf{w}_1,\ldots,\mathbf{w}_m) - Z_L(\mathbf{w}_1,\ldots,\mathbf{w}_m)\big|. \tag{B.49}$$

By the bound (3.16) in Proposition 3.4.2, $\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|Z(\mathbf{w}_1,\ldots,\mathbf{w}_m) - Z_L(\mathbf{w}_1,\ldots,\mathbf{w}_m)\big| = O_p(1/n)$ hence $|\mathscr{L} - \mathscr{L}_L| = O_p(1/n) = o_p(1/\sqrt{n})$. Analogously $|\mathscr{U} - \mathscr{U}_L| = o_p(1/\sqrt{n})$. In particular, the representation (3.22) holds for $\mathscr{L}, \mathscr{U}$ as well, i.e.

$$\begin{aligned} \mathscr{L} &= Z^* + \sum_{i=1}^m \bar{G}_i - z_{1-\alpha/2}\sigma_I + o_p\big(\frac{1}{\sqrt{n}}\big) \\ \mathscr{U} &= Z^* + \sum_{i=1}^m \bar{G}_i + z_{1-\alpha/2}\sigma_I + o_p\big(\frac{1}{\sqrt{n}}\big). \end{aligned} \tag{B.50}$$

Now we show that (B.50) guarantees the asymptotic exactness of $[\mathscr{L}, \mathscr{U}]$ as a CI for Z^*. For convenience, assume $\mathrm{Var}(G_i(X_i)) > 0$ for all i without loss of generality. The standard central limit theorem entails that $\frac{\bar{G}_i}{\sqrt{\mathrm{Var}(G_i(X_i))/n_i}} \Rightarrow \mathcal{N}(0,1)$. Since the data across different input models are independent, we have the joint convergence

$$\Big(\frac{\bar{G}_1}{\sqrt{\mathrm{Var}(G_1(X_1))/n_1}}, \ldots, \frac{\bar{G}_m}{\sqrt{\mathrm{Var}(G_m(X_m))/n_m}}\Big) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_m),$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix. To proceed, we need the following result:

Lemma B.4.2 (Uniform convergence of measures, Theorem 4.2 in Rao 1962) *Let μ^* and $\{\mu_n\}_{n=1}^\infty$ be probability measures on $\mathbb{R}^d$. If μ^* is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, then $\mu_n \Rightarrow \mu^*$ if and only if*

$$\lim_{n\to\infty} \sup_{C\in\mathscr{C}} |\mu_n(C) - \mu^*(C)| = 0,$$

where $\mathscr{C}$ denotes the set of all measurable convex sets.

Let $(W_1,\ldots,W_m)$ be an m dimensional standard normal vector, then $\sum_{i=1}^m \frac{1}{\sigma_I}\sqrt{\mathrm{Var}(G_i(X_i))/n_i} W_i$

follows $\mathcal{N}(0,1)$. Hence

$$\begin{aligned}
&\left|P\Big(\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \le z\Big) - \Phi(z)\right| \\
= &\left|P\Big(\sum_{i=1}^m \frac{\sqrt{\mathrm{Var}(G_i(X_i))/n_i}}{\sigma_I} \cdot \frac{\bar{G}_i}{\sqrt{\mathrm{Var}(G_i(X_i))/n_i}} \le z\Big) - P\Big(\sum_{i=1}^m \frac{\sqrt{\mathrm{Var}(G_i(X_i))/n_i}}{\sigma_I} W_i \le z\Big)\right| \\
= &\left|P\Big(\Big(\frac{\bar{G}_1}{\sqrt{\mathrm{Var}(G_1(X_1))/n_1}}, \ldots, \frac{\bar{G}_m}{\sqrt{\mathrm{Var}(G_m(X_m))/n_m}}\Big) \in \tilde{C}\Big) - P\Big((W_1, \ldots, W_m) \in \tilde{C}\Big)\right| \\
&\text{where } \tilde{C} = \Big\{(x_1, \ldots, x_m) \in \mathbb{R}^m \Big| \sum_{i=1}^m \frac{\sqrt{\mathrm{Var}(G_i(X_i))/n_i}}{\sigma_I} x_i \le z\Big\}. \qquad \text{(B.51)}
\end{aligned}$$

Since the set $\tilde{C}$ is a half-space and in particular a convex set, Lemma B.4.2 implies

$$\begin{aligned}
&\left|P\Big(\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \le z\Big) - \Phi(z)\right| \\
\le &\sup_{C \in \mathscr{C}} \left|P\Big(\Big(\frac{\bar{G}_1}{\sqrt{\mathrm{Var}(G_1(X_1))/n_1}}, \ldots, \frac{\bar{G}_m}{\sqrt{\mathrm{Var}(G_m(X_m))/n_m}}\Big) \in C\Big) - P\Big((W_1, \ldots, W_m) \in C\Big)\right| \to 0.
\end{aligned}$$

Therefore

$$\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \Rightarrow \mathcal{N}(0,1). \qquad \text{(B.52)}$$

Now (B.50) forces

$$\begin{aligned}
P(\mathscr{L} \le Z^*) &= P(\sum_{i=1}^m \bar{G}_i + o_p\big(\frac{1}{\sqrt{n}}\big) \le z_{1-\alpha/2}\sigma_I) \\
&= P\Big(\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} + o_p(1) \le z_{1-\alpha/2}\Big) \\
&\to P(\mathcal{N}(0,1) \le z_{1-\alpha/2}) = 1 - \frac{\alpha}{2} \text{ by Slutsky's theorem.}
\end{aligned}$$

Similarly we have $P(\mathscr{U} \ge Z^*) \to 1 - \alpha/2$. Moreover, $\mathscr{U} - \mathscr{L} = 2z_{1-\alpha/2}\sigma_I + o_p(1/\sqrt{n})$ hence

$$P(\mathscr{U} < Z^* < \mathscr{L}) \le P(\mathscr{U} < \mathscr{L}) = P(2z_{1-\alpha/2}\sigma_I + o_p(1/\sqrt{n}) < 0) = P(2z_{1-\alpha/2} < o_p(1)) \to 0.$$

Combining the limit probabilities gives

$$\begin{aligned} P(\mathscr{L} \le Z^* \le \mathscr{U}) &= P(\mathscr{L} \le Z^*) + P(\mathscr{U} \ge Z^*) - P(\mathscr{L} \le Z^* \text{ or } \mathscr{U} \ge Z^*) \\ &= P(\mathscr{L} \le Z^*) + P(\mathscr{U} \ge Z^*) - 1 + P(\mathscr{U} < Z^* < \mathscr{L}) \\ &\to 1 - \frac{\alpha}{2} + 1 - \frac{\alpha}{2} - 1 + 0 = 1 - \alpha. \end{aligned}$$

This completes the proof. □

B.5 Proofs of Results in Section 3.4.5

Proof of Proposition 3.4.10. It suffices to show the first part

$$\mathbb{E}_{D,\xi_1}\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|\widehat{Z_L}(\mathbf{w}_1,\ldots,\mathbf{w}_m) - \widehat{\widehat{Z_L}}(\mathbf{w}_1,\ldots,\mathbf{w}_m)\big|^2\big] = O\big(\frac{1}{R_1}\big),$$

because the second part then follows from (3.17) and the simple inequality $\sup|Z - \widehat{\widehat{Z_L}}| \le \sup|Z - \widehat{Z_L}| + \sup|\widehat{Z_L} - \widehat{\widehat{Z_L}}|$. First we present two lemmas.

Lemma B.5.1 *Under Assumptions 3.3.1 and 3.3.3, as $n \to \infty$ for $k = 1, 2, 3, 4$ we have*

$$\mathbb{E}_D\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \Big|\mathbb{E}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}[h^k(\mathbf{X}_1,\ldots,\mathbf{X}_m)] - \mathbb{E}_{P_1,\ldots,P_m}[h^k(\mathbf{X}_1,\ldots,\mathbf{X}_m)]\Big|^2\big] = O\big(\frac{1}{n}\big). \quad \text{(B.53)}$$

In particular for $k = 1, 2, 3, 4$ it holds

$$\mathbb{E}_D\big[\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \Big|\mathbb{E}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}[h^k(\mathbf{X}_1,\ldots,\mathbf{X}_m)]\Big|^2\big] = O(1). \quad \text{(B.54)}$$

Proof of Lemma B.5.1. (B.53) is argued using the proof of Proposition 3.4.2. Note that the proof for Proposition 3.4.2 goes through as long as the maximum second moment $\mathcal{M}$ defined in (B.6) is finite, a weaker condition than Assumption 3.3.3. In particular, Assumption 3.3.3 remains valid if the target performance measure is changed to $\mathbb{E}_{P_1,\ldots,P_m}[h^k(\mathbf{X}_1,\ldots,\mathbf{X}_m)]$ for $k = 2, 3, 4$, except that the maximum second moment $\mathcal{M}$ has to be replaced by the 4-th, 6-th and 8-th moments

respectively. Below we will argue for the case $k = 1$ only, and the cases $k = 2, 3, 4$ follow from the same reasoning. Recall the expansion (B.5). The term with $d = 0$ is simply Z^*. The argument leading to the bound (B.16) works for all $d \geq 1$, and hence (B.16) is valid for all $d \geq 1$. The leading remainders with $d = 1$ then give rise to the order $O(1/n)$ in (B.53), as opposed to $d = 2$ giving the order $O(1/n^2)$ in (3.16).

To prove (B.54), use the inequality

$$\begin{aligned} & \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \Big|\mathbb{E}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}[h^k(\mathbf{X}_1,\ldots,\mathbf{X}_m)]\Big| \\ \leq & \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \Big|\mathbb{E}_{\mathbf{w}_1,\ldots,\mathbf{w}_m}[h^k(\mathbf{X}_1,\ldots,\mathbf{X}_m)] - \mathbb{E}_{P_1,\ldots,P_m}[h^k(\mathbf{X}_1,\ldots,\mathbf{X}_m)]\Big| \\ & \qquad + \mathbb{E}_{P_1,\ldots,P_m}[h^k(\mathbf{X}_1,\ldots,\mathbf{X}_m)] \end{aligned}$$

and the Minkowski inequality. □

Lemma B.5.2 *Under Assumptions 3.3.1 and 3.3.3, as the input data size $n \to \infty$, the gradient estimator $\hat{\hat{G}}_i(X_{i,j})$ in (3.2) satisfies*

$$\mathbb{E}_{D,\xi_1}\big[\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\hat{\hat{G}}_i(X_{i,j}) - \hat{G}_i(X_{i,j}))^2\big] = O\big(\frac{1}{R_1}\big).$$

Proof of Lemma B.5.2. We first note that due to the symmetry between the i.i.d. data

$$\mathbb{E}_{D,\xi_1}\big[\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\hat{\hat{G}}_i(X_{i,j}) - \hat{G}_i(X_{i,j}))^2\big] = \sum_{i=1}^{m}\frac{1}{n_i}\mathbb{E}_{D,\xi_1}\big[(\hat{\hat{G}}_i(X_{i,1}) - \hat{G}_i(X_{i,1}))^2\big],$$

and therefore it suffices to bound each $\mathbb{E}\big[(\hat{\hat{G}}_i(X_{i,1}) - \hat{G}_i(X_{i,1}))^2\big]$. Since $\hat{\hat{G}}_i(X_{i,1})$ differs from the unbiased sample covariance by only a factor of $\frac{R_1-1}{R_1}$, its bias (conditioned on the input data) can be easily identified as $\hat{G}_i(X_{i,1})/R_1$. By the variance formula for the unbiased sample covariance,

and suppressing the arguments in h for notational simplicity, we have

$$\begin{aligned}\mathrm{Var}_{\xi_1|D}\big(\hat{\hat{G}}_i(X_{i,1})\big) &= \frac{(R_1-1)^2}{R_1^3}\Big(\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m)-Z(\hat{P}_1,\ldots,\hat{P}_m))^2(S_{i,1}(\mathbf{X}_i))^2\big]+\\ &\qquad \frac{1}{R_1-1}\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(S_{i,1}(\mathbf{X}_i))-\frac{R_1-2}{R_1-1}(\hat{G}_i(X_{i,1}))^2\Big).\end{aligned}$$

Hence the mean squared error

$$\begin{aligned}&\mathbb{E}_{\xi_1|D}[(\hat{\hat{G}}_i(X_{i,1})-\hat{G}_i(X_{i,1}))^2]\\ =\;& \mathrm{Var}_{\xi_1|D}\big(\hat{\hat{G}}_i(X_{i,j})\big)+\big(\frac{\hat{G}_i(X_{i,1})}{R_1}\big)^2\\ \leq\;& \frac{1}{R_1}\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m)-Z(\hat{P}_1,\ldots,\hat{P}_m))^2(S_{i,1}(\mathbf{X}_i))^2\big]+\\ &\quad\frac{1}{R_1^2}\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(S_{i,1}(\mathbf{X}_i)) \qquad \text{(B.55)}\\ \leq\;& \frac{1}{R_1}\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m)-Z(\hat{P}_1,\ldots,\hat{P}_m))^2(S_{i,1}(\mathbf{X}_i))^2\big]+\frac{n_iT_i}{R_1^2}\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) \qquad \text{(B.56)}\end{aligned}$$

To tackle the first term in (B.56)

$$
\begin{aligned}
& \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m) - Z(\hat{P}_1,\ldots,\hat{P}_m))^2 (S_{i,1}(\mathbf{X}_i))^2\big] \\
= \; & \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\Bigg[(h - Z(\hat{P}_1,\ldots,\hat{P}_m))^2 \Bigg(T_i^2 + n_i^2 \left(\sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,1}\}\right)^2 \\
& \quad -2T_i n_i \sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,1}\}\Bigg)\Bigg] \\
\leq \; & T_i^2 \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\left[(h - Z(\hat{P}_1,\ldots,\hat{P}_m))^2 n_i^2 \left(\sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,1}\}\right)^2\right] \\
\leq \; & T_i^2 \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\left[2h^2 n_i^2 \left(\sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,1}\}\right)^2\right] + \\
& \quad \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\left[2(Z(\hat{P}_1,\ldots,\hat{P}_m))^2 n_i^2 \left(\sum_{t=1}^{T_i} \mathbf{1}\{X_i(t) = X_{i,1}\}\right)^2\right] \\
= \; & T_i^2 \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + 2n_i^2 \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\left[h^2 \left(\sum_{s,t=1}^{T_i} \mathbf{1}\{X_i(t) = X_i(s) = X_{i,1}\}\right)\right] + \\
& \quad 2(T_i n_i + T_i(T_i - 1))(Z(\hat{P}_1,\ldots,\hat{P}_m))^2 \\
\leq \; & T_i^2 \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + 2n_i^2 \sum_{s,t=1}^{T_i} \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2 \cdot \mathbf{1}\{X_i(t) = X_i(s) = X_{i,1}\}] \\
& \quad +2(T_i n_i + T_i^2)(Z(\hat{P}_1,\ldots,\hat{P}_m))^2 \\
= \; & T_i^2 \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + 2(T_i n_i + T_i^2)(Z(\hat{P}_1,\ldots,\hat{P}_m))^2 + 2n_i \sum_{t=1}^{T_i} \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2 | X_i(t) = X_{i,1}] + \\
& \quad 2\sum_{s \neq t} \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2 | X_i(t) = X_i(s) = X_{i,1}].
\end{aligned}
$$

Like in Assumption 3.3.3, denoting by $J_i = (J_i(1),\ldots,J_i(T_i)) \in \{1,2,\ldots,n_i\}^{T_i}$ the second subscripts of the input data points we can rewrite each conditional expectation $\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2 | X_i(t) = X_{i,1}]$ as

$$
\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2 | X_i(t) = X_{i,1}] = \frac{1}{n_i^{T_i - 1} \prod_{i' \neq i} n_{i'}^{T_{i'}}} \sum_{J_1,\ldots,J_m \text{ such that } J_i(t)=1} h^2(\mathbf{X}_{1,J_1},\ldots,\mathbf{X}_{m,J_m}).
$$

Therefore under Assumption 3.3.3 we have

$$\mathbb{E}_D\big[\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2|X_i(t)=X_{i,1}]\big]\le\mathcal{M}$$

where $\mathcal{M}$ is the maximum second moment defined in (B.6). The same argument also gives that $\mathbb{E}_D\big[\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2|X_i(t)=X_i(s)=X_{i,1}]\big]\le\mathcal{M}$. Also note that $\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)\le\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2]$ and $(Z(\hat{P}_1,\ldots,\hat{P}_m))^2\le\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[h^2]$ by Jensen's inequality. Hence by (B.54) with $k=2$ from Lemma (B.5.1) it holds that $\mathbb{E}_D\big[\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)\big]=O(1)$ and $\mathbb{E}_D\big[(Z(\hat{P}_1,\ldots,\hat{P}_m))^2\big]=O(1)$.

Now we take expectation of (B.56) with respect to the input data and use the upper bounds derived above to get

$$\begin{aligned}
&\mathbb{E}_{D,\xi_1}\big[(\hat{\hat{G}}_i(X_{i,1})-\hat{G}_i(X_{i,1}))^2\big]\\
=\;&\mathbb{E}_D\Big[\mathbb{E}_{\xi_1|D}\big[(\hat{\hat{G}}_i(X_{i,1})-\hat{G}_i(X_{i,1}))^2\big]\Big]\\
=\;&\frac{1}{R_1}\mathbb{E}_D\big[\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m)-Z(\hat{P}_1,\ldots,\hat{P}_m))^2(S_{i,1}(\mathbf{X}_i))^2\big]\big]+\frac{n_iT_i}{R_1^2}O(1)\\
=\;&\frac{1}{R_1}(T_i^2O(1)+(T_in_i+T_i^2)O(1)+O(n_iT_i)+O(T_i^2))+O\big(\frac{n_iT_i}{R_1^2}\big)\\
=\;&O\big(\frac{T_i^2}{R_1}+\frac{n_iT_i}{R_1}+\frac{n_iT_i}{R_1^2}\big)\\
=\;&O\big(\frac{n_i}{R_1}\big)\;\text{ since each } T_i \text{ is treated as constant.}
\end{aligned}$$

Dividing each side by n_i and summing up over $i=1,\ldots,m$ gives the bound $O(1/R_1)$. □

Now we can prove Proposition 3.4.10. We bound the maximum deviation as follows

$$\begin{aligned}
&\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}\Big|\widehat{Z_L}(\mathbf{w}_1,\ldots,\mathbf{w}_m)-\widehat{\widehat{Z}}_L(\mathbf{w}_1,\ldots,\mathbf{w}_m)\Big|\\
\le\;&\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}\left|\sum_{i=1}^m\sum_{j=1}^{n_i}(\hat{\hat{G}}_i(X_{i,j})-\hat{G}_i(X_{i,j}))w_{i,j}\right|+\Big|Z(\hat{P}_1,\ldots,\hat{P}_m)-\hat{Z}(\hat{P}_1,\ldots,\hat{P}_m)\Big| \qquad \text{(B.57)}
\end{aligned}$$

On one hand, using conditioning and the moment bound (B.54) with $k=2$ from Lemma B.5.1, we

bound the second moment of the second term in (B.57) as

$$\begin{aligned}
& \mathbb{E}_{D,\xi_1}\big[\big|Z(\hat{P}_1,\ldots,\hat{P}_m)-\hat{Z}(\hat{P}_1,\ldots,\hat{P}_m)\big|^2\big] \\
= \;& \mathbb{E}_D\big[\mathbb{E}_{\xi_1|D}\big[\big|Z(\hat{P}_1,\ldots,\hat{P}_m)-\hat{Z}(\hat{P}_1,\ldots,\hat{P}_m)\big|^2\big]\big] \\
= \;& \frac{1}{R_1}\mathbb{E}_D\big[\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)\big] \\
= \;& O\big(\frac{1}{R_1}\big) \;\text{ by (B.54) in Lemma B.5.1.}
\end{aligned}$$

On the other hand, letting $Q_i^1=\hat{P}_i$ in Proposition 3.4.1 reveals that $\sum_{j=1}^{n_i}\hat{G}_i(X_{i,j})=0$ for all i. Note that the estimator (3.2) also has this property, i.e. $\sum_{j=1}^{n_i}\hat{\hat{G}}_i(X_{i,j})=0$ for all i. Hence the first term in (B.57) can be bounded as

$$\begin{aligned}
& \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}\left|\sum_{i=1}^m\sum_{j=1}^{n_i}(\hat{\hat{G}}_i(X_{i,j})-\hat{G}_i(X_{i,j}))w_{i,j}\right| \\
= \;& \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}\left|\sum_{i=1}^m\sum_{j=1}^{n_i}(\hat{\hat{G}}_i(X_{i,j})-\hat{G}_i(X_{i,j}))(w_{i,j}-\frac{1}{n_i})\right| \;\text{ by } \sum_{j=1}^{n_i}\hat{G}_i(X_{i,j})=\sum_{j=1}^{n_i}\hat{\hat{G}}_i(X_{i,j})=0 \\
= \;& \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}\left|\sum_{i=1}^m\sum_{j=1}^{n_i}\frac{1}{n_i}(\hat{\hat{G}}_i(X_{i,j})-\hat{G}_i(X_{i,j}))\cdot n_i(w_{i,j}-\frac{1}{n_i})\right| \\
\le \;& \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}\sqrt{\sum_{i=1}^m\sum_{j=1}^{n_i}\frac{1}{n_i^2}(\hat{\hat{G}}_i(X_{i,j})-\hat{G}_i(X_{i,j}))^2\sum_{i=1}^m\sum_{j=1}^{n_i}n_i^2(w_{i,j}-\frac{1}{n_i})^2} \\
\le \;& \sqrt{u(\alpha)^2\mathcal{X}^2_{1,1-\alpha}\sum_{i=1}^m\sum_{j=1}^{n_i}\frac{1}{n_i^2}(\hat{\hat{G}}_i(X_{i,j})-\hat{G}_i(X_{i,j}))^2} \;\text{ by Lemma B.2.2.}
\end{aligned}$$

After combining the above bounds, the desired conclusion follows from an application of the Minkowski inequality to (B.57) and using Lemma B.5.2. □

Proof of Theorem 3.4.11. In the proof of Theorem 3.4.8, if we replace the linear approximation Z_L by $\widehat{\widehat{Z}}_L$ then by the exactly same argument we have the following counterpart of (B.49) where on one hand

$$\big|\mathscr{L}-\widehat{\widehat{Z}}_L(\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min})\big|\le\sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha}\big|Z(\mathbf{w}_1,\ldots,\mathbf{w}_m)-\widehat{\widehat{Z}}_L(\mathbf{w}_1,\ldots,\mathbf{w}_m)\big|. \tag{B.58}$$

On the other hand the following bound trivially holds

$$|Z^{\min} - \widehat{\widehat{Z_L}}(\mathbf{w}_1^{\min}, \ldots, \mathbf{w}_m^{\min})| \leq \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - \widehat{\widehat{Z_L}}(\mathbf{w}_1, \ldots, \mathbf{w}_m)\big|.$$

Therefore

$$\big|\mathscr{L} - Z^{\min}\big| \leq 2 \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - \widehat{\widehat{Z_L}}(\mathbf{w}_1, \ldots, \mathbf{w}_m)\big|.$$

The desired conclusion for $Z^{\min}$ then immediately follows from the maximum deviation result (3.24) in Proposition 3.4.10. The conclusion for $Z^{\max}$ can be established similarly. □

The following result presents an alternate CI constructed directly from a linear approximation that is discussed at the end of Section 3.4.5.

Theorem B.5.3 *Suppose Assumptions 3.3.1, 3.3.2 and 3.3.3 hold. Replace the outputs in Step 3 of Algorithm 4 by*

$$L = \hat{Z}(\hat{P}_1, \ldots, \hat{P}_m) + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \hat{\hat{G}}_i(X_{i,j}) w_{i,j}^{\min},\ U = \hat{Z}(\hat{P}_1, \ldots, \hat{P}_m) + \sum_{i=1}^{m}\sum_{j=1}^{n_i} \hat{\hat{G}}_i(X_{i,j}) w_{i,j}^{\max},$$

where $\hat{Z}(\hat{P}_1, \ldots, \hat{P}_m)$ is the same sample mean from Step 1. Then as $n \to \infty$ and $R_1 \to \infty$

$$\mathbb{E}_{D,\xi_1}[(L - \mathscr{L})^2] = O\big(\frac{1}{n^2} + \frac{1}{R_1}\big),\ \mathbb{E}_{D,\xi_1}[(U - \mathscr{U})^2] = O\big(\frac{1}{n^2} + \frac{1}{R_1}\big)$$

where $\mathscr{L}, \mathscr{U}$ are the ideal confidence bounds defined in (3.9)*. Moreover, if R_1 satisfies $\frac{R_1}{n} \to \infty$ then*

$$\lim_{n\to\infty, \frac{R_1}{n}\to\infty} P(L \leq Z^* \leq U) = 1 - \alpha.$$

Proof of Theorem B.5.3. The bound (B.58) derived in the proof of Theorem 3.4.11 is exactly $|L - \mathscr{L}| \leq \sup_{(\mathbf{w}_1,\ldots,\mathbf{w}_m)\in\mathcal{U}_\alpha} \big|Z(\mathbf{w}_1, \ldots, \mathbf{w}_m) - \widehat{\widehat{Z_L}}(\mathbf{w}_1, \ldots, \mathbf{w}_m)\big|$. A direct application of result (3.24) from Proposition 3.4.10 then gives $\mathbb{E}_{D,\xi_1}[(L - \mathscr{L})^2] = O(1/n^2 + 1/R_1)$. The error bound of U with respect to $\mathscr{U}$ can be obtained similarly. To establish the asymptotic exactness of $[L, U]$ when R_1 grows at a faster rate than n, note that when $R_1/n \to \infty$ we have $1/R_1 = o(1/n)$ hence

$L - \mathscr{L} = o_p(1/\sqrt{n})$ and $U - \mathscr{U} = o_p(1/\sqrt{n})$. In this case the representation (B.50) holds for L, U as well. The rest of the proof is the same as that of Theorem 3.4.8. □

B.6 Proofs of Results in Section 3.4.6

Proof of Proposition 3.4.12. Recall that $\sigma^2_{\min}$ and $\sigma^2_{\max}$ are the output variances under the extreme weighted input models. We have

$$\begin{aligned}
& \mathbb{E}[(\mathscr{L}^{BEL} - \mathscr{L})^2] \\
= \; & \mathbb{E}[(\mathscr{L}^{BEL} - Z^{\min})^2] + 2\mathbb{E}[(\mathscr{L}^{BEL} - Z^{\min})(Z^{\min} - \mathscr{L})] + \mathbb{E}_{D,\xi_1}[(Z^{\min} - \mathscr{L})^2] \\
= \; & \mathbb{E}_{D,\xi_1}[\mathbb{E}_{\xi_2|D,\xi_1}[(\mathscr{L}^{BEL} - Z^{\min})^2]] + 2\mathbb{E}_{D,\xi_1}\big[\mathbb{E}_{\xi_2|D,\xi_1}[(\mathscr{L}^{BEL} - Z^{\min})(Z^{\min} - \mathscr{L})]\big] + \\
& \qquad O\big(\frac{1}{n^2} + \frac{1}{R_1}\big) \;\text{ by Theorem 3.4.11} \\
= \; & \mathbb{E}_{D,\xi_1}\big[\frac{1}{R_2}\sigma^2_{\min}\big] + \mathbb{E}_{D,\xi_1}\big[(Z^{\min} - \mathscr{L})\mathbb{E}_{\xi_2|D,\xi_1}[(\mathscr{L}^{BEL} - Z^{\min})]\big] + O\big(\frac{1}{n^2} + \frac{1}{R_1}\big) \\
\le \; & \frac{1}{R_2}\mathbb{E}_{D,\xi_1}\big[\mathbb{E}_{\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min}}[h^2(\mathbf{X}_1,\ldots,\mathbf{X}_m)]\big] + 0 + O\big(\frac{1}{n^2} + \frac{1}{R_1}\big) \\
= \; & O\big(\frac{1}{R_2}\big) + O\big(\frac{1}{n^2} + \frac{1}{R_1}\big) \;\text{ by (B.54) with } k = 2 \text{ from Lemma B.5.1} \\
= \; & O\big(\frac{1}{n^2} + \frac{1}{R_1} + \frac{1}{R_2}\big).
\end{aligned}$$

The bound for $\mathbb{E}[(\mathscr{U}^{BEL} - \mathscr{U})^2]$ can be obtained by the same argument. □

Proof of Proposition 3.4.13. We first establish the representations for $\mathscr{L}^{EEL}, \mathscr{U}^{EEL}$. The uniform moment convergence result (B.53) from Lemma B.5.1 implies that $\sigma^2_{\min} = \sigma^2 + O_p(1/\sqrt{n})$. By calculating the variance of sample variance, one can show that the $\hat{\sigma}^2_{\min}$ in Algorithm 5 satisfies $\mathbb{E}_{\xi_2|D,\xi_1}[(\hat{\sigma}^2_{\min} - \sigma^2_{\min})^2] \le C\mathbb{E}_{\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min}}[h^4(\mathbf{X}_1,\ldots,\mathbf{X}_m)]/R_2$ for some universal constant C. Using the result (B.54) with $k = 4$ we have $\mathbb{E}_{D,\xi_1}\big[\mathbb{E}_{\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min}}[h^4(\mathbf{X}_1,\ldots,\mathbf{X}_m)]\big] = O(1)$. Therefore we have $\mathbb{E}[(\hat{\sigma}^2_{\min} - \sigma^2_{\min})^2] = \mathbb{E}_{D,\xi_1}\big[\mathbb{E}_{\xi_2|D,\xi_1}[(\hat{\sigma}^2_{\min} - \sigma^2_{\min})^2]\big] = O(1/R_2)$, whereby

$$\hat{\sigma}^2_{\min} = \sigma^2_{\min} + O_p\big(\frac{1}{\sqrt{R_2}}\big) = \sigma^2 + O_p\big(\frac{1}{\sqrt{n}}\big) + O_p\big(\frac{1}{\sqrt{R_2}}\big) = \sigma^2 + o_p(1). \tag{B.59}$$

Now the lower confidence bound $\mathscr{L}^{EEL}$ from Algorithm 5 can be expressed as

$$\begin{aligned}
\mathscr{L}^{EEL} &= \hat{Z}^{\min} - z_{1-\alpha/2}\frac{\hat{\sigma}_{\min}}{\sqrt{R_2}} \\
&= \mathscr{L} + (Z^{\min} - \mathscr{L}) + \hat{Z}^{\min} - Z^{\min} - z_{1-\alpha/2}\frac{\hat{\sigma}_{\min}}{\sqrt{R_2}} \\
&= \mathscr{L} + O_p\big(\frac{1}{n} + \frac{1}{\sqrt{R_1}}\big) + \hat{Z}^{\min} - Z^{\min} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{R_2}} + o_p\big(\frac{1}{\sqrt{R_2}}\big) \\
&\quad \text{by (B.59) and Theorem 3.4.11} \\
&= \mathscr{L} + \hat{Z}^{\min} - Z^{\min} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{R_2}} + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big) \ \text{ because } \frac{R_1}{n} \to \infty \\
&= Z^* + \sum_{i=1}^{m} \bar{G}_i - z_{1-\alpha/2}\sigma_I + \hat{Z}^{\min} - Z^{\min} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{R_2}} + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big) \\
&\quad \text{because of (B.50).}
\end{aligned}$$

Rearranging the above gives the desired conclusion for $\mathscr{L}^{EEL}$. The representation for $\mathscr{U}^{EEL}$ can be obtained via a similar way.

To justify the representation for $\mathscr{L}^{FEL}$ and $\mathscr{U}^{FEL}$, we first need to establish the consistency of our input-induced variance estimate (3.6). Specifically, we have:

Lemma B.6.1 *Under Assumptions 3.3.1, 3.3.2 and 3.3.3, as $n \to \infty$ and $R_1/n \to \infty$ the input-induced variance estimate* (3.6) *is relatively consistent, i.e., $\hat{\sigma}_I^2/\sigma_I^2 \to 1$ in probability with respect to the joint randomness of both input data and simulation.*

Proof of Lemma B.6.1. Since the input-induced variance σ_I^2 is of order $1/n$ and the strong law of large numbers ensures that $\big(\sum_{i=1}^m \sum_{j=1}^{n_i} \big(G_i(X_{i,j})\big)^2/n_i^2\big)/\sigma_I^2 \to 1$ almost surely, it suffices to show

$$\sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \big(\hat{G}_i(X_{i,j})\big)^2 - \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \big(G_i(X_{i,j})\big)^2 = o_p\big(\frac{1}{n}\big), \tag{B.60}$$

$$\hat{\sigma}_I^2 - \sum_{i=1}^{m} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \big(\hat{G}_i(X_{i,j})\big)^2 = o_p\big(\frac{1}{n}\big). \tag{B.61}$$

We bound the left hand side of (B.60) as

$$\begin{aligned}
&\big|\text{left hand side of (B.60)}\big| \\
=\ &\Big|\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(2G_i(X_{i,j})(\hat{G}_i(X_{i,j})-G_i(X_{i,j}))+(\hat{G}_i(X_{i,j})-G_i(X_{i,j}))^2)\Big| \\
\le\ &\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\hat{G}_i(X_{i,j})-G_i(X_{i,j}))^2+2\sqrt{\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(G_i(X_{i,j}))^2\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\hat{G}_i(X_{i,j})-G_i(X_{i,j}))^2}.
\end{aligned}$$

Hence it suffices to bound the error $(\hat{G}_i(X_{i,j})-G_i(X_{i,j}))^2$ for each i,j. Seeing that both G_i and $\hat{G}_i$ take the form of a sum of conditional expectations, we can control this error via a similar analysis in proving Proposition 3.4.2. In particular, for all i,j we have $\mathbb{E}_D[(\hat{G}_i(X_{i,j})-G_i(X_{i,j}))^2]\le C/n$ for some constant C depending on h (a similar observation has been proved in equation (EC.10) in Lemma EC.1 of Lam and Qian (2018d)). Therefore $\big|\text{left hand side of (B.60)}\big| = O_p(1/n^2) + 2\sqrt{O_p(1/n)O_p(1/n^2)} = O_p(1/n^{\frac{3}{2}}) = o_p(1/n)$. Thus (B.60) follows.

(B.61) can be established in two steps. First we show that the bias correction term $\sum_{i=1}^{m}\frac{T_i\hat{\sigma}^2}{R_1} = o_p(1/n)$. Note that $\hat{\sigma}^2=\sigma^2+o_p(1)=O_p(1)$ can be proved via the same argument used to prove (B.59) but with the minimal weights $\mathbf{w}_i^{\min}, i=1,\ldots,m$ replaced by the uniform weights. When $R_1/n\to\infty$, we have each $\frac{T_i\hat{\sigma}^2}{R_1}=O_p(1/R_1)=o_p(1/n)$. Second, we examine the error

$$\begin{aligned}
&\Big|\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\big(\hat{\hat{G}}_i(X_{i,j})\big)^2-\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2\Big| \\
\le\ &\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\hat{\hat{G}}_i(X_{i,j})-\hat{G}_i(X_{i,j}))^2+2\sqrt{\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}(\hat{\hat{G}}_i(X_{i,j})-\hat{G}_i(X_{i,j}))^2} \\
=\ &O_p\big(\frac{1}{R_1}\big)+2\sqrt{O_p\big(\frac{1}{n}\big)O_p\big(\frac{1}{R_1}\big)}\ \text{ by Lemma B.5.2} \\
=\ &o_p\big(\frac{1}{n}\big)+2\sqrt{O_p\big(\frac{1}{n}\big)o_p\big(\frac{1}{n}\big)} \\
=\ &o_p\big(\frac{1}{n}\big).
\end{aligned}$$

This concludes (B.61). □

Given the relative consistency of the input-induced variance estimate $\hat{\sigma}_I^2$ in estimating σ_I^2, if we

couple the simulation runs of Algorithms 5 and 6, then

$$\begin{aligned}
\mathscr{L}^{FEL} &= \mathscr{L}^{EEL} + z_{1-\alpha/2}\frac{\hat{\sigma}_{\min}}{\sqrt{R_2}} - z_{1-\alpha/2}\Big(\sqrt{\hat{\sigma}_I^2 + \frac{\hat{\sigma}_{\min}^2}{R_2}} - \hat{\sigma}_I\Big) \\
&= \mathscr{L}^{EEL} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{R_2}} + o_p\big(\frac{1}{\sqrt{R_2}}\big) - z_{1-\alpha/2}\Big(\sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}} - \sigma_I\Big) + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big) \\
&= \mathscr{L}^{EEL} - z_{1-\alpha/2}\Big(\sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}} - \sigma_I - \frac{\sigma}{\sqrt{R_2}}\Big) + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big) \\
&= Z^* + \sum_{i=1}^{m}\bar{G}_i + \hat{Z}^{\min} - Z^{\min} - z_{1-\alpha/2}\sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}} + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big)
\end{aligned}$$

where in the last equality we use the representation for $\mathscr{L}^{EEL}$. The representation for the upper bound $\mathscr{U}^{FEL}$ can be similarly obtained. □

Note that the analysis in Lemma B.6.1 treats each $\frac{n_i T_i \hat{\sigma}^2}{R_1}$ in the estimator (3.6) as an asymptotically diminishing quantity instead of a bias correction term. The following proposition shows their effectiveness in correcting the bias of $\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\big(\hat{\hat{G}}_i(X_{i,j})\big)^2$ as a Monte Carlo estimator of $\sum_{i=1}^{m}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2$:

Proposition B.6.2 *Under Assumptions 3.3.1 and 3.3.3, as $n \to \infty$ and $R_1 \to \infty$, we have for each input model i that*

$$\mathbb{E}_{\xi_1|D}\big[\frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{\hat{G}}_i(X_{i,j})\big)^2\big] - \frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2 = \frac{n_i T_i \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)}{R_1} + o_p\big(\frac{n_i}{R_1}\big).$$

where the o_p is with respect to the randomness in the input data.

Our input variance estimator $\hat{\sigma}_I^2$ in (3.6) results from a direct use of Proposition B.6.2 to correct the bias of each $\frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{\hat{G}}_i(X_{i,j})\big)^2$ with $n_i T_i \hat{\sigma}^2/R_1$, a simulation estimate of $n_i T_i \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)/R_1$. Here we provide the proof of Proposition B.6.2: *Proof of Proposition B.6.2.* We have analyzed the bias and variance of each gradient component $\hat{\hat{G}}_i(X_{i,j})$ in the proof of Lemma B.5.2. The expected value is

$$\mathbb{E}_{\xi_1|D}[\hat{\hat{G}}_i(X_{i,j})] = \frac{R_1 - 1}{R_1}\hat{G}_i(X_{i,j})$$

and the variance is

$$\begin{aligned}\mathrm{Var}_{\xi_1|D}\big(\hat{\hat{G}}_i(X_{i,j})\big) &= \frac{(R_1-1)^2}{R_1^3}\Big(\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m)-Z(\hat{P}_1,\ldots,\hat{P}_m))^2(S_{i,j}(\mathbf{X}_i))^2\big]+\\ &\qquad \frac{n_iT_i(1-1/n_i)}{R_1-1}\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)-\frac{R_1-2}{R_1-1}(\hat{G}_i(X_{i,j}))^2\Big).\end{aligned}$$

Therefore the overall bias is

$$\begin{aligned}&\mathbb{E}_{\xi_1|D}\big[\frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{\hat{G}}_i(X_{i,j})\big)^2\big]-\frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2\\ =\;&\frac{1}{n_i}\sum_{j=1}^{n_i}\big[\big(\frac{R_1-1}{R_1}\hat{G}_i(X_{i,j})\big)^2+\mathrm{Var}_{\xi_1|D}\big(\hat{\hat{G}}_i(X_{i,j})\big)\big]-\frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2\\ =\;&\frac{(R_1-1)^2}{R_1^3n_i}\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m)-Z(\hat{P}_1,\ldots,\hat{P}_m))^2\sum_{j=1}^{n_i}(S_{i,j}(\mathbf{X}_i))^2\big]+\\ &\frac{R_1-1}{R_1^3}n_iT_i(1-\frac{1}{n_i})\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)-\frac{3R_1^2-4R_1+2}{R_1^3}\frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2\\ =\;&\frac{(R_1-1)^2}{R_1^3n_i}\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m)-Z(\hat{P}_1,\ldots,\hat{P}_m))^2\sum_{j=1}^{n_i}(S_{i,j}(\mathbf{X}_i))^2\big]+O_p\big(\frac{n_i}{R_1^2}+\frac{1}{R_1}\big)\end{aligned}\tag{B.62}$$

where the last equality holds because $\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h)=O_p(1)$, $\frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2=O_p(1)$ and T_i is a constant. We deal with the first term by expanding the score function $S_{i,j}(\mathbf{X}_i)$ as in the proof

of Lemma B.5.2, arriving at

$$
\begin{aligned}
& \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\big[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m) - Z(\hat{P}_1,\ldots,\hat{P}_m))^2 \sum_{j=1}^{n_i}(S_{i,j}(\mathbf{X}_i))^2\big] \\
= \;& \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\Bigg[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m) - Z(\hat{P}_1,\ldots,\hat{P}_m))^2 \sum_{j=1}^{n_i}\Bigg(T_i^2 + n_i^2 \sum_{s\neq t}\mathbf{1}\{X_i(t) = X_i(s) = X_{i,j}\} + \\
& \qquad (n_i^2 - 2T_i n_i)\sum_{t=1}^{T_i}\mathbf{1}\{X_i(t) = X_{i,j}\}\Bigg)\Bigg] \\
= \;& \mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}\Bigg[(h(\mathbf{X}_1,\ldots,\mathbf{X}_m) \\
& \qquad -Z(\hat{P}_1,\ldots,\hat{P}_m))^2\left(T_i^2 n_i + n_i^2\sum_{s\neq t}\mathbf{1}\{X_i(t) = X_i(s)\} + (n_i - 2T_i)n_i T_i\right)\Bigg] \\
= \;& n_i T_i(n_i - T_i)\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + n_i^2\sum_{s\neq t}\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[(h - Z(\hat{P}_1,\ldots,\hat{P}_m))^2 | X_i(t) = X_i(s)]\cdot\frac{1}{n_i} \\
= \;& n_i^2 T_i \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + O_p(n_i)
\end{aligned}
$$

where the last equality follows because each $\mathbb{E}_{\hat{P}_1,\ldots,\hat{P}_m}[(h - Z(\hat{P}_1,\ldots,\hat{P}_m))^2 | X_i(t) = X_i(s)] = O_p(1)$ and $\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) = O_p(1)$. Substituting the above bound into (B.62), we get the leading term of the bias

$$
\begin{aligned}
& \mathbb{E}_{\xi_1|D}\big[\frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{\hat{G}}_i(X_{i,j})\big)^2\big] - \frac{1}{n_i}\sum_{j=1}^{n_i}\big(\hat{G}_i(X_{i,j})\big)^2 \\
= \;& \frac{(R_1-1)^2}{R_1^3 n_i}\big(n_i^2 T_i \mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + O_p(n_i)\big) + O_p\big(\frac{n_i}{R_1^2} + \frac{1}{R_1}\big) \\
= \;& \frac{n_i T_i}{R_1}\mathrm{Var}_{\hat{P}_1,\ldots,\hat{P}_m}(h) + O_p\big(\frac{n_i}{R_1^2} + \frac{1}{R_1}\big).
\end{aligned}
$$

Noting that the O_p term is $o_p(n_i/R_1)$ as $n_i, R_1 \to \infty$ concludes the proposition. □

B.7 Proofs of Proposition 3.3.1 and Theorems 3.3.2, 3.3.3, 3.3.4

Proof of Proposition 3.3.1. It suffices to prove the theorem for the minimization problem. Since $w_{i,j} = \frac{1}{n_i}$ for each i, j is a solution in the (relative) interior of the feasible set, Slater's conditions

holds for (3.3). It is also clear, by a compactness argument, that the optimal value of the program is finite and attainable. By Corollary 28.3.1 of Rockafellar (1970), $(\mathbf{w}_1^{\min}, \ldots, \mathbf{w}_m^{\min})$ is a minimizer if and only if there exist Lagrange multipliers $\beta^*, \lambda_i^* \in \mathbb{R}, i = 1, \ldots, m$ such that the following KKT conditions are satisfied

$$
\begin{aligned}
&2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}^{\min}) + \mathcal{X}^2_{1,1-\alpha} \geq 0,\ \beta^* \geq 0 \\
&\beta^*\Big(2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}^{\min}) + \mathcal{X}^2_{1,1-\alpha}\Big) = 0 \\
&\sum_{j=1}^{n_i} w_{i,j}^{\min} = 1 \text{ for all } i = 1, \ldots, m \\
&\hat{\hat{G}}_i(X_{i,j}) + \lambda_i^* - \frac{2\beta^*}{w_{i,j}^{\min}} = 0 \text{ for all } i, j.
\end{aligned}
$$

When $\hat{\hat{G}}_{i_0}(X_{i_0,j_1}) \neq \hat{\hat{G}}_{i_0}(X_{i_0,j_2})$ for some $1 \leq i_0 \leq m$ and $1 \leq j_1 < j_2 \leq n_{i_0}$, the objective is a non-constant linear function and thus any minimizer must lie on the (relative) boundary of the feasible set, i.e. $2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}^{\min}) + \mathcal{X}^2_{1,1-\alpha} = 0$. Since the constraint $-2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log(n_i w_{i,j}) \leq \mathcal{X}^2_{1,1-\alpha}$ is strictly convex, the minimizer must be unique. Moreover, we show that β^* must be strictly positive in this case. Suppose $\beta^* = 0$ then the last equation of KKT conditions requires $\hat{\hat{G}}_i(X_{i,j}) = -\lambda_i^*$ for all i, j, which is a contradiction. Note that the minimizer must have positive components $w_{i,j}^{\min} > 0$ due to the logarithm in the constraint, hence

$$
w_{i,j}^{\min} = \frac{2\beta^*}{\hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*},\ \beta^* > 0,\ \hat{\hat{G}}_i(X_{i,j}) + \lambda_i^* > 0 \text{ for all } i, j, \tag{B.63}
$$

$$
2\sum_{i=1}^{m}\sum_{j=1}^{n_i}\log\frac{2n_i\beta^*}{\hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*} + \mathcal{X}^2_{1,1-\alpha} = 0,\ \sum_{j=1}^{n_i}\frac{2\beta^*}{\hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*} = 1 \text{ for all } i. \tag{B.64}
$$

To show that such $(\beta^*, \lambda_1^*, \ldots, \lambda_m^*)$ is also unique, let i_0, j_1, j_2 be the indices mentioned in the theorem. Then (B.63) stipulates $w_{i_0,j_1}^{\min}/w_{i_0,j_2}^{\min} = (\hat{\hat{G}}_{i_0,j_2} + \lambda_{i_0}^*)/(\hat{\hat{G}}_{i_0,j_1} + \lambda_{i_0}^*)$. Since the right hand side is strictly monotone in $\lambda_{i_0}^*$, the uniqueness of $w_{i,j}^{\min}$ implies the uniqueness of $\lambda_{i_0}^*$, which in turn implies the uniqueness of β^* and other λ_i^*'s due to the second equation of line (B.64).

We further show that β^* must lie in the interval given in the proposition. We first argue that

there is at least one $i \in \{1, \ldots, m\}$ such that

$$\frac{\min_j \hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*}{\max_j \hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*} < e^{-\frac{\chi^2_{1,1-\alpha}}{2N}}. \tag{B.65}$$

Suppose $(\min_j \hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*)/(\max_j \hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*) \geq e^{-\frac{\chi^2_{1,1-\alpha}}{2N}}$ for all i, then $\sum_{j=1}^{n_i} 2\beta^*/(\hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*) = 1$ implies that $2\beta^*/(\hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*) \geq \frac{1}{n_i} e^{-\frac{\chi^2_{1,1-\alpha}}{2N}}$ for all i, j and the inequality must be strict for some i, j because $e^{-\frac{\chi^2_{1,1-\alpha}}{2N}} < 1$. Therefore

$$2\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log \frac{2n_i\beta^*}{\hat{\hat{G}}_i(X_{i,j}) + \lambda_i^*} + \mathcal{X}^2_{1,1-\alpha} > -2\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{\mathcal{X}^2_{1,1-\alpha}}{2N} + \mathcal{X}^2_{1,1-\alpha} = 0$$

which contradicts (B.64). Now let $\lambda_{i'}^*$ be a multiplier that satisfies (B.65). Rearranging (B.65) gives

$$\lambda_{i'}^* < \frac{e^{-\frac{\chi^2_{1,1-\alpha}}{2N}} \max_j \hat{\hat{G}}_{i'}(X_{i',j}) - \min_j \hat{\hat{G}}_{i'}(X_{i',j})}{1 - e^{-\frac{\chi^2_{1,1-\alpha}}{2N}}}. \tag{B.66}$$

Hence

$$\begin{aligned}
1 = \sum_{j=1}^{n_{i'}} \frac{2\beta^*}{\hat{\hat{G}}_{i'}(X_{i',j}) + \lambda_{i'}^*} &\geq \frac{2n_{i'}\beta^*}{\max_j \hat{\hat{G}}_{i'}(X_{i',j}) + \lambda_{i'}^*} \\
&> \frac{2n_{i'}\beta^*(1 - e^{-\frac{\chi^2_{1,1-\alpha}}{2N}})}{\max_j \hat{\hat{G}}_{i'}(X_{i',j}) - \min_j \hat{\hat{G}}_{i'}(X_{i',j})} \text{ by using the upper bound (B.66)} \\
&\geq \frac{2\min_i n_i\beta^*(1 - e^{-\frac{\chi^2_{1,1-\alpha}}{2N}})}{\max\{\max_j \hat{\hat{G}}_i(X_{i,j}) - \min_j \hat{\hat{G}}_i(X_{i,j}) | i = 1, \ldots, m\}}.
\end{aligned}$$

Rearranging the above inequality gives the desired upper bound for β^*.

If $\hat{\hat{G}}_i(X_{i,j}) = c_i$ for some constant c_i, then the objective is the constant function $\sum_{i=1}^m c_i$, hence every feasible solution is optimal. □

Proof of Theorem 3.3.2. When $R_1/n \to \infty$ and $R_2/n \to \infty$, Proposition 3.4.12 stipulates that $\mathscr{L}^{BEL} = \mathscr{L} + o_p(1/\sqrt{n})$ and $\mathscr{U}^{BEL} = \mathscr{U} + o_p(1/\sqrt{n})$. Theorem 3.4.8 then implies that the asymptotic representation (B.50) holds for $\mathscr{L}^{BEL}$ and $\mathscr{U}^{BEL}$. The rest of the proof is the same as

that of Theorem 3.4.8 from (B.50) onwards. □

Proof of Theorems 3.3.3 and 3.3.4. For convenience, all limit statements are understood to be for $n, R_1, R_2 \to \infty$ such that $\frac{R_1}{n} \to \infty, \frac{R_2}{n} \leq M$ (e.g., (B.67) and (B.68)), unless stated otherwise. We need the Berry-Esseen Theorem stated as:

Lemma B.7.1 (Theorem 3.4.9 in Durrett 2010) *Let $\{\eta_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. random variables such that $\mathbb{E}[\eta_1] = 0, \mathbb{E}[\eta_1^2] = \sigma_\eta^2, \mathbb{E}[|\eta_1|^3] = \rho_\eta < \infty$, and $S_n = \sum_{i=1}^n \eta_i/(\sigma_\eta\sqrt{n})$. Let $F_n(\cdot)$ be the cumulative distribution function of S_n. Then*

$$\sup_{x\in\mathbb{R}}|F_n(x) - \Phi(x)| \leq \frac{3\rho_\eta}{\sigma_\eta^3\sqrt{n}}.$$

We first show the following weak convergence to the joint standard normal

$$\Big(\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I}, \frac{\sqrt{R_2}(\hat{Z}^{\min} - Z^{\min})}{\sigma}, \frac{\sqrt{R_2}(\hat{Z}^{\max} - Z^{\max})}{\sigma}\Big) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_3). \tag{B.67}$$

Since $\sigma_{\min}^2 = \sigma^2 + o_p(1)$ and $\sigma_{\max}^2 = \sigma^2 + o_p(1)$ as argued in (B.59), to show (B.67) it suffices to show

$$\Big(\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I}, \frac{\sqrt{R_2}(\hat{Z}^{\min} - Z^{\min})}{\sigma_{\min}}, \frac{\sqrt{R_2}(\hat{Z}^{\max} - Z^{\max})}{\sigma_{\max}}\Big) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{I}_3) \tag{B.68}$$

and then apply Slutsky's theorem. For any $(x, y, z) \in \mathbb{R}^3$, we compute the joint probability

$$
\begin{aligned}
& P\Big(\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \le x, \frac{\sqrt{R_2}(\hat{Z}^{\min} - Z^{\min})}{\sigma_{\min}} \le y, \frac{\sqrt{R_2}(\hat{Z}^{\max} - Z^{\max})}{\sigma_{\max}} \le z\Big) \\
= \; & \mathbb{E}\left[\mathbf{1}\left\{\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \le x\right\} \cdot \mathbf{1}\left\{\frac{\sqrt{R_2}(\hat{Z}^{\min} - Z^{\min})}{\sigma_{\min}} \le y\right\} \cdot \mathbf{1}\left\{\frac{\sqrt{R_2}(\hat{Z}^{\max} - Z^{\max})}{\sigma_{\max}} \le z\right\}\right] \\
= \; & \mathbb{E}_{D,\xi_1}\left[\mathbf{1}\left\{\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \le x\right\} \mathbb{E}_{\xi_2|D,\xi_1}\left[\mathbf{1}\left\{\frac{\sqrt{R_2}(\hat{Z}^{\min} - Z^{\min})}{\sigma_{\min}} \le y\right\}\right]\right. \\
& \left. \cdot \mathbb{E}_{\xi_2|D,\xi_1}\left[\mathbf{1}\left\{\frac{\sqrt{R_2}(\hat{Z}^{\max} - Z^{\max})}{\sigma_{\max}} \le z\right\}\right]\right] \\
& \quad \text{by conditional independence of } \hat{Z}^{\min} \text{ and } \hat{Z}^{\max} \text{ given input data and Step 1} \\
= \; & \mathbb{E}_{D,\xi_1}\left[\mathbf{1}\left\{\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \le x\right\} (\Phi(y) + \epsilon^{\min})(\Phi(z) + \epsilon^{\max})\right] \quad \text{for some error terms } \epsilon^{\min} \text{ and } \epsilon^{\max} \\
= \; & P\Big(\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \le x\Big)\Phi(y)\Phi(z) + \mathbb{E}_{D,\xi_1}\left[\mathbf{1}\left\{\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} \le x\right\} (\Phi(y)\epsilon^{\max} + \Phi(z)\epsilon^{\min} + \epsilon^{\min}\epsilon^{\max})\right] \qquad \text{(B.69)}
\end{aligned}
$$

We need to show that the second term in (B.69) vanishes. Denoting by

$$
\begin{aligned}
\rho_{\min} &:= \mathbb{E}_{\mathbf{w}_1^{\min},\ldots,\mathbf{w}_m^{\min}}[|h(\mathbf{X}_1, \ldots, \mathbf{X}_m) - Z^{\min}|^3] \\
\rho_{\max} &:= \mathbb{E}_{\mathbf{w}_1^{\max},\ldots,\mathbf{w}_m^{\max}}[|h(\mathbf{X}_1, \ldots, \mathbf{X}_m) - Z^{\max}|^3]
\end{aligned}
$$

the third order central moments of the output h driven by the extreme weighted inputs. The errors $\epsilon^{\min}, \epsilon^{\max}$ then satisfy $|\epsilon^{\min}| \le \min\big\{1, \frac{3\rho_{\min}}{\sigma_{\min}^3\sqrt{R_2}}\big\}, |\epsilon^{\max}| \le \min\big\{1, \frac{3\rho_{\max}}{\sigma_{\max}^3\sqrt{R_2}}\big\}$. On one hand (B.54) entails that $\rho_{\min} = O_p(1)$ and $\rho_{\max} = O_p(1)$. On the other hand, $\sigma_{\min}^2 = \sigma^2 + o_p(1)$ and $\sigma_{\max}^2 = \sigma^2 + o_p(1)$ as mentioned before. These two facts together lead to $\epsilon^{\min} = O_p(1/\sqrt{R_2})$ and $\epsilon^{\max} = O_p(1/\sqrt{R_2})$. Since both errors do not exceed 1, by the dominated convergence theorem, the second term in (B.69) converges to zero asymptotically. Moreover, the probability $P\big(\sum_{i=1}^m \bar{G}_i \le x\sigma_I\big) \to \Phi(x)$ which has been shown in (B.52). Therefore the joint probability converges to $\Phi(x)\Phi(y)\Phi(z)$, hence weak convergence (B.68) holds by definition.

Secondly, we prove that $[\mathscr{L}^{FEL}, \mathscr{U}^{FEL}]$ is asymptotically valid, i.e., the lim inf part in Theorem 3.3.4. The lim inf result for $[\mathscr{L}^{EEL}, \mathscr{U}^{EEL}]$ is then a direct consequence of $[\mathscr{L}^{FEL}, \mathscr{U}^{FEL}]$ by a coupling argument as follows. If Algorithms 5 and 6 use the same $R_1 + 2R_2$ simulation runs, then

the two different adjustments in Step 3 satisfy $\frac{\hat{\sigma}_{\min}}{\sqrt{R_2}} \geq \sqrt{\hat{\sigma}_I^2 + \frac{\hat{\sigma}_{\min}^2}{R_2}} - \hat{\sigma}_I$ almost surely, therefore $\mathscr{L}^{EEL} \leq \mathscr{L}^{FEL}$ and $\mathscr{U}^{EEL} \geq \mathscr{U}^{FEL}$ almost surely. We proceed to prove the lim inf statement for $[\mathscr{L}^{FEL}, \mathscr{U}^{FEL}]$. We write

$$\begin{aligned} & P(\mathscr{L}^{FEL} \leq Z^* \leq \mathscr{U}^{FEL}) \\ = \; & P(\mathscr{L}^{FEL} \leq Z^*) + P(Z^* \leq \mathscr{U}^{FEL}) - P(\mathscr{L}^{FEL} \leq Z^* \text{ or } Z^* \leq \mathscr{U}^{FEL}) \\ = \; & P(\mathscr{L}^{FEL} \leq Z^*) + P(Z^* \leq \mathscr{U}^{FEL}) - 1 + P(\mathscr{U}^{FEL} < Z^* < \mathscr{L}^{FEL}). \end{aligned} \tag{B.70}$$

To compute the probabilities in (B.70), we use the representation from Proposition 3.4.13 to get

$$\begin{aligned} & P(\mathscr{L}^{FEL} \leq Z^*) \\ = \; & P\Big(\sum_{i=1}^m \bar{G}_i + \hat{Z}^{\min} - Z^{\min} - z_{1-\alpha/2}\sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}} + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big) \leq 0\Big) \\ = \; & P\Big(\frac{1}{\sqrt{\sigma_I^2 + \sigma^2/R_2}}\big(\sum_{i=1}^m \bar{G}_i + \hat{Z}^{\min} - Z^{\min}\big) + o_p(1) \leq z_{1-\alpha/2}\Big) \\ = \; & P\Big(\frac{\sigma_I}{\sqrt{\sigma_I^2 + \sigma^2/R_2}}\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} + \frac{\sigma/\sqrt{R_2}}{\sqrt{\sigma_I^2 + \sigma^2/R_2}}\frac{\sqrt{R_2}(\hat{Z}^{\min} - Z^{\min})}{\sigma} + o_p(1) \leq z_{1-\alpha/2}\Big) \\ \to \; & 1 - \frac{\alpha}{2}. \end{aligned}$$

The limit here is valid because, by rewriting the last probability above as the probability of a half-space of $\mathbb{R}^3$ like in (B.51), we can conclude from (B.67) and Lemma B.4.2 that

$$\frac{\sigma_I}{\sqrt{\sigma_I^2 + \sigma^2/R_2}}\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} + \frac{\sigma/\sqrt{R_2}}{\sqrt{\sigma_I^2 + \sigma^2/R_2}}\frac{\sqrt{R_2}(\hat{Z}^{\min} - Z^{\min})}{\sigma} \Rightarrow \mathcal{N}(0,1)$$

which continues to hold with an additional $o_p(1)$ term on the left hand side by Slutsky's Theorem. Similary, one can show that $P(\mathscr{U}^{FEL} \geq Z^*) \to 1 - \alpha/2$. Neglecting the last probability in (B.70) gives

$$P(\mathscr{L}^{FEL} \leq Z^* \leq \mathscr{U}^{FEL}) \geq P(\mathscr{L}^{FEL} \leq Z^*) + P(Z^* \leq \mathscr{U}^{FEL}) - 1 \to 2\big(1 - \frac{\alpha}{2}\big) - 1 = 1 - \alpha$$

from which the lim inf result follows.

Thirdly, we prove the lim sup results by further analyzing the last probability in (B.70). Using the representation from Proposition 3.4.13 again we have

$$\begin{aligned}
& P(\mathscr{U}^{FEL} < Z^* < \mathscr{L}^{FEL}) \\
= \; & P\Big(\frac{\sigma_I}{\sqrt{\sigma_I^2+\sigma^2/R_2}}\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} + \frac{\sigma/\sqrt{R_2}}{\sqrt{\sigma_I^2+\sigma^2/R_2}}\frac{\sqrt{R_2}(\hat{Z}^{\min}-Z^{\min})}{\sigma} + o_p(1) > z_{1-\alpha/2} \text{ and} \\
& \quad -\frac{\sigma_I}{\sqrt{\sigma_I^2+\sigma^2/R_2}}\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} - \frac{\sigma/\sqrt{R_2}}{\sqrt{\sigma_I^2+\sigma^2/R_2}}\frac{\sqrt{R_2}(\hat{Z}^{\max}-Z^{\max})}{\sigma} + o_p(1) > z_{1-\alpha/2}\Big) \\
= \; & P\Big(\frac{\sigma_I}{\sqrt{\sigma_I^2+\sigma^2/R_2}}\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} + \frac{\sigma/\sqrt{R_2}}{\sqrt{\sigma_I^2+\sigma^2/R_2}}\Big(\frac{\sqrt{R_2}(\hat{Z}^{\min}-Z^{\min})}{\sigma} + o_p(1)\Big) > z_{1-\alpha/2} \text{ and} \\
& \quad -\frac{\sigma_I}{\sqrt{\sigma_I^2+\sigma^2/R_2}}\frac{\sum_{i=1}^m \bar{G}_i}{\sigma_I} - \frac{\sigma/\sqrt{R_2}}{\sqrt{\sigma_I^2+\sigma^2/R_2}}\Big(\frac{\sqrt{R_2}(\hat{Z}^{\max}-Z^{\max})}{\sigma} + o_p(1)\Big) > z_{1-\alpha/2}\Big)
\end{aligned}$$

where the second equality is valid because $\frac{R_2}{n} \le M < \infty$ implies $\frac{\sigma/\sqrt{R_2}}{\sqrt{\sigma_I^2+\sigma^2/R_2}} \ge \epsilon > 0$ for some fixed constant ϵ. By Slutsky's theorem, if the three-dimensional random vector in (B.67) is contaminated by a negligible noise of size $o_p(1)$ in each component, it still converges weakly to the joint standard normal. This convergence, together with Lemma B.4.2, leads to the following limit

$$P(\mathscr{U}^{FEL} < Z^* < \mathscr{L}^{FEL}) - P(\tilde{W}_1 > z_{1-\alpha/2}, \tilde{W}_2 > z_{1-\alpha/2}) \to 0$$

where $(\tilde{W}_1, \tilde{W}_2)$ is the joint normal $\mathcal{N}\Big(\mathbf{0}, \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\Big)$ and $\rho = \sigma_I^2/(\sigma_I^2+\sigma^2/R_2) > 0$. To compute the limit probability, note that the conditional distribution $\tilde{W}_2|\tilde{W}_1$ is $\mathcal{N}(-\rho\tilde{W}_1, 1-\rho^2)$, therefore

$$\begin{aligned}
& P(\tilde{W}_1 > z_{1-\alpha/2}, \tilde{W}_2 > z_{1-\alpha/2}) \\
= & \int_{z_{1-\alpha/2}}^{\infty} \phi(x) P(\mathcal{N}(-\rho x, 1-\rho^2) > z_{1-\alpha/2}) dx \le \frac{\alpha}{2}\int_{z_{1-\alpha/2}}^{\infty} \phi(x) dx = \frac{\alpha^2}{4}.
\end{aligned}$$

Here ϕ denotes the density of the standard normal, and the inequality follows since $-\rho x < 0$ and

$1-\rho^2 < 1$ and hence $P(\mathcal{N}(-\rho x, 1-\rho^2) > z_{1-\alpha/2}) \leq P(\mathcal{N}(0,1) > z_{1-\alpha/2}) = \alpha/2$. This establishes

$$\limsup P(\mathscr{U}^{FEL} < Z^* < \mathscr{L}^{FEL}) \leq \frac{\alpha^2}{4}.$$

Substituting it into (B.70) gives the lim sup statement of Theorem 3.3.4.

Following the above line of analysis, the lim sup statement of Theorem 3.3.3 can be derived. We use the representation from Proposition 3.4.13. Since $\sigma_I + \frac{\sigma}{\sqrt{R_2}} \leq \sqrt{2}\sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}}$, we have

$$\mathscr{L}^{EEL} \geq \tilde{\mathscr{L}} := Z^* + \sum_{i=1}^{m} \bar{G}_i + (\hat{Z}^{\min} - Z^{\min}) - \sqrt{2} z_{1-\alpha/2} \sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}} + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big)$$
$$\mathscr{U}^{EEL} \leq \tilde{\mathscr{U}} := Z^* + \sum_{i=1}^{m} \bar{G}_i + (\hat{Z}^{\max} - Z^{\max}) + \sqrt{2} z_{1-\alpha/2} \sqrt{\sigma_I^2 + \frac{\sigma^2}{R_2}} + o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big)$$

almost surely, where the $o_p\big(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R_2}}\big)$ terms are those from Proposition 3.4.13. Repeating the above analysis for $\tilde{\mathscr{L}}, \tilde{\mathscr{U}}$ reveals that

$$\limsup P\left(\tilde{\mathscr{L}} \leq Z^* \leq \tilde{\mathscr{U}}\right) \leq 1 - \tilde{\alpha} + \frac{\tilde{\alpha}^2}{4}.$$

The same lim sup bound then holds for $\mathscr{L}^{EEL}, \mathscr{U}^{EEL}$ because $\mathscr{L}^{EEL} \geq \tilde{\mathscr{L}}$ and $\mathscr{U}^{EEL} \leq \tilde{\mathscr{U}}$.

Lastly, when R_2 also grows at a faster rate than n, the adjustments in Algorithms 5 and 6 relative to Algorithm 4 are of order $o_p(1/\sqrt{n})$, i.e., $\hat{\sigma}_{\min}/\sqrt{R_2} = o_p(1/\sqrt{n}), \hat{\sigma}_{\max}/\sqrt{R_2} = o_p(1/\sqrt{n})$ and $\sqrt{\hat{\sigma}_I^2 + \hat{\sigma}_{\min}^2/R_2} - \hat{\sigma}_I = o_p(1/\sqrt{n}), \sqrt{\hat{\sigma}_I^2 + \hat{\sigma}_{\max}^2/R_2} - \hat{\sigma}_I = o_p(1/\sqrt{n})$. Therefore, by coupling the simulation runs in Step 3 with Algorithm 4, the confidence bounds from Algorithms 5 and 6 differ from those from Algorithm 4 by $o_p(1/\sqrt{n})$. Using the proof for Theorem 3.3.2 concludes asymptotic exactness. □

Appendix C

Technical Proofs for Chapter 4

This chapter presents proofs for results in Chapter 4.

C.1 Proof of Lemma 4.5.1

We show the lemma under the two alternate sets of listed assumptions. First is under Assumptions 4.4.1 and 4.5.5. Second is under Assumptions 4.2.1, 4.4.1, 4.5.6 and that $\mathcal{X}$ is compact. *Proof of Lemma 4.5.1 (using Assumption 4.5.5).* Under Assumption 4.4.1, by (4.16) we have $EH_k(\xi_1, \ldots, \xi_k)^2 < \infty$ and hence $Var(g_k(\xi)) = Var(E[H_k(\xi_1, \ldots, \xi_k)|\xi_1]) \leq Var(H_k(\xi_1, \ldots, \xi_k))$ is

well-defined and finite. By the Chebyshev inequality, we have

$$
\begin{aligned}
& k^2 Var(g_k(\xi)) \\
= \;& k^2 Var(E[H_k(\xi_1,\ldots,\xi_k)|\xi_1]) \\
= \;& Var\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\Big|\xi_1\right]\right) \\
\geq \;& \eta^2 P\left(\left|E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\Big|\xi_1\right] - E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\right]\right| > \eta\right) \\
& \text{since } E\left[E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\Big|\xi_1\right]\right] = E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\right] \\
= \;& \eta^2\Bigg[P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\Big|\xi_1\right] - E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\right] > \eta\right) \\
& + P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\Big|\xi_1\right] - E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\right] < -\eta\right)\Bigg]
\end{aligned} \tag{C.1}
$$

Consider the two terms in (C.1). We use a coupling argument to bound them and show that they lead to the two terms in (4.17) that are independent of k. For the first term in (C.1),

$$
\begin{aligned}
& P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\Big|\xi_1\right] - E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\right] > \eta\right) \\
= \;& P\left(E\left[\min_{x\in\mathcal{X}}\left\{h(x,\xi_1') + \sum_{i\neq 1} h(x,\xi_i)\right\}\Big|\xi_1'\right] - E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\right] > \eta\right) \\
& \text{where } \xi_1, \xi_1', \xi_2, \ldots, \xi_k \text{ are all independent} \\
= \;& P\left(E\left[\min_{x\in\mathcal{X}}\left\{h(x,\xi_1') + \sum_{i\neq 1} h(x,\xi_i)\right\} - \min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\Big|\xi_1'\right] > \eta\right) \\
\geq \;& P\left(E\left[h(x_\epsilon(\boldsymbol{\xi}'),\xi_1') + \sum_{i\neq 1} h(x_\epsilon(\boldsymbol{\xi}'),\xi_i) - \sum_{i=1}^{k} h(x_\epsilon(\boldsymbol{\xi}'),\xi_i)\Big|\xi_1'\right] > \eta+\epsilon\right)
\end{aligned} \tag{C.2}
$$

where $x_\epsilon(\boldsymbol{\xi}')$ is an ϵ-optimal solution for the optimization $\min_{x\in\mathcal{X}}\left\{h(x,\xi_1') + \sum_{i\neq 1} h(x,\xi_i)\right\}$ that only depends on $\boldsymbol{\xi}' = \{\xi_1', \xi_2, \ldots, \xi_k\}$. The last inequality follows since by the definition of $x_\epsilon(\boldsymbol{\xi}')$ it holds that $\min_{x\in\mathcal{X}}\left\{h(x,\xi_1') + \sum_{i\neq 1} h(x,\xi_i)\right\} \geq h(x_\epsilon(\boldsymbol{\xi}'),\xi_1') + \sum_{i\neq 1} h(x_\epsilon(\boldsymbol{\xi}'),\xi_i) - \epsilon$, and

$\min_{x\in\mathcal{X}} \sum_{i=1}^{k} h(x,\xi_i) \leq \sum_{i=1}^{k} h(x_\epsilon(\boldsymbol{\xi}'),\xi_i)$ by the definition of minimization. Note that (C.2) is equal to

$$\begin{aligned}
& P\left(E\left[h(x_\epsilon(\boldsymbol{\xi}'),\xi_1') - h(x_\epsilon(\boldsymbol{\xi}'),\xi_1)\middle|\xi_1'\right] > \eta+\epsilon\right) \\
= \;& P\left(E\left[h(x_\epsilon(\boldsymbol{\xi}'),\xi_1') - Z(x_\epsilon(\boldsymbol{\xi}'))\middle|\xi_1'\right] > \eta+\epsilon\right) \text{ since } \xi_1 \text{ is independent of } x_\epsilon(\boldsymbol{\xi}') \text{ and } \xi_1' \\
\geq \;& P\left(E\left[\min_{x\in\mathcal{X}}\{h(x,\xi_1') - Z(x)\}\middle|\xi_1'\right] > \eta+\epsilon\right) \\
= \;& P\left(\min_{x\in\mathcal{X}}\{h(x,\xi_1') - Z(x)\} > \eta+\epsilon\right) \qquad \text{(C.3)}
\end{aligned}$$

Similarly, for the second term in (C.1), we have

$$
\begin{aligned}
& P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)\middle|\xi_1\right]-E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)\right]<-\eta\right)\\
=\;& P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)\right]-E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)\middle|\xi_1\right]>\eta\right)\\
=\;& P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)\right]-E\left[\min_{x\in\mathcal{X}}\left\{h(x,\xi_1')+\sum_{i\neq 1}h(x,\xi_i)\right\}\middle|\xi_1'\right]>\eta\right)\\
& \text{where } \xi_1,\xi_1',\xi_2,\ldots,\xi_k \text{ are all independent}\\
=\;& P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)-\min_{x\in\mathcal{X}}\left\{h(x,\xi_1')+\sum_{i\neq 1}h(x,\xi_i)\right\}\middle|\xi_1'\right]>\eta\right)\\
\geq\;& P\left(E\left[\sum_{i=1}^{k}h(x_\epsilon(\boldsymbol{\xi}),\xi_i)-h(x_\epsilon(\boldsymbol{\xi}),\xi_1')-\sum_{i\neq 1}h(x_\epsilon(\boldsymbol{\xi}),\xi_i)\middle|\xi_1'\right]>\eta+\epsilon\right)\\
& \text{where } x_\epsilon(\boldsymbol{\xi}), \text{ with } \boldsymbol{\xi}=\{\xi_1,\ldots,\xi_n\}, \text{ is an } \epsilon\text{-optimal solution of } \min_{x\in\mathcal{X}}\sum_{i=1}^{n}h(x,\xi_i);\\
& \text{this follows since } \min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)+\epsilon\geq\sum_{i=1}^{k}h(x_\epsilon(\boldsymbol{\xi}),\xi_i) \text{ by the definition of } x_\epsilon(\boldsymbol{\xi}) \text{ and}\\
& \min_{x\in\mathcal{X}}\left\{h(x,\xi_1')+\sum_{i\neq 1}h(x,\xi_i)\right\}\leq h(x_\epsilon(\boldsymbol{\xi}),\xi_1')+\sum_{i\neq 1}h(x_\epsilon(\boldsymbol{\xi}),\xi_i)\\
=\;& P\left(E\left[h(x_\epsilon(\boldsymbol{\xi}),\xi_1)-h(x_\epsilon(\boldsymbol{\xi}),\xi_1')\middle|\xi_1'\right]>\eta+\epsilon\right)\\
\geq\;& P\left(E\left[\min_{x\in\mathcal{X}}\{h(x,\xi_1)-h(x,\xi_1')\}\middle|\xi_1'\right]>\eta+\epsilon\right) \qquad \text{(C.4)}
\end{aligned}
$$

Combining (C.3) and (C.4) into (C.1), we get (4.17). □

Proof of Lemma 4.5.1 (using Assumption 4.5.6). We first argue consistency of the SAA solutions. Since Assumptions 4.2.1 and 4.4.1 hold, by Theorem 7.48 in Shapiro et al. (2014) we have $\sup_{x\in\mathcal{X}}\left|\frac{1}{k}\sum_{i=1}^{k}h(x,\xi_i)-Z(x)\right|\to 0$ almost surely as $k\to\infty$. Denote by $\hat{\mathcal{X}}_k^*$ the set of optimal solutions for the SAA problem formed by $\xi_1,\ldots,\xi_k$. Note that $\hat{\mathcal{X}}_k^*\neq\emptyset$ because of Lipschitz continuity and compactness of $\mathcal{X}$. Assumption 4.2.1 also implies Lipschitzness of Z, i.e. $|Z(x_1)-Z(x_2)|\leq EM(\xi)\|x_1-x_2\|$. With all these ingredients, Theorem 5.3 in Shapiro et al. (2014) then ensures

that almost surely $\sup_{x\in\hat{\mathcal{X}}_k^*}\inf_{x'\in\mathcal{X}^*}\|x-x'\|\to 0$. Moreover, since $\sup_{x\in\hat{\mathcal{X}}_k^*}\inf_{x'\in\mathcal{X}^*}\|x-x'\|\le D_{\mathcal{X}}$, where $D_{\mathcal{X}}$ is the diameter of $\mathcal{X}$, we have $E[\sup_{x\in\hat{\mathcal{X}}_k^*}\inf_{x'\in\mathcal{X}^*}\|x-x'\|^2]\to 0$ by bounded convergence theorem.

We now follow the line of arguments in the proof that uses Assumption 4.5.5. Here, we can work with exact optimal solutions in place of ϵ-optimal solutions because $\hat{\mathcal{X}}_k^*\neq\emptyset$. Following the coupling argument in the previous proof, we have for the first term in (C.1)

$$\begin{aligned}
& P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^k h(x,\xi_i)\Big|\xi_1\right]-E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^k h(x,\xi_i)\right]>\eta\right)\\
\ge\ & P\left(E\left[h(x(\boldsymbol{\xi}'),\xi_1')-Z(x(\boldsymbol{\xi}'))\Big|\xi_1'\right]>\eta\right)\\
\ge\ & P\left(E\left[h(x'(\boldsymbol{\xi}'),\xi_1')-Z(x'(\boldsymbol{\xi}'))-(M(\xi_1')+EM(\xi))\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\|\Big|\xi_1'\right]>\eta\right)\\
\ge\ & P\left(E\left[\min_{x\in\mathcal{X}^*}\{h(x,\xi_1')-Z^*\}-(M(\xi_1')+EM(\xi))\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\|\Big|\xi_1'\right]>\eta\right)\\
\ge\ & P\left(\min_{x\in\mathcal{X}^*}\{h(x,\xi_1')-Z^*\}-(M(\xi_1')+EM(\xi))E\left[\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\|\Big|\xi_1'\right]>\eta\right)
\end{aligned}$$

where $x(\boldsymbol{\xi}')$ is an optimal solution for $\min_{x\in\mathcal{X}}\{h(x,\xi_1')+\sum_{i=2}^k h(x,\xi_i)\}$ and $x'(\boldsymbol{\xi}')\in\mathcal{X}^*$ minimizes $\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\|$ (minimum is achieved because $\mathcal{X}^*$ is compact). Since

$$E\left[E\left[\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\|\Big|\xi_1'\right]\right]\le E\left[\sup_{x\in\hat{\mathcal{X}}_k^*}\inf_{x'\in\mathcal{X}^*}\|x-x'\|\right]\le\left(E\left[\sup_{x\in\hat{\mathcal{X}}_k^*}\inf_{x'\in\mathcal{X}^*}\|x-x'\|^2\right]\right)^{\frac{1}{2}}\to 0$$

we know $E[\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\||\xi_1']=o_p(1)$ on one hand. On the other hand, $M(\xi_1')+E\,|M(\xi)|=O_p(1)$, hence $(M(\xi_1')+EM(\xi))E[\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\||\xi_1']=o_p(1)$. By Slutsky's theorem

$$\min_{x\in\mathcal{X}^*}\{h(x,\xi_1')-Z^*\}-(M(\xi_1')+EM(\xi))E\left[\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\|\Big|\xi_1'\right]\Rightarrow\min_{x\in\mathcal{X}^*}\{h(x,\xi_1')-Z^*\}$$

which leads to

$$\begin{aligned}
& \liminf_{k\to\infty}P\left(\min_{x\in\mathcal{X}^*}\{h(x,\xi_1')-Z^*\}-(M(\xi_1')+EM(\xi))E\left[\|x(\boldsymbol{\xi}')-x'(\boldsymbol{\xi}')\|\Big|\xi_1'\right]>\eta\right)\\
\ge\ & P\left(\min_{x\in\mathcal{X}^*}\{h(x,\xi_1')-Z^*\}>\eta\right).
\end{aligned}$$

For the second term in (C.1), we have the following lower bound by a similar argument

$$\begin{aligned} & P\left(E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)\middle|\xi_1\right]-E\left[\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)\right]<-\eta\right) \\ \geq\ & P\left(E\left[h(x(\boldsymbol{\xi}),\xi_1)-h(x(\boldsymbol{\xi}),\xi_1')\middle|\xi_1'\right]>\eta\right) \\ \geq\ & P\left(E\left[\min_{x\in\mathcal{X}^*}\{h(x,\xi_1)-h(x,\xi_1')\}\middle|\xi_1'\right]-E[M(\xi_1)\|x(\boldsymbol{\xi})-x'(\boldsymbol{\xi})\|]-M(\xi_1')E\|x(\boldsymbol{\xi})-x'(\boldsymbol{\xi})\|>\eta\right) \end{aligned}$$

where $x(\boldsymbol{\xi})$ is an optimal solution for $\min_{x\in\mathcal{X}}\sum_{i=1}^{k}h(x,\xi_i)$ and $x'(\boldsymbol{\xi})\in\mathcal{X}^*$ minimizes $\|x(\boldsymbol{\xi})-x'(\boldsymbol{\xi})\|$. Again $E[M(\xi_1)\|x(\boldsymbol{\xi})-x'(\boldsymbol{\xi})\|]$ and $M(\xi_1')E\|x(\boldsymbol{\xi})-x'(\boldsymbol{\xi})\|$ are both $o_p(1)$, and by convergence in distribution we obtain the lower bound

$$P\left(E\left[\min_{x\in\mathcal{X}^*}\{h(x,\xi_1)-h(x,\xi_1')\}\middle|\xi_1'\right]>\eta\right)$$

in place of (C.4). This completes the proof. □

C.2 Proof of Theorems 4.5.2 and 4.5.3

We need the following result from Van der Vaart (2000):

Theorem C.2.1 (Theorem 11.2 in Van der Vaart (2000)) *Let $\mathcal{L}_n$ be a linear space of random variables with finite second moment that contains the constants. Let T_n be a random variable with projection S_n onto $\mathcal{L}_n$. If*

$$\frac{Var(T_n)}{Var(S_n)}\to 1 \quad as\ n\to\infty$$

then

$$\frac{T_n-ET_n}{sd(T_n)}-\frac{S_n-ES_n}{sd(S_n)}\xrightarrow{p}0 \quad as\ n\to\infty$$

where $sd(\cdot)$ denotes the standard deviation.

For any random variable in the form $T=T(\xi_1,\ldots,\xi_n)$, we also use the notation $\mathring{T}$ to denote the Hajek projection, namely, the projection of T onto the space spanned by $\sum_{i=1}^{n}f_i(\xi_i)$ where f_i's

are any measurable functions. By Van der Vaart (2000), we know that, if $\xi_1, \ldots, \xi_n$ are i.i.d. and T has finite second moment,

$$\mathring{T} = \sum_{i=1}^{n} E[T|\xi_i] - (n-1)ET$$

To proceed, we also define

$$g_{k,c}(\tilde{\xi}_1, \ldots, \tilde{\xi}_c) = E[H_k(\xi_1, \ldots, \xi_k)|\xi_1 = \tilde{\xi}_1, \ldots, \xi_c = \tilde{\xi}_c]$$

as the conditional expectation of H_k given the first c variables. In particular, by our definition before, $g_k(\xi) = g_{k,1}(\xi)$ and $H_k(\xi_1, \ldots, \xi_k) = g_{k,k}(\xi_1, \ldots, \xi_k)$.

We have the following lemma on the estimate of $g_{k,c}(\cdot)$:

Lemma C.2.2 *Suppose Assumption 4.4.1 holds. For $\xi_1, \ldots, \xi_c, \xi'_1, \ldots, \xi'_c \overset{i.i.d.}{\sim} F$, we have*

$$|g_{k,c}(\xi'_1, \ldots, \xi'_c) - E[g_{k,c}(\xi_1, \ldots, \xi_c)]| \leq \frac{1}{k} \sum_{i=1}^{c} E\left[\sup_{x \in \mathcal{X}} |h(x, \xi'_i) - h(x, \xi_i)| \middle| \xi'_i\right]$$

Proof. Let ξ_i, ξ'_i be i.i.d. variables generated from F. Assumption 4.4.1 ensures $EH_k(\xi_1, \ldots, \xi_k)^2 < \infty$ and hence $Var(g_{k,c}(\xi_1, \ldots, \xi_c)) = Var(E[H_k(\xi_1, \ldots, \xi_k)|\xi_1, \ldots, \xi_c]) \leq Var(H_k(\xi_1, \ldots, \xi_k))$ is well-defined and finite. Consider

$$\begin{aligned}
& g_{k,c}(\xi'_1, \ldots, \xi'_c) - E[g_{k,c}(\xi_1, \ldots, \xi_c)] \\
= \; & E[H_k(\xi'_1, \ldots, \xi'_c, \xi_1, \ldots, \xi_k)|\xi'_1, \ldots, \xi'_c] - E[H_k(\xi_1, \ldots, \xi_k)] \\
= \; & E\left[\min_{x \in \mathcal{X}} \left\{ \frac{1}{k} \sum_{i=1}^{c} h(x, \xi'_i) + \frac{1}{k} \sum_{i=c+1}^{k} h(x, \xi_i) \right\} \middle| \xi'_1, \ldots, \xi'_c \right] - E\left[\min_{x \in \mathcal{X}} \frac{1}{k} \sum_{i=1}^{k} h(x, \xi_i)\right] \\
= \; & E\left[\min_{x \in \mathcal{X}} \left\{ \frac{1}{k} \sum_{i=1}^{c} h(x, \xi'_i) + \frac{1}{k} \sum_{i=c+1}^{k} h(x, \xi_i) \right\} - \min_{x \in \mathcal{X}} \frac{1}{k} \sum_{i=1}^{k} h(x, \xi_i) \middle| \xi'_1, \ldots, \xi'_c \right] \\
& \text{since } \xi_i, \xi'_i \text{ are all independent} \\
\leq \; & E\left[\frac{1}{k} \sum_{i=1}^{c} h(x_{\epsilon,k}(\boldsymbol{\xi}), \xi'_i) + \frac{1}{k} \sum_{i=c+1}^{k} h(x_{\epsilon,k}(\boldsymbol{\xi}), \xi_i) - \frac{1}{k} \sum_{i=1}^{k} h(x_{\epsilon,k}(\boldsymbol{\xi}), \xi_i) \middle| \xi'_1, \ldots, \xi'_c \right] + \epsilon \quad \text{(C.5)}
\end{aligned}$$

where $x_{\epsilon,k}(\boldsymbol{\xi})$ is an ϵ-optimal solution of $\min_{x\in\mathcal{X}} \frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i)$. The last inequality follows since

$$\min_{x\in\mathcal{X}}\left\{\frac{1}{k}\sum_{i=1}^{c} h(x,\xi_i') + \frac{1}{k}\sum_{i=c+1}^{k} h(x,\xi_i)\right\} \le \frac{1}{k}\sum_{i=1}^{c} h(x_{\epsilon,k}(\boldsymbol{\xi}),\xi_i') + \frac{1}{k}\sum_{i=c+1}^{k} h(x_{\epsilon,k}(\boldsymbol{\xi}),\xi_i)$$

by the definition of minimization and

$$\min_{x\in\mathcal{X}} \frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i) \ge \frac{1}{k}\sum_{i=1}^{k} h(x_{\epsilon,k}(\boldsymbol{\xi}),\xi_i) - \epsilon$$

by the definition of $x_\epsilon(\boldsymbol{\xi})$. Note that (C.5) is equal to

$$\begin{aligned}
& E\left[\frac{1}{k}\sum_{i=1}^{c}(h(x_{\epsilon,k}(\boldsymbol{\xi}),\xi_i') - h(x_{\epsilon,k}(\boldsymbol{\xi}),\xi_i))\middle|\xi_1',\ldots,\xi_c'\right] + \epsilon \\
\le\ & E\left[\frac{1}{k}\sum_{i=1}^{c}\sup_{x\in\mathcal{X}}|h(x,\xi_i') - h(x,\xi_i)|\middle|\xi_1',\ldots,\xi_c'\right] + \epsilon \qquad \text{(C.6)} \\
=\ & \frac{1}{k}\sum_{i=1}^{c} E\left[\sup_{x\in\mathcal{X}}|h(x,\xi_i') - h(x,\xi_i)|\middle|\xi_i'\right] + \epsilon \qquad \text{(C.7)} \\
& \text{since } \xi_1,\ldots,\xi_c,\xi_1',\ldots,\xi_c' \text{ are independent}
\end{aligned}$$

Similarly,

$$
\begin{aligned}
& E[g_{k,c}(\xi_1,\ldots,\xi_c)] - g_{k,c}(\xi'_1,\ldots,\xi'_c) \\
= \;& E[H_k(\xi_1,\ldots,\xi_k)] - E[H_k(\xi'_1,\ldots,\xi'_c,\xi_1,\ldots,\xi_k)|\xi'_1,\ldots,\xi'_c] \\
= \;& E\left[\min_{x\in\mathcal{X}} \frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i)\right] - E\left[\min_{x\in\mathcal{X}}\left\{\frac{1}{k}\sum_{i=1}^{c} h(x,\xi'_i) + \frac{1}{k}\sum_{i=c+1}^{k} h(x,\xi_i)\right\}\middle|\xi'_1,\ldots,\xi'_c\right] \\
= \;& E\left[\min_{x\in\mathcal{X}} \frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i) - \min_{x\in\mathcal{X}}\left\{\frac{1}{k}\sum_{i=1}^{c} h(x,\xi'_i) + \frac{1}{k}\sum_{i=c+1}^{k} h(x,\xi_i)\right\}\middle|\xi'_1,\ldots,\xi'_c\right] \\
& \text{since } \xi_i, \xi'_i \text{ are all independent} \\
\leq \;& E\left[\frac{1}{k}\sum_{i=1}^{k} h(x_{\epsilon,k}(\boldsymbol{\xi}'),\xi_i) - \frac{1}{k}\sum_{i=1}^{c} h(x_{\epsilon,k}(\boldsymbol{\xi}'),\xi'_i) - \frac{1}{k}\sum_{i=c+1}^{k} h(x_{\epsilon,k}(\boldsymbol{\xi}'),\xi_i)\middle|\xi'_1,\ldots,\xi'_c\right] + \epsilon \\
& \text{where } x_{\epsilon,k}(\boldsymbol{\xi}') \text{ is the } \epsilon\text{-optimal solution of } \min_{x\in\mathcal{X}}\left\{\frac{1}{k}\sum_{i=1}^{c} h(x,\xi'_i) + \frac{1}{k}\sum_{i=c+1}^{k} h(x,\xi_i)\right\} \text{ since} \\
& \min_{x\in\mathcal{X}}\left\{\frac{1}{k}\sum_{i=1}^{c} h(x,\xi'_i) + \frac{1}{k}\sum_{i=c+1}^{k} h(x,\xi_i)\right\} + \epsilon \geq \frac{1}{k}\sum_{i=1}^{c} h(x_{\epsilon,k}(\boldsymbol{\xi}'),\xi'_i) + \frac{1}{k}\sum_{i=c+1}^{k} h(x_{\epsilon,k}(\boldsymbol{\xi}'),\xi_i) \\
& \text{by the definition of } x_{\epsilon,k}(\boldsymbol{\xi}') \text{ and } \min_{x\in\mathcal{X}} \frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i) \leq \frac{1}{k}\sum_{i=1}^{k} h(x_{\epsilon,k}(\boldsymbol{\xi}'),\xi_i) \\
= \;& E\left[\frac{1}{k}\sum_{i=1}^{c}(h(x_{\epsilon,k}(\boldsymbol{\xi}'),\xi_i) - h(x_{\epsilon,k}(\boldsymbol{\xi}'),\xi'_i))\middle|\xi'_1,\ldots,\xi'_c\right] + \epsilon \\
\leq \;& \frac{1}{k}\sum_{i=1}^{c} E\left[\sup_{x\in\mathcal{X}}|h(x,\xi_i) - h(x,\xi'_i)|\middle|\xi'_i\right] + \epsilon \qquad \text{(C.8)}
\end{aligned}
$$

Combining (C.7) and (C.8), and noting that ϵ is arbitrary, we have

$$
|g_{k,c}(\xi'_1,\ldots,\xi'_c) - E[g_{k,c}(\xi_1,\ldots,\xi_c)]| \leq \frac{1}{k}\sum_{i=1}^{c} E\left[\sup_{x\in\mathcal{X}}|h(x,\xi'_i) - h(x,\xi_i)|\middle|\xi'_i\right]
$$

which concludes the lemma. □

We are now ready to prove Theorem 4.5.2: *Proof of Theorem 4.5.2.* By Assumption 4.4.1, we have $EH_k(\xi_1,\ldots,\xi_k)^2 < \infty$ by (4.16) and hence the centered U-statistic $U_{n,k} - W_k$ satisfies

$E(U_{n,k} - W_k)^2 < \infty$. Following Section 12.1 in van der Vaart, the Hajek projection of $U_{n,k} - W_k$ is

$$(U_{n,k} \mathring{-} W_k) = \sum_{i=1}^{n} E[U_{n,k} - W_k|\xi_i] = \sum_{i=1}^{n} \frac{1}{\binom{n}{k}} \sum_{(i_1,\ldots,i_k)\in\mathcal{C}_k} E[H_k(\xi_{i_1},\ldots,\xi_{i_k}) - W_k|\xi_i] \tag{C.9}$$

Note that

$$E[H_k(\xi_{i_1},\ldots,\xi_{i_k}) - W_k|\xi_i] = \begin{cases} E[H_k(\xi_1,\ldots,\xi_k) - W_k|\xi_i] = g_k(\xi_i) - W_k & \text{if } i \in \{i_1,\ldots,i_k\} \\ 0 & \text{otherwise} \end{cases}$$

For each i, the number of $E[H_k(\xi_{i_1},\ldots,\xi_{i_k}) - W_k|\xi_i]$ in which the first case above happens, among all summands in the inner summation in the left hand side of (C.9), is $\binom{n-1}{k-1}$. Therefore, (C.9) is equal to

$$\frac{\binom{n-1}{k-1}}{\binom{n}{k}} \sum_{i=1}^{n} (g_k(\xi_i) - W_k) = \frac{k}{n} \sum_{i=1}^{n} (g_k(\xi_i) - W_k) \tag{C.10}$$

Since $\xi_1,\ldots,\xi_n$ are i.i.d., we have

$$Var((U_{n,k} \mathring{-} W_k)) = \frac{k^2}{n} Var(g_k(\xi)) \tag{C.11}$$

where $\xi \sim F$.

By Theorem C.2.1, if we can prove that

$$\frac{Var(U_{n,k})}{Var((U_{n,k} \mathring{-} W_k))} \to 1 \tag{C.12}$$

and

$$\frac{(U_{n,k} \mathring{-} W_k) - E(U_{n,k} \mathring{-} W_k)}{sd((U_{n,k} \mathring{-} W_k))} \Rightarrow N(0,1) \tag{C.13}$$

Then

$$\begin{aligned} & \frac{U_{n,k} - W_k}{sd(U_{n,k})} \\ = & \left(\frac{U_{n,k} - W_k}{sd(U_{n,k})} - \frac{(U_{n,k} \mathring{-} W_k) - E(U_{n,k} \mathring{-} W_k)}{sd((U_{n,k} \mathring{-} W_k))} \right) + \frac{(U_{n,k} \mathring{-} W_k) - E(U_{n,k} \mathring{-} W_k)}{sd((U_{n,k} \mathring{-} W_k))} \Rightarrow N(0,1) \end{aligned}$$

by Slutsky's Theorem. Furthermore, by (C.12) and Slutsky's Theorem again, we have

$$\frac{U_{n,k} - W_k}{sd((U_{n,k} \overset{\circ}{-} W_k))} = \frac{U_{n,k} - W_k}{sd(U_{n,k})} \frac{sd(U_{n,k})}{sd((U_{n,k} \overset{\circ}{-} W_k))} \Rightarrow N(0,1)$$

Note that, by (C.11),

$$\frac{U_{n,k} - W_k}{sd((U_{n,k} \overset{\circ}{-} W_k))} = \frac{\sqrt{n}(U_{n,k} - W_k)}{k\sqrt{Var(g_k(\xi))}}$$

and hence we conclude the theorem.

By (C.10) and (C.11), the left hand side of (C.13) can be written as

$$\frac{(k/n)\sum_{i=1}^n (g_k(\xi_i) - W_k)}{(k/\sqrt{n})sd(g_k(\xi))} = \frac{\sqrt{n}((1/n)\sum_{i=1}^n (g_k(\xi_i) - W_k)}{sd(g_k(\xi))}$$

Thus (C.13) is equivalent to

$$\frac{\sqrt{n}((1/n)\sum_{i=1}^n (g_k(\xi_i) - W_k)}{sd(g_k(\xi))} \Rightarrow N(0,1) \tag{C.14}$$

The rest of the proof focuses on showing (C.12) and (C.14).

Proof of (C.12). Consider

$$\begin{aligned} Var(U_{n,k}) &= \frac{1}{\binom{n}{k}^2} \sum_{(i_1,\ldots,i_k),(i'_1,\ldots,i'_k)\in\mathcal{C}_k} Cov(H_k(\xi_{i_1},\ldots,\xi_{i_k}), H_k(\xi_{i'_1},\ldots,\xi_{i'_k})) \\ &= \frac{1}{\binom{n}{k}^2} \sum_{c=1}^k \binom{n}{k}\binom{k}{c}\binom{n-k}{k-c} Var(g_{k,c}(\xi_1,\ldots,\xi_c)) \qquad \text{(C.15)} \\ &= \sum_{c=1}^k \frac{\binom{k}{c}\binom{n-k}{k-c}}{\binom{n}{k}} Var(g_{k,c}(\xi_1,\ldots,\xi_c)) \qquad \text{(C.16)} \end{aligned}$$

where the second equality follows by counting the number of combinations of $(i_1,\ldots,i_k)$ and $(i'_1,\ldots,i'_k)$ in the summation that have c overlapping indices. For each c, this number follows by first picking k out of n indices from $\{1,\ldots,n\}$ to place in $(i_1,\ldots,i_k)$, then choosing c from these k numbers to place in $(i'_1,\ldots,i'_k)$ and $k-c$ from the remaining $n-k$ numbers to place in

the remaining spots in $(i'_1, \ldots, i'_k)$. Note also that if $(i_1, \ldots, i_k)$ and $(i'_1, \ldots, i'_k)$ have $c \geq 1$ overlapping indices, say by relabeling and the symmetry of H_k we write the indices as $(i_1, \ldots, i_k)$ and $(i_1, \ldots, i_c, i'_{c+1}, \ldots, i'_k)$, then

$$\begin{aligned}
& Cov(H_k(\xi_{i_1}, \ldots, \xi_{i_k}), H_k(\xi_{i_1}, \ldots, \xi_{i_c}, \xi_{i'_{c+1}}, \ldots, \xi_{i'_k})) \\
= \; & Cov(E[H_k(\xi_{i_1}, \ldots, \xi_{i_k})|\xi_1, \ldots, \xi_c], E[H_k(\xi_{i_1}, \ldots, \xi_{i_c}, \xi_{i'_{c+1}}, \ldots, \xi_{i'_k})|\xi_1, \ldots, \xi_c]) \\
& + E[Cov(H_k(\xi_{i_1}, \ldots, \xi_{i_k}), H_k(\xi_{i_1}, \ldots, \xi_{i_c}, \xi_{i'_{c+1}}, \ldots, \xi_{i'_k})|\xi_1, \ldots, \xi_c)] \\
= \; & Var(g_{k,c}(\xi_1, \ldots, \xi_c))
\end{aligned}$$

since $\xi_{c+1}, \ldots, \xi_k$ and $\xi'_{c+1}, \ldots, \xi'_k$ are independent. Finally, if $(i_1, \ldots, i_k)$ and $(i'_1, \ldots, i'_k)$ have no overlapping index then $Cov(H_k(\xi_{i_1}, \ldots, \xi_{i_k}), H_k(\xi_{i'_1}, \ldots, \xi_{i'_k})) = 0$. Therefore the equality in (C.15) holds.

On the other hand, by (C.11), we have $Var((U_{n,k} \mathring{-} W_k)) = \frac{k^2}{n} Var(g_k(\xi))$. Also, by Lemma 4.5.1, $Var(g_k(\xi)) > 0$. Combining these with (C.16) gives

$$\begin{aligned}
\frac{Var(U_{n,k})}{Var((U_{n,k} \mathring{-} W_k))} &= \frac{n \sum_{c=1}^{k} \frac{\binom{k}{c}\binom{n-k}{k-c}}{\binom{n}{k}} Var(g_{k,c}(\xi_1, \ldots, \xi_c))}{k^2 Var(g_k(\xi))} \\
&= \frac{n}{k^2} \cdot \frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}} + \frac{n \sum_{c=2}^{k} \frac{\binom{k}{c}\binom{n-k}{k-c}}{\binom{n}{k}} Var(g_{k,c}(\xi_1, \ldots, \xi_c))}{k^2 Var(g_k(\xi))} \qquad \text{(C.17)}
\end{aligned}$$

Consider the first term in (C.17). We have

$$\begin{aligned}
& \frac{n}{k^2} \cdot \frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}} \\
= \; & \frac{n}{k^2} \frac{k \cdot (n-k)(n-k-1) \cdots (n-2k+2)/(k-1)!}{n(n-1) \cdots (n-k+1)/k!} \\
= \; & \frac{n-k}{n-1} \cdot \frac{n-k-1}{n-2} \cdots \frac{n-2k+2}{n-k+1} \\
= \; & \left(1 - \frac{k-1}{n-1}\right)\left(1 - \frac{k-1}{n-2}\right) \cdots \left(1 - \frac{k-1}{n-k+1}\right) \qquad \text{(C.18)}
\end{aligned}$$

For $k = o(\sqrt{n})$,

$$\begin{aligned}\left|\sum_{j=n-k+1}^{n-1} \log\left(1 - \frac{k-1}{j}\right)\right| &\leq C \sum_{j=n-k+1}^{n-1} \frac{k-1}{j} \\ &\leq C \int_{n-k}^{n-1} \frac{k-1}{u} du \\ &= C(k-1) \log \frac{n-1}{n-k} \\ &= -C(k-1) \log\left(1 - \frac{k-1}{n-1}\right) \\ &\leq \tilde{C}(k-1)\frac{k-1}{n-1} \\ &= o(1)\end{aligned}$$

where $C, \tilde{C} > 0$ are some constants. Therefore, from (C.18), we get

$$\frac{n}{k^2} \cdot \frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}} \to 1 \tag{C.19}$$

as $n \to \infty$.

Now consider the second term in (C.17). By Lemma C.2.2, for $c \geq 1$,

$$\begin{aligned}Var(g_{k,c}(\xi_1, \ldots, \xi_c)) &\leq E\left(\frac{1}{k}\sum_{i=1}^{c} E\left[\sup_{x \in \mathcal{X}} |h(x, \xi_i') - h(x, \xi_i)| \middle| \xi_i'\right]\right)^2 \\ &\leq \frac{1}{k^2}\left(\sum_{i=1}^{c} \sqrt{E\left(\sup_{x \in \mathcal{X}} |h(x, \xi_i') - h(x, \xi_i)|\right)^2}\right)^2 \text{ by the Minkowski inequality} \\ &\leq \frac{c^2 M}{k^2} \qquad (C.20)\end{aligned}$$

for some $M > 0$ by Assumption 4.5.1.

Note also that

$$\frac{\binom{k}{c}\binom{n-k}{k-c}}{\binom{n}{k}}$$

is the probability mass at c of a hypergeometric variable with parameters (n, k, k). Note that such a variable takes domain $\{\max(2k - n, 0), \ldots, k\}$, which equals $\{0, \ldots, k\}$ for n sufficiently large

since $k = o(\sqrt{n})$. This hypergeometric variable has second moment equal to

$$\frac{k^2(n-k)^2}{n^2(n-1)} + \frac{k^4}{n^2} \tag{C.21}$$

Thus, for n large enough, and using Lemma 4.5.1, we have

$$\begin{aligned}
& \frac{n\sum_{c=2}^{k} \frac{\binom{k}{c}\binom{n-k}{k-c}}{\binom{n}{k}} Var(g_{k,c}(\xi_1,\ldots,\xi_c))}{k^2 Var(g_k(\xi))} \\
\leq\ & \frac{n\sum_{c=2}^{k} \frac{\binom{k}{c}\binom{n-k}{k-c}}{\binom{n}{k}} \frac{c^2 M}{k^2}}{\epsilon} \text{ for some constant } \epsilon > 0 \\
=\ & n\left(\frac{k^2(n-k)^2}{n^2(n-1)} + \frac{k^4}{n^2} - \frac{k^2}{n}(1+o(1))\right)\frac{M}{\epsilon k^2} \qquad \text{(C.22)} \\
& \text{by using (C.21) and } \frac{\binom{k}{1}\binom{n-k}{k-1}}{\binom{n}{k}} = \frac{k^2}{n}(1+o(1)) \text{ that we have proven using (C.19)} \\
=\ & \left(\frac{(n-k)^2}{n(n-1)} + \frac{k^2}{n} - (1+o(1))\right)\frac{M}{\epsilon} \\
=\ & o(1) \qquad \text{(C.23)}
\end{aligned}$$

since $k = o(\sqrt{n})$.

Combining (C.19) and (C.22) into (C.17), we get

$$\frac{Var(U_{n,k})}{Var((U_{n,k} \overset{\circ}{-} W_k))} \to 1.$$

Proof of (C.14). By Lemma C.2.2, denoting $\xi, \xi' \overset{i.i.d.}{\sim} F$, we have

$$\begin{aligned}
& E|g_k(\xi_i) - W_k|^{2+\delta} \\
\leq\ & E\left(\frac{1}{k}E\left[\sup_{x\in\mathcal{X}}|h(x,\xi') - h(x,\xi)|\middle|\xi'\right]\right)^{2+\delta} \leq \frac{1}{k^{2+\delta}}E\sup_{x\in\mathcal{X}}|h(x,\xi') - h(x,\xi)|^{2+\delta} \leq \frac{\tilde{M}}{k^{2+\delta}}
\end{aligned}$$

for some $\tilde{M} > 0$ by Assumption 4.5.1. Moreover, by Lemma 4.5.1 we have $Var(g_k(\cdot)) \geq \epsilon/k^2$ for

some $\epsilon > 0$ for k sufficiently large. Hence

$$\frac{nE|g_k(\xi) - W_k|^{2+\delta}}{(nVar(g_k(\xi)))^{1+\delta/2}} \leq \frac{n\tilde{M}/k^{2+\delta}}{(n\epsilon/k^2)^{1+\delta/2}} = \frac{\tilde{M}}{n^{\delta/2}\epsilon^{1+\delta/2}} \to 0$$

as $n \to \infty$. The Lyapunov condition then implies the central limit theorem in (C.14). □

Proof of Theorem 4.5.3. Let $c(n,k,s)$ count the number of mappings $\phi : \{1,2,\ldots,k\} \to \{1,2,\ldots,n\}$ such that $|\phi(\{1,2,\ldots,k\})| = s$, or equivalently, count the number of $\xi_{i_1},\ldots,\xi_{i_k}$ such that $i_1,\ldots,i_k$ covers s distinct indices, and let $A_{n,s}$ be the average of all $H_k(\xi_{i_1},\ldots,\xi_{i_k})$ with s distinct indices. In particular, $A_{n,k} = U_{n,k}$. The V-statistic can be expressed for a fixed $l \geq 0$ as

$$n^k V_{n,k} = \sum_{s=k-l}^{k} c(n,k,s)A_{n,s} + \big(n^k - \sum_{s=k-l}^{k} c(n,k,s)\big)R_{n,l}$$

where $R_{n,l}$ is the average of all $H_k(\xi_{i_1},\ldots,\xi_{i_k})$ with at most $k-l-1$ distinct indices. We have

$$\begin{aligned} n^k(U_{n,k} - V_{n,k}) &= n^k U_{n,k} - \sum_{s=k-l}^{k} c(n,k,s)(U_{n,k} + A_{n,s} - U_{n,k}) - \big(n^k - \sum_{s=k-l}^{k} c(n,k,s)\big)R_{n,l} \\ &= \big(n^k - \sum_{s=k-l}^{k} c(n,k,s)\big)(U_{n,k} - R_{n,l}) - \sum_{s=k-l}^{k-1} c(n,k,s)(A_{n,s} - U_{n,k}) \\ &= \big(\sum_{s=1}^{k-l-1} c(n,k,s)\big)(U_{n,k} - R_{n,l}) - \sum_{s=k-l}^{k-1} c(n,k,s)(A_{n,s} - U_{n,k}). \end{aligned} \tag{C.24}$$

We want to show that the two terms in (C.24) are both $o_p(n^{k-1/2})$ so that the desired conclusion follows by Slutsky's theorem. To this end, we let

$$l = \lfloor \frac{1}{2(1-2\gamma)} \rfloor \tag{C.25}$$

the reason for which shall be clear later.

To bound the first term in (C.24), note that $c(n,k,s)$ can be written as

$$c(n,k,s) = S(k,s)n(n-1)\cdots(n-s+1)$$

where $S(k,s)$ is the Stirling number of the second kind with parameters k, s, which is the number of partitions of a set of size k into s non-empty subsets. It's shown in Rennie and Dobson (1969) that for $k \geq 2$ and $1 \leq s \leq k-1$

$$S(k,s) \leq \frac{1}{2}\binom{k}{s} s^{k-s}. \tag{C.26}$$

Hence

$$\sum_{s=1}^{k-l-1} c(n,k,s) \leq \frac{1}{2} \sum_{s=1}^{k-l-1} \binom{k}{s} s^{k-s} n^s.$$

Note that the ratio between two neighboring $\binom{k}{s} s^{k-s} n^s$ is

$$\binom{k}{s-1}(s-1)^{k-s+1} n^{s-1} \Big/ \binom{k}{s} s^{k-s} n^s = \frac{(s-1)^{k-s+1}}{(k-s+1)s^{k-s-1}n} \leq \frac{s^2}{n} \leq \frac{k^2}{n} = o(1),$$

therefore

$$\begin{aligned} \sum_{s=1}^{k-l-1} c(n,k,s) &\leq \frac{1}{2}\big(1 + \sum_{s=1}^{k-l-2} (\frac{k^2}{n})^s\big)\binom{k}{l+1}(k-l-1)^{l+1} n^{k-l-1} \\ &\leq \frac{1}{2(1-k^2/n)}\binom{k}{l+1}(k-l-1)^{l+1} n^{k-l-1} = O(k^{2l+2} n^{k-l-1}) = O\big((\frac{k^2}{n})^{l+1} n^k\big). \end{aligned}$$

For the particular choice of l shown in (C.25), the above bound is $o(n^{k-1/2})$. Since both $U_{n,k}$ and $R_{n,l}$ are $O_p(1)$ by Assumption 4.4.1, the first term in (C.24) is $O_p(n^{k-1})$.

For the second term in (C.24), it suffices to show that for each $k-l \leq s \leq k-1$ it holds $c(n,k,s)(A_{n,s} - U_{n,k}) = o_p(n^{k-1/2})$ since there are only l of them. Since l is now viewed as a constant, from the upper bound (C.26) for $s \geq k-l$ it follows that $S(k,s) = O(k^{2(k-s)})$, resulting in $c(n,k,s) = O(k^{2(k-s)} n^s)$. If we can argue that $A_{n,s} - U_{n,k} = O_p(k^{-1})$, then each summand can be bounded as

$$O_p(k^{2(k-s)-1} n^s) = O_p(n^{2\gamma(k-s)-\gamma+s}) = O_p(n^{k+\gamma-1})$$

where the last equality holds because $\gamma < 1/2$ hence $2\gamma(k-s) - \gamma + s$ increases in s. This implies an upper bound of order $o_p(n^{k-1/2})$ again because $\gamma < 1/2$. Now we show $A_{n,s} - U_{n,k} = O_p(k^{-1})$ by a coupling argument. The value of $A_{n,s}$ can be computed from the same resamples $\xi_{i_1}, \ldots, \xi_{i_k}$ (with k distinct data points) used to compute $U_{n,k}$, by removing $k-s$ of them and fill in with those

remaining in the resample. To be specific, we use $I_k = (I(1), \ldots, I(k))$ to represent a sequence of length k where $I(j) \in \{1, \ldots, n\}$ for each $j \leq k$, define $|I_k|$ to be the number of distinct indices in I_k. For convenience we denote by $I_k(j_1 : j_2) = (I_k(j_1), \ldots, I_k(j_2))$ the sub-sequence for $1 \leq j_1 \leq j_2 \leq k$ and $\boldsymbol{\xi}_{I_k} = (\xi_{I_k(1)}, \ldots, \xi_{I_k(k)})$. Then

$$A_{n,s} = \frac{(n-k)!}{n!} \sum_{|I_k|=k} \frac{1}{s^{k-s}} \sum_{|I'_k|=s, I'_k(1:s)=I_k(1:s)} H_k(\boldsymbol{\xi}_{I'_k}).$$

This leads to

$$\begin{aligned}
|A_{n,s} - U_{n,k}| &\leq \frac{(n-k)!}{n!} \sum_{|I_k|=k} \frac{1}{s^{k-s}} \sum_{|I'_k|=s, I'_k(1:s)=I_k(1:s)} \left| H_k(\boldsymbol{\xi}_{I'_k}) - H_k(\boldsymbol{\xi}_{I_k}) \right| \\
&\leq \frac{(n-k)!}{n!} \sum_{|I_k|=k} \frac{1}{s^{k-s}} \sum_{|I'_k|=s, I'_k(1:s)=I_k(1:s)} \sup_{x \in \mathcal{X}} \left| \sum_{j=s+1}^{k} \frac{1}{k} (h(x, \xi_{I'_k(j)}) - h(x, \xi_{I_k(j)})) \right| \\
&\leq \frac{(n-k)!}{n!} \sum_{|I_k|=k} \frac{1}{s^{k-s}} \sum_{|I'_k|=s, I'_k(1:s)=I_k(1:s)} \sum_{j=s+1}^{k} \frac{1}{k} \sup_{x \in \mathcal{X}} \left| h(x, \xi_{I'_k(j)}) - h(x, \xi_{I_k(j)}) \right| \\
&\leq \frac{1}{k} \sum_{j=s+1}^{k} \frac{(n-k)!}{n! s^{k-s}} \sum_{|I_k|=k} \sum_{|I'_k|=s, I'_k(1:s)=I_k(1:s)} \sup_{x \in \mathcal{X}} \left| h(x, \xi_{I'_k(j)}) - h(x, \xi_{I_k(j)}) \right| \\
&= \frac{k-s}{k} \frac{2}{n(n-1)} \sum_{1 \leq i_1 < i_2 \leq n} \sup_{x \in \mathcal{X}} |h(x, \xi_{i_1}) - h(x, \xi_{i_2})|
\end{aligned}$$

where the last equality is because $I'_k(j)$ and $I_k(j)$ are distinct indices and the gross sum over I_k, I'_k puts equal weight on each pair (i_1, i_2). Due to Assumption 4.5.1, we have

$$E[|A_{n,s} - U_{n,k}|] \leq \frac{k-s}{k} E[\sup_{x \in \mathcal{X}} |h(x, \xi) - h(x, \xi')|] = O(\frac{l}{k}) = O(\frac{1}{k}).$$

This completes the proof. □

C.3 Proof of Theorems 4.5.4 and 4.5.5

The proof relies on the following ANOVA decomposition of a symmetric statistic:

Lemma C.3.1 (Adapted from Efron and Stein (1981)) *For any symmetric function T mapping from $\Xi^k \to \mathbb{R}$ and i.i.d. random elements $\xi_1, \ldots, \xi_k \in \Xi$ such that $Var[T(\xi_1, \ldots, \xi_k)] < \infty$, there exist functions $T_1, \ldots, T_k$ such that*

$$T(\xi_1, \ldots, \xi_k) = E[T] + \sum_{i=1}^{k} T_1(\xi_1) + \sum_{i_1 < i_2} T_2(\xi_{i_1}, \xi_{i_2}) + \cdots + T_k(\xi_1, \ldots, \xi_k).$$

Moreover, all the $2^k - 1$ random variables on the right hand side have mean zero and are mutually uncorrelated.

Note that $T_1(x)$ must be $E[T(\xi_1, \ldots, \xi_k)|\xi_1 = x] - ET$ by the property in the lemma, and the total variance of a symmetric statistic $T(\xi_1, \ldots, \xi_k)$ can be decomposed as $Var(T) = \sum_{s=1}^{k} \binom{k}{s} V_s$, where $V_s := Var(T_s(\xi_1, \ldots, \xi_s))$. The Hajek projection is defined as $\mathring{T} := ET + \sum_{i=1}^{k} T_1(\xi_1)$, i.e. the first order effect in the ANOVA decomposition. In particular, the Hajek projections of the symmetric kernel H_k and the symmetric statistic $U_{n,k}$ are

$$\mathring{H}_k = W_k + \sum_{i=1}^{k} (g_k(\xi_i) - W_k)$$
$$\mathring{U}_{n,k} = W_k + \frac{k}{n} \sum_{i=1}^{n} (g_k(\xi_i) - W_k).$$

As discussed in Section 4.5, we will use the ANOVA decomposition (Efron and Stein (1981)) of the symmetric kernel H_k to allow for a larger resample size k in obtaining Theorem 4.5.4. We have the following variance bound from Wager and Athey (2018) in analyzing random forests:

Lemma C.3.2 ***(Adapted from Lemma 7 of Wager and Athey (2018))*** *Under Assumption 4.4.1, for any $k \leq n$ it holds*

$$E(U_{n,k} - \mathring{U}_{n,k})^2 \leq \frac{k^2}{n^2} E(H_k - \mathring{H}_k)^2.$$

Proof. Wager and Athey (2018) prove this bound in the context of random forests where H_k is a regression tree and $U_{n,k}$ is the random forest obtained from aggregating the resampled trees (without replacement). Although the context they focus on is different from ours, their proof works

for general symmetric kernels and U-statistics including the SAA values considered in this paper. Note that in Lemma 7 of Wager and Athey (2018) the right hand side is the total variance $Var(H_k)$ instead of $E(H_k - \mathring{H}_k)^2$, however this comes from upper bounding $E(H_k - \mathring{H}_k)^2$ by $Var(H_k)$ in their proof so the bound with $E(H_k - \mathring{H}_k)^2$ remains true. □

This allows us to derive a CLT under an additional assumption on k:

Theorem C.3.3 *Under Assumptions 4.4.1, 4.5.1 and 4.5.5, if the resample size k is chosen such that*

$$k^2 E(H_k - \mathring{H}_k)^2 = o(n) \tag{C.27}$$

then

$$\frac{\sqrt{n}(U_{n,k} - W_k)}{k\sqrt{Var(g_k(\xi))}} \Rightarrow N(0,1)$$

where $N(0,1)$ is the standard normal.

Proof. According to the proof of Theorem 4.5.2, we only need to show $E(U_{n,k} - \mathring{U}_{n,k})^2 / Var(U_{n,k}) \to 0$, or equivalently $E(U_{n,k} - \mathring{U}_{n,k})^2 / Var(\mathring{U}_{n,k}) \to 0$. Under the choice of k we have $E(U_{n,k} - \mathring{U}_{n,k})^2 = o(1/n)$ due to Proposition C.3.2, whereas $Var(\mathring{U}_{n,k}) = k^2 Var(g_k(\xi))/n \geq \epsilon/n$ for k large enough. This completes the proof. □

We state an upper bound for the left hand side of (C.27) in terms of the maximum deviation of the cost function from its mean.

Lemma C.3.4 *We have $|H_k - Z^*| \leq \sup_{x\in\mathcal{X}} \left|\frac{1}{k}\sum_{i=1}^k h(x,\xi_i) - Z(x)\right|$, hence*

$$E(H_k - \mathring{H}_k)^2 \leq Var(H_k) \leq E(H_k - Z^*)^2 \leq E\big[\sup_{x\in\mathcal{X}} \Big|\frac{1}{k}\sum_{i=1}^k h(x,\xi_i) - Z(x)\Big|^2\big].$$

Proof. Let x^* be an optimal solution of the original optimization (4.1), and x_k^* be an optimal solution of the SAA formed by $\xi_1, \ldots, \xi_k$. If $H_k \leq Z^*$, since $Z(x_k^*) \geq Z^*$, we have $|H_k - Z^*| \leq |H_k - Z(x_k^*)| \leq \sup_{x\in\mathcal{X}} \left|\frac{1}{k}\sum_{i=1}^k h(x,\xi_i) - Z(x)\right|$. Otherwise, if $H_k > Z^*$, then obviously $Z^* < H_k \leq \frac{1}{k}\sum_{i=1}^k h(x^*,\xi_i)$, hence again $|H_k - Z^*| \leq |\frac{1}{k}\sum_{i=1}^k h(x^*,\xi_i) - Z(x^*)| \leq \sup_{x\in\mathcal{X}} \left|\frac{1}{k}\sum_{i=1}^k h(x,\xi_i) - Z(x)\right|$. This proves the first inequality. For the second part of the lemma, the inequality $E(H_k -$

$\mathring{H}_k)^2 \leq Var(H_k)$ follows from the projection property of Hajek projection and the other two are obvious. □

To proceed, we need to introduce concepts in empirical processes and some notations. Denote by

$$\mathcal{F} := \{h(x, \cdot) - Z(x) : x \in \mathcal{X}\}$$

the family of centered cost functions indexed by the decision $x \in \mathcal{X}$. Note that for centered functions the Lipschitz condition holds with a slightly larger constant than $M(\xi)$

$$|h(x_1, \xi) - Z(x_1) - (h(x_2, \xi) - Z(x_2))| \leq (M(\xi) + EM(\xi))\|x_1 - x_2\|.$$

For a vector $x \in R^d$, let $\|x\|$ be its L_2 norm, and for a random variable X we define $\|X\|_p := (E\,|X|^p)^{1/p}$ for $p \geq 1$. We equip the function space $\mathcal{F}$ defined above with the norm $\|\cdot\|_2$. We denote by $N(\epsilon, \mathcal{X}, \|\cdot\|)$ the covering number, with ball size ϵ, of the decision space, and by $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_2)$ the bracketing number, with bracket size ϵ, of the function space $\mathcal{F}$.

We need a few results adapted from Van der Vaart and Wellner (1996). The first result connects the complexity of the function space $\mathcal{F}$ to that of the decision space $\mathcal{X}$:

Lemma C.3.5 (Adapted from Theorem 2.7.11 in Van der Vaart and Wellner (1996)) *Suppose Assumption 4.2.1 holds and the decision space $\mathcal{X}$ is compact, then for any $\epsilon > 0$*

$$N_{[]}(4\epsilon\|M(\xi)\|_2, \mathcal{F}, \|\cdot\|_2) \leq N(\epsilon, \mathcal{X}, \|\cdot\|).$$

The second result gives an upper bound of the covering number of the decision space $\mathcal{X}$, hence an upper bound of the bracketing number of $\mathcal{F}$ because of the first result.

Lemma C.3.6 *Let $D_{\mathcal{X}}$ be the diameter of the decision space $\mathcal{X}$ with respect to the L_2 norm $\|\cdot\|$, then $N(\epsilon, \mathcal{X}, \|\cdot\|) \leq \left(3D_{\mathcal{X}}/\epsilon\right)^d$ for all $\epsilon \leq D_{\mathcal{X}}$.*

Proof. Problem 6 in Section 2.1 of Van der Vaart and Wellner (1996) states that the ϵ-packing number of a Euclidean ball of radius R in $\mathbb{R}^d$ is bounded above by $\left(3R/\epsilon\right)^d$, and the lemma follows

from the fact that the covering number is always no more than the packing number and that $\mathcal{X}$ can be contained in a Euclidean ball of radius $D_{\mathcal{X}}$. □

The third result concerns the first order moment of the maximum deviation.

Lemma C.3.7 (Adapted from Theorem 2.14.2 of Van der Vaart and Wellner (1996)) *Let $\tilde{h}(\xi) = \sup_{x\in\mathcal{X}} |h(x,\xi) - Z(x)|$. We have for all k*

$$\sqrt{k}E\Big[\sup_{x\in\mathcal{X}} \big|\frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i) - Z(x)\big|\Big] \leq C\|\tilde{h}(\xi)\|_2 \int_0^1 \sqrt{1+\log N_{[]}(\epsilon\|\tilde{h}(\xi)\|_2, \mathcal{F}, \|\cdot\|_2)}d\epsilon$$

where C is a universal constant.

We also need the following result that translates an upper bound of the first order moment to one for higher order moments:

Lemma C.3.8 (Adapted from Theorem 2.14.5 of Van der Vaart and Wellner (1996)) *For any $p \geq 2$ it holds*

$$\sqrt{k}\Big(E\Big[\sup_{x\in\mathcal{X}} \big|\frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i) - Z(x)\big|^p\Big]\Big)^{\frac{1}{p}} \leq C\Big(\sqrt{k}E\Big[\sup_{x\in\mathcal{X}} \big|\frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i) - Z(x)\big|\Big] + k^{-\frac{1}{2}+\frac{1}{p}}\|\tilde{h}(\xi)\|_p\Big)$$

where C is a constant depending only on p, and $\tilde{h}$ is the same as in Lemma C.3.7.

Now we turn to the problem of further bounding the upper bound in Lemma C.3.4, which can be viewed as the maximum deviation of the empirical process generated by the cost function. Specifically, we show that this can be controlled at the canonical rate $1/\sqrt{k}$ in the case of Lipschitz continuous cost function. We have:

Theorem C.3.9 *Suppose Assumptions 4.2.1 and 4.4.1 hold, and that the decision space $\mathcal{X}$ is compact, then we have*

$$E\Big[\sup_{x\in\mathcal{X}} \big|\frac{1}{k}\sum_{i=1}^{k} h(x,\xi_i) - Z(x)\big|^2\Big] = O\big(\frac{1}{k}\big).$$

Proof. First we conclude the following upper bound of the maximum deviation

$$
\begin{aligned}
&\sqrt{k}E\Big[\sup_{x\in\mathcal{X}}\big|\frac{1}{k}\sum_{i=1}^{k}h(x,\xi_i)-Z(x)\big|\Big]\\
\leq& C\|\tilde{h}(\xi)\|_2\int_0^1\sqrt{1+\log N\big(\frac{\epsilon\|\tilde{h}(\xi)\|_2}{4\|M(\xi)\|_2},\mathcal{X},\|\cdot\|\big)}d\epsilon \quad \text{by Lemmas C.3.7 and C.3.5}\\
\leq& C\|\tilde{h}(\xi)\|_2\Big(1+\int_0^1\sqrt{\log N\big(\frac{\epsilon\|\tilde{h}(\xi)\|_2}{4\|M(\xi)\|_2},\mathcal{X},\|\cdot\|\big)}d\epsilon\Big) \quad \text{since } \sqrt{a+b}\leq\sqrt{a}+\sqrt{b}\\
\leq& C\|\tilde{h}(\xi)\|_2\Big(1+\int_0^{\frac{4D_{\mathcal{X}}\|M(\xi)\|_2}{\|\tilde{h}(\xi)\|_2}\wedge 1}\sqrt{d\log\frac{12D_{\mathcal{X}}\|M(\xi)\|_2}{\epsilon\|\tilde{h}(\xi)\|_2}}d\epsilon\Big)\\
&\quad \text{by Lemma C.3.6 and } N(\epsilon,\mathcal{X},\|\cdot\|)=1 \text{ for } \epsilon\geq D_{\mathcal{X}}\\
=& C\|\tilde{h}(\xi)\|_2+12CD_{\mathcal{X}}\|M(\xi)\|_2\int_0^{\frac{1}{3}\wedge\frac{\|\tilde{h}(\xi)\|_2}{12D_{\mathcal{X}}\|M(\xi)\|_2}}\sqrt{d\log\frac{1}{\epsilon}}d\epsilon\\
\leq& C'\Big(\|\tilde{h}(\xi)\|_2+\sqrt{d\log\big(3\vee\frac{12D_{\mathcal{X}}\|M(\xi)\|_2}{\|\tilde{h}(\xi)\|_2}\big)}(4D_{\mathcal{X}}\|M(\xi)\|_2\wedge\|\tilde{h}(\xi)\|_2)\Big)<\infty
\end{aligned}
\tag{C.28}
$$

where C' is another universal constant. Then we apply the Lemma C.3.8 with $p=2$ to get

$$
kE\Big[\sup_{x\in\mathcal{X}}\big|\frac{1}{k}\sum_{i=1}^{k}h(x,\xi_i)-Z(x)\big|^2\Big]\leq C^2(\sqrt{k}E\Big[\sup_{x\in\mathcal{X}}\big|\frac{1}{k}\sum_{i=1}^{k}h(x,\xi_i)-Z(x)\big|\Big]+\|\tilde{h}(\xi)\|_2)^2<\infty
$$

which concludes Theorem C.3.9. □

With all these preparations, Theorem 4.5.4 can be readily proved: *Proof of Theorem 4.5.4.* By Lemma C.3.4 and Theorem C.3.9 we have $E(H_k-\mathring{H}_k)^2=O(1/k)$ hence $k^2E(H_k-\mathring{H}_k)^2=O(k)$, which is $o(n)$ when $k=o(n)$. The CLT then follows from Theorem C.3.3. □

We now prove Theorem 4.5.5: *Proof of Theorem 4.5.5.* If we show $Var(H_k)/Var(\mathring{H}_k)\to 1$ as $k\to\infty$, then the conclusion follows from

$$
E(H_k-\mathring{H}_k)^2=Var(H_k)-Var(\mathring{H}_k)=o(Var(\mathring{H}_k))=o(kVar(g_k(\xi)))=o\big(\frac{1}{k}\big)
$$

and Theorem C.3.3, where the last equality is due to Proposition 4.6.1. Recall that $EH_k=E\mathring{H}_k=$

W_k. In fact we will show the stronger results

$$kE(H_k - W_k)^2 \to Var(h(x^*, \xi)) \tag{C.29}$$

$$kE(\mathring{H}_k - W_k)^2 = k^2 Var(g_k(\xi)) \to Var(h(x^*, \xi)) \tag{C.30}$$

where x^* is the unique optimal solution. The way we prove these two moment convergence results is to first show that the left hand side weakly converges to some variable that has the desired variance and then use uniform integrability to conclude convergence in moments.

We first prove the $\leq$ direction of (C.29). Under the depicted conditions, Theorem 4.2.1 entails that $\sqrt{k}(H_k - Z^*) \Rightarrow N(0, Var(h(x^*, \xi)))$ on one hand. On the other hand, from Lemma C.3.4 we have $|H_k - Z^*|^{2+\delta} \leq \sup_{x\in\mathcal{X}} \left|\frac{1}{k}\sum_{i=1}^k h(x, \xi_i) - Z(x)\right|^{2+\delta}$ and Lemma C.3.8 with $p = 2 + \delta$ implies

$$\begin{aligned}
&E\Big[k^{1+\frac{\delta}{2}} \sup_{x\in\mathcal{X}} \left|\frac{1}{k}\sum_{i=1}^k h(x, \xi_i) - Z(x)\right|^{2+\delta}\Big] \\
\leq &C\Big(\sqrt{k}E\Big[\sup_{x\in\mathcal{X}} \left|\frac{1}{k}\sum_{i=1}^k h(x, \xi_i) - Z(x)\right|\Big] + k^{-\frac{\delta}{2(2+\delta)}} \|\tilde{h}(\xi)\|_{2+\delta}\Big)^{2+\delta}.
\end{aligned}$$

Note that the first term on the right hand side is bounded because of (C.28). Let ξ, ξ' be i.i.d. copies of the uncertain variable, then the second term

$$\begin{aligned}
\|\tilde{h}(\xi)\|_{2+\delta}^{2+\delta} &= E \sup_{x\in\mathcal{X}} |h(x, \xi) - Z(x)|^{2+\delta} \\
&\leq E_\xi \sup_{x\in\mathcal{X}} E_{\xi'} \left|h(x, \xi) - h(x, \xi')\right|^{2+\delta} \text{ by Jensen's inequality} \\
&\leq E_\xi E_{\xi'} \sup_{x\in\mathcal{X}} \left|h(x, \xi) - h(x, \xi')\right|^{2+\delta} < \infty. \text{ by Assumption 4.5.1}
\end{aligned}$$

This guarantees that

$$\sup_k (\sqrt{k}\,|H_k - Z^*|)^{2+\delta} \leq \sup_k E\Big[k^{1+\frac{\delta}{2}} \sup_{x\in\mathcal{X}} \left|\frac{1}{k}\sum_{i=1}^k h(x, \xi_i) - Z(x)\right|^{2+\delta}\Big] < \infty$$

therefore the sequence of random variables $k(H_k - Z^*)^2$ is uniformly integrable. Since $\sqrt{k}(H_k - Z^*)$

is asymptotically normal we conclude $kE(H_k - Z^*)^2 \to Var(h(x^*, \xi))$, and hence

$$\limsup_k kE(H_k - W_k)^2 \leq \limsup_k kE(H_k - Z^*)^2 = Var(h(x^*, \xi)). \tag{C.31}$$

Next we show (C.30). Recall from the proof of Lemma 4.5.1 that when $\mathcal{X}^* = \{x^*\}$

$$k(g_k(\xi_1') - W_k) \begin{cases} \geq h(x^*, \xi_1') - Z^* - (M(\xi_1') + EM(\xi))E\left[\|x(\boldsymbol{\xi}') - x'(\boldsymbol{\xi}')\| \middle| \xi_1'\right] \\ \leq h(x^*, \xi_1') - Z^* + E[M(\xi_1)\|x(\boldsymbol{\xi}) - x'(\boldsymbol{\xi})\|] + M(\xi_1')E\|x(\boldsymbol{\xi}) - x'(\boldsymbol{\xi})\| \end{cases}$$

where $x(\boldsymbol{\xi}), x'(\boldsymbol{\xi}), x(\boldsymbol{\xi}'), x'(\boldsymbol{\xi}')$ are the same as those in the proof of Lemma 4.5.1. We have shown that the errors are all $o_p(1)$, hence $k(g_k(\xi_1') - W_k) \Rightarrow h(x^*, \xi_1') - Z^*$. On the other hand, when verifying Lyapunov condition in proving Theorem 4.5.2 we have already seen that $\sup_k k^{2+\delta} |g_k(\xi_1') - W_k|^{2+\delta} < \infty$. Therefore uniform integrability of $k^2(g_k(\xi_1') - W_k)^2$ follows and as $k \to \infty$

$$kE(\mathring{H}_k - W_k)^2 = k^2 E(g_k(\xi) - W_k)^2 \to E(h(x^*, \xi) - Z^*)^2 = Var(h(x^*, \xi))$$

which is exactly (C.30).

Now (C.31), (C.30) and the relation $E(H_k - W_k)^2 \geq E(\mathring{H}_k - W_k)^2$ together imply (C.29).

To justify the order of bias, note that (C.31) and (C.29) force that as $k \to \infty$

$$k(Z^* - W_k)^2 = kE(H_k - Z^*)^2 - kE(H_k - W_k)^2 \to 0$$

hence $Z^* - W_k = o(1/\sqrt{k})$.

The CLT when $k \geq \epsilon n$ follows from $Z^* - W_k = o(1/\sqrt{k}) = o(1/\sqrt{n})$, variance convergence (C.30) and Slutsky's theorem. □

C.4 Proof of Proposition 4.6.1

Proof of Proposition 4.6.1. Using Lemma C.2.2 with $c = 1$, we have

$$\begin{aligned}
Var(g_k(\xi)) &= E(E[H_k(\xi_1,\ldots,\xi_k)|\xi_1] - E[H_k(\xi_1,\ldots,\xi_k)])^2 \\
&\le \frac{1}{k^2} E\left(E\left[\sup_{x\in\mathcal{X}} |h(x,\xi_1') - h(x,\xi_1)| \Big| \xi_1'\right]\right)^2 \\
&\le \frac{1}{k^2} E\left(\sup_{x\in\mathcal{X}} |h(x,\xi_1') - h(x,\xi_1)|\right)^2 \quad \text{by Jensen's inequality} \\
&\le \frac{1}{k^2} E\left(\sup_{x\in\mathcal{X}} |h(x,\xi_1')| + \sup_{x\in\mathcal{X}} |h(x,\xi_1)|\right)^2 \\
&\le \frac{4}{k^2} E \sup_{x\in\mathcal{X}} |h(x,\xi)|^2 = O\big(\frac{1}{k^2}\big) \quad \text{by Minkowski inequality and Assumption 4.4.1}
\end{aligned}$$

This concludes the proposition. □

C.5 Proof of Theorem 4.6.2 and the Claim in Example 4.6.1

Proof of Theorem 4.6.2. From the batching procedure it is clear that $Var(\tilde{Z}_k) = Var(H_k)/m$ where m is the number of batches such that $mk = n$ if rounding errors are ignored. For our U-statistic, note that $Var(\mathring{U}_{n,k}) = k^2 Var(g_k(\xi))/n = kVar(\mathring{H}_k)/n$, and that the resample sizes in Theorems 4.5.2 and 4.5.4 satisfy the relation (C.27) hence by Lemma C.3.2 it holds $Var(U_{n,k} - \mathring{U}_{n,k}) = o(1/n) = o(Var(\mathring{U}_{n,k}))$. So the asymptotic ratio

$$\limsup_{n,k\to\infty} \frac{Var(U_{n,k})}{Var(\tilde{Z}_k)} = \limsup_{n,k\to\infty} \frac{Var(\mathring{U}_{n,k})}{Var(H_k)/m} = \limsup_{n,k\to\infty} \frac{kVar(\mathring{H}_k)/n}{Var(H_k)/m} = \limsup_{k\to\infty} \frac{Var(\mathring{H}_k)}{Var(H_k)}. \tag{C.32}$$

Then $r_U \le 1$ follows from the fact that $Var(\mathring{H}_k) \le Var(H_k)$. Under the conditions and resample sizes of Theorem 4.5.3 we have $E(V_{n,k} - U_{n,k})^2 = o(1/n) = o(Var(U_{n,k}))$ from the proof of Theorem 4.6.4, hence $r_V = r_U$ follows.

When $kVar(H_k) \to \infty$, it's obvious that (C.32) is equal to 0 since $kVar(\mathring{H}_k) = O(1)$ by Proposition 4.6.1. This proves the first case.

To show $r_U < 1$ in the second case, suppose $\limsup_{k\to\infty} Var(\mathring{H}_k)/Var(H_k) = 1$, then there

exists a subsequence $\{H_{k_s}\}_s$ such that

$$Var(\mathring{H}_{k_s})/Var(H_{k_s}) \to 1 \text{ as } s \to \infty$$

which implies that $Var(H_{k_s} - \mathring{H}_{k_s}) = o(Var(\mathring{H}_{k_s})) = o(1/k_s)$. By Theorem C.3.3, this ensures that U_{k_s,k_s}, or equivalently H_{k_s}, has a Gaussian limit as $s \to \infty$. However, the subsequence $\{H_{k_s}\}_s$ must have the same weak limit as the full sequence which is assumed non-Gaussian, thus leading to a contradiction.

For the last case, from (C.29) and (C.30) in the proof of Theorem 4.5.5 we know (C.32) is equal to 1. □

Proof of the claim in Example 4.6.1. Like in the proof of Theorem 4.5.5, we can follow the coupling argument for Lemma 4.5.1 to get

$$\begin{aligned} k(g_k(\xi_1') - W_k) &= E\left[\min_{x\in\mathcal{X}}\left\{h(x,\xi_1') + \sum_{i\neq 1} h(x,\xi_i)\right\} - \min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)\middle|\xi_1'\right] \\ &\leq E\left[h(x(\boldsymbol{\xi}),\xi_1') - h(x(\boldsymbol{\xi}),\xi_1)\middle|\xi_1'\right] \end{aligned}$$

where $x(\boldsymbol{\xi})$ is the optimal solution for $\min_{x\in\mathcal{X}}\sum_{i=1}^{k} h(x,\xi_i)$ hence is independent of ξ_1'. Note that $x(\boldsymbol{\xi})$ is uniformly distributed among $\{1,2,\ldots,d\}$, and that for any fixed ξ_1 the solution $x(\boldsymbol{\xi})$ will weakly converge to the same uniform distribution. Therefore

$$\begin{aligned} E\left[h(x(\boldsymbol{\xi}),\xi_1') - h(x(\boldsymbol{\xi}),\xi_1)\middle|\xi_1'\right] &= \frac{1}{d}\sum_{j=1}^{d}\xi_{1,j}' - E[h(x(\boldsymbol{\xi}),\xi_1)] \\ &= \frac{1}{d}\sum_{j=1}^{d}\xi_{1,j}' - E\left[\frac{1}{d}\sum_{j=1}^{d}\xi_{1,j} + \sum_{j=1}^{d}\big(P(x(\boldsymbol{\xi})=j) - \frac{1}{d}\big)\xi_{1,j}\right] \\ &= \frac{1}{d}\sum_{j=1}^{d}\xi_{1,j}' - E\left[E\left[\sum_{j=1}^{d}\big(P(x(\boldsymbol{\xi})=j) - \frac{1}{d}\big)\xi_{1,j}\middle|\xi_1\right]\right] \\ &= \frac{1}{d}\sum_{j=1}^{d}\xi_{1,j}' + o(1) \text{ by dominated convergence theorem} \end{aligned}$$

where $\xi'_{1,j}$ and $\xi_{1,j}$ are the j-th components of ξ'_1 and ξ_1 respectively. Therefore we have shown $k(g_k(\xi'_1) - W_k) \leq \sum_{j=1}^d \xi'_{1,j}/d + o(1)$. Similarly, denoting by $x(\boldsymbol{\xi}')$ the optimal solution for the optimization $\min_{x\in\mathcal{X}}\{h(x,\xi'_1) + \sum_{i=2}^k h(x,\xi_i)\}$, the lower bound can be obtained as

$$\begin{aligned} k(g_k(\xi'_1) - W_k) &\geq E\left[h(x(\boldsymbol{\xi}'),\xi'_1) - h(x(\boldsymbol{\xi}'),\xi_1)\Big|\xi'_1\right] \\ &= E\left[h(x(\boldsymbol{\xi}'),\xi'_1)\Big|\xi'_1\right] \text{ by independence between } \xi_1 \text{ and } x(\boldsymbol{\xi}') \\ &= \frac{1}{d}\sum_{j=1}^d \xi'_{1,j} + \sum_{j=1}^d \big(P(x(\boldsymbol{\xi}') = j|\xi'_1) - \frac{1}{d}\big)\xi'_{1,j} \\ &= \frac{1}{d}\sum_{j=1}^d \xi'_{1,j} + o_p(1). \end{aligned}$$

The lower and upper bounds agree so

$$k(g_k(\xi'_1) - W_k) \Rightarrow \frac{1}{d}\sum_{j=1}^d \xi'_{1,j} = N\big(0, \frac{1}{d}\big).$$

On the other hand $k^2(g_k(\xi'_1) - W_k)^2$ is uniformly integrable as argued in the proof of Theorem 4.5.5, hence $k^2 Var(g_k(\xi)) = E[k^2(g_k(\xi'_1) - W_k)^2] \to 1/d$. □

C.6 Proof of Theorem 4.6.4

Proof of Theorem 4.6.4. For $U_{n,k}$, note that each summand in its definition is an SAA value with distinct i.i.d. data, and thus has mean exactly W_k. For $V_{n,k}$, recall the relation (C.24)

$$n^k(U_{n,k} - V_{n,k}) = \big(\sum_{s=1}^{k-l-1} c(n,k,s)\big)(U_{n,k} - R_{n,l}) - \sum_{s=k-l}^{k-1} c(n,k,s)(A_{n,s} - U_{n,k}).$$

Note that $U_{n,k}$ is unbiased for W_k, and that $ER_{n,l} = O(1)$ since Assumption 4.4.1 implies for any indices $i_1,\ldots,i_k \in \{1,\ldots,n\}$ that $|EH_k(\xi_{i_1},\ldots,\xi_{i_k})| \leq E\sup_{x\in\mathcal{X}}|h(x,\xi)|$. In the proof of Theorem 4.5.3 we have shown that $E|A_{n,s} - U_{n,k}| = O(1/k)$, $\sum_{s=1}^{k-l-1} c(n,k,s) = O((k^2/n)^{l+1}n^k)$

when $k = o(\sqrt{n})$, and that $c(n,k,s) = O(k^{2(k-s)}n^s)$ for $s \geq k-l$. Therefore

$$n^k \left| EV_{n,k} - W_k \right| \leq O\big(\big(\frac{k^2}{n}\big)^{l+1} n^k\big) + O(1/k) \sum_{s=k-l}^{k-1} O(k^{2(k-s)}n^s).$$

Since $k^2/n = o(1)$, it holds $\sum_{s=k-l}^{k-1} O(k^{2(k-s)}n^s) = O(k^2 n^{k-1})$, which leads to $EV_{n,k} - W_k = O((k^2/n)^{l+1} + k/n)$ for any fixed $l \geq 0$. □

C.7 Proof of Theorems 4.7.1 and 4.7.2

Proof of Theorem 4.7.1. Wager and Athey (2018) provides a proof in the context of random forests. Since their proof can be adapted to our optimization context, we shall directly borrow some intermediate results there which hold for general symmetric kernels and U-statistics, and only focus on parts that rely on the particular SAA kernel considered there. Readers are referred to the proof of Theorem 9 in Wager and Athey (2018) for explanations of the borrowed results.

Note that Theorems 4.5.2, 4.5.4 and 4.5.5 can be viewed as special cases of Theorem C.3.3 where $E(H_k - \mathring{H}_k)^2$ is $O(1)$, $O(1/k)$ and $o(1/k)$ respectively. So it suffices to show consistency in the more general setting under Theorem C.3.3 and, if not implied by (C.27), the additional requirement $k \leq \theta n$ for some $\theta < 1$. The IJ variance estimator now can be expressed as

$$\begin{aligned}
\frac{n^2}{(n-k)^2} \sum_{i=1}^{n} \mathrm{Cov}_*^2(N_i^*, H_k^*) &= \frac{n^2}{(n-k)^2} \sum_{i=1}^{n} (E_*[H_k^* \sum_{j=1}^{k} \mathbf{1}(\xi_{i_j} = \xi_i)] - E_*[N_i^*]E_*[H_k^*])^2 \\
&= \frac{n^2}{(n-k)^2} \sum_{i=1}^{n} (kE_*[H_k^* \mathbf{1}(\xi_{i_1} = \xi_i)] - \frac{k}{n} U_{n,k})^2 \\
&= \frac{n^2}{(n-k)^2} \frac{k^2}{n^2} \sum_{i=1}^{n} (E_*[H_k^* | \xi_{i_1} = \xi_i] - U_{n,k})^2 \qquad \text{(C.33)} \\
&= \frac{k^2}{(n-k)^2} \sum_{i=1}^{n} (A_i + R_i)^2
\end{aligned}$$

where $\xi_{i_1}, \ldots, \xi_{i_k}$ are resampled from $\xi_1, \ldots, \xi_n$ without replacement, and

$$A_i = E_*[\mathring{H}_k^* | \xi_{i_1} = \xi_i] - E_*[\mathring{H}_k^*]$$
$$R_i = E_*[H_k^* - \mathring{H}_k^* | \xi_{i_1} = \xi_i] - E_*[H_k^* - \mathring{H}_k^*].$$

We aim to show that

$$\frac{k^2}{(n-k)^2} \sum_{i=1}^n A_i^2 \Big/ \frac{k^2 Var(g_k(\xi))}{n} \xrightarrow{p} 1, \quad \frac{k^2}{(n-k)^2} \sum_{i=1}^n R_i^2 = o_p\big(\frac{1}{n}\big) \tag{C.34}$$

so that consistency follows by an application of Cauchy Schwartz inequality to the cross term $\sum_{i=1}^n 2A_i R_i$.

First we deal with R_i's. Lemma 13 in Wager and Athey (2018) shows that

$$ER_i^2 = \sum_{s=2}^k (a_s + b_s) V_s^H$$

where

$$a_s = \binom{n-1}{s-1}\Big(\binom{k-1}{s-1}\Big/\binom{n-1}{s-1} - \binom{k}{s}\Big/\binom{n}{s}\Big)^2$$
$$b_s = \binom{n-1}{s}\Big(\binom{k-1}{s}\Big/\binom{n-1}{s} - \binom{k}{s}\Big/\binom{n}{s}\Big)^2$$

with $b_k = 0$, and V_s^H is the variance of the s-th order function in the ANOVA decomposition of H_k (see the discussion after Lemma C.3.1). Note that $Var(H_k) = \sum_{s=1}^k \binom{k}{s} V_s^H$ and $Var(\mathring{H}_k) = kV_1^H$. Some basic algebra shows that

$$\frac{a_{s+1}/\binom{k}{s+1}}{a_s/\binom{k}{s}} = \frac{(s+1)(k-s)}{s(n-s)}, \quad \frac{b_{s+1}/\binom{k}{s+1}}{b_s/\binom{k}{s}} = \frac{(s+1)^2(k-s)}{s^2(n-s-1)}.$$

Therefore, if $k \le \theta n$ for $\theta < 1$, the above two ratios are both less than one when $s \ge s^* := \max\{2, \lceil \sqrt{\theta}/(1-\sqrt{\theta}) \rceil\}$, meaning that the maximum of $a_s/\binom{k}{s}$ or $b_s/\binom{k}{s}$ over s is attained at some $s \le s^*$. Moreover, by upper bounding $(k-s)/(n-s-1) < 1$ we have for all $s \le s^*$ that

$b_s/\binom{k}{s}/\big(b_2/\binom{k}{2}\big) \le s^2/4 \le s^{*2}/4$ and that $a_s/\binom{k}{s}/\big(a_2/\binom{k}{2}\big) \le s/2 \le s^*/2 \le s^{*2}/4$. Hence

$$\begin{aligned} ER_i^2 &\le \frac{s^{*2}}{4}\frac{a_2+b_2}{\binom{k}{2}}\sum_{s=2}^{s^*}\binom{k}{s}V_s^H + \sum_{s=s^*+1}^{k}\frac{a_s+b_s}{\binom{k}{s}}\binom{k}{s}V_s^H \\ &\le \frac{s^{*2}}{4}\frac{a_2+b_2}{\binom{k}{2}}\sum_{s=2}^{k}\binom{k}{s}V_s^H \le C(\theta)\frac{(n-k)^2}{n^3}E(H_k-\mathring{H}_k)^2 \end{aligned}$$

where $C(\theta)$ is a constant that only depends on θ. This bound implies

$$E\big[\frac{k^2}{(n-k)^2}\sum_{i=1}^{n}R_i^2\big] = O\big(\frac{k^2}{n^2}E(H_k-\mathring{H}_k)^2\big) = o\big(\frac{1}{n}\big) \tag{C.35}$$

where the second equality follows from the requirement on resample size in Theorem C.3.3.

Now we analyze the A_i's. Lemma 12 in Wager and Athey (2018) shows that

$$A_i = \big(1-\frac{k}{n}\big)(g_k(\xi_i)-W_k) + \big(\frac{k-1}{n-1}-\frac{k}{n}\big)\sum_{j\neq i}(g_k(\xi_j)-W_k)$$

therefore one can write

$$\frac{(n-1)^2k^2}{n^2(n-k)^2}\sum_{i=1}^{n}A_i^2 = \frac{k^2}{n}\Big(\frac{1}{n}\sum_{i=1}^{n}(g_k(\xi_i)-W_k)^2 - (\bar{g}_k-W_k)^2\Big), \text{ where } \bar{g}_k = \frac{1}{n}\sum_{i=1}^{n}g_k(\xi_i).$$

Since $E[k^2(\bar{g}_k-W_k)^2/n] = k^2Var(g_k(\xi))/n^2 = O(1/n^2) = o(1/n)$ it suffices to prove

$$\frac{1}{n}\sum_{i=1}^{n}(g_k(\xi_i)-W_k)^2/Var(g_k(\xi)) \xrightarrow{p} 1 \tag{C.36}$$

in order to justify the first limit equality in (C.34). To proceed, we need the following weak law of large numbers:

Lemma C.7.1 (Theorem 2.2.9 from Durrett (2010)) *For each n let $Y_{n,i}, 1 \le i \le n$ be independent. Let $b_n > 0$ with $b_n \to \infty$, and let $\bar{Y}_{n,i} = Y_{n,i}\mathbf{1}(|Y_{n,i}| \le b_n)$. Suppose that, as $n \to \infty$,*

$\sum_{i=1}^n P(|Y_{n,i}| > b_n) \to 0$ *and* $b_n^{-2}\sum_{i=1}^n E\bar{Y}_{n,i}^2 \to 0$, *then*

$$\frac{\sum_{i=1}^n Y_{n,i} - \sum_{i=1}^n E\bar{Y}_{n,i}}{b_n} \xrightarrow{p} 0.$$

We apply the weak law to $Y_{n,i} = (g_k(\xi_i) - W_k)^2/Var(g_k(\xi))$ with $b_n = n$. To verify the conditions

$$\begin{aligned}
nP\big(\frac{(g_k(\xi_i)-W_k)^2}{Var(g_k(\xi))} > n\big) &= nP(|g_k(\xi_i)-W_k|^{2+\delta} > (nVar(g_k(\xi)))^{1+\frac{\delta}{2}}) \\
&\le \frac{n}{(nVar(g_k(\xi)))^{1+\frac{\delta}{2}}} E\,|g_k(\xi_i)-W_k|^{2+\delta} \text{ by Markov inequality} \\
&\le \frac{n}{(nVar(g_k(\xi)))^{1+\frac{\delta}{2}}} \frac{\tilde{M}}{k^{2+\delta}} \text{ by the proof of Theorem 4.5.2} \\
&= \frac{\tilde{M}}{n^{\frac{\delta}{2}}(k^2Var(g_k(\xi)))^{1+\frac{\delta}{2}}} = O(n^{-\frac{\delta}{2}}) \to 0
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n}E\left[\frac{(g_k(\xi_i)-W_k)^4}{(Var(g_k(\xi)))^2}\mathbf{1}\big(\frac{(g_k(\xi_i)-W_k)^2}{Var(g_k(\xi))} \le n\big)\right] \\
\le\ &\frac{1}{n}E\left[\frac{|g_k(\xi_i)-W_k|^{2+\delta}}{(Var(g_k(\xi)))^{1+\frac{\delta}{2}}} n^{1-\frac{\delta}{2}}\mathbf{1}\big(\frac{(g_k(\xi_i)-W_k)^2}{Var(g_k(\xi))} \le n\big)\right] \\
\le\ &\frac{1}{n^{\frac{\delta}{2}}}E\left[\frac{|g_k(\xi_i)-W_k|^{2+\delta}}{(Var(g_k(\xi)))^{1+\frac{\delta}{2}}}\right] \\
\le\ &\frac{\tilde{M}}{n^{\frac{\delta}{2}}(k^2Var(g_k(\xi)))^{1+\frac{\delta}{2}}} \to 0.
\end{aligned}$$

It remains to show that

$$\begin{aligned}
&\left|1 - E\left[\frac{(g_k(\xi_i)-W_k)^2}{Var(g_k(\xi))}\mathbf{1}\big(\frac{(g_k(\xi_i)-W_k)^2}{Var(g_k(\xi))} \le n\big)\right]\right| \\
=\ &\left|E\left[\frac{(g_k(\xi_i)-W_k)^2}{Var(g_k(\xi))}\mathbf{1}\big(\frac{(g_k(\xi_i)-W_k)^2}{Var(g_k(\xi))} > n\big)\right]\right| \\
\le\ &\left(E\left[\frac{|g_k(\xi_i)-W_k|^{2+\delta}}{(Var(g_k(\xi)))^{1+\frac{\delta}{2}}}\right]\right)^{\frac{2}{2+\delta}}\left(P\big(\frac{(g_k(\xi_i)-W_k)^2}{Var(g_k(\xi))} > n\big)\right)^{\frac{\delta}{2+\delta}} \text{ by Holder's inequality} \\
\le\ &\left(\frac{\tilde{M}}{(k^2Var(g_k(\xi)))^{1+\frac{\delta}{2}}}\right)^{\frac{2}{2+\delta}}\left(\frac{1}{n}\right)^{\frac{\delta}{2+\delta}} \to 0 \text{ by Markov inequality.}
\end{aligned}$$

With all these conditions verified, we can conclude (C.36) from Lemma C.7.1 and complete the proof. □

Proof of Theorem 4.7.2. Given Theorem 4.7.1, it suffices to show that the IJ variance estimator under resampling with replacement differs by only $o_p(1/n)$ from the one without replacement. Since quantities under both resampling with and without replacement will be involved in this proof, we attach $*$ to quantities under resampling without replacement, and $\tilde{*}$ to those with replacement. Note that $k = O(n^\gamma)$ for some $\gamma < 1/2$ which implies $n^2/(n-k)^2 \to 1$, so the without-replacement IJ variance estimate without the factor $n^2/(n-k)^2$, i.e. $\sum_{i=1}^n \mathrm{Cov}_*^2(N_i^*, H_k^*)$, is also consistent. We have

$$\sum_{i=1}^n \mathrm{Cov}_{\tilde{*}}^2(N_i^{\tilde{*}}, H_k^{\tilde{*}}) = \frac{k^2}{n^2}\sum_{i=1}^n (E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1} = \xi_i] - V_{n,k})^2 \tag{C.37}$$

where $\xi_{i_1}, \ldots, \xi_{i_k}$ are resampled from $\xi_1, \ldots, \xi_n$ with replacement. By comparing (C.33) (without $n^2/(n-k)^2$) and (C.37) and using Cauchy Schwartz inequality

$$\left|\sum_{i=1}^n \mathrm{Cov}_{\tilde{*}}^2(N_i^{\tilde{*}}, H_k^{\tilde{*}}) - \sum_{i=1}^n \mathrm{Cov}_*^2(N_i^*, H_k^*)\right| \le \frac{k^2}{n^2}\sum_{i=1}^n (v_i - u_i)^2 + 2\sqrt{\sum_{i=1}^n \mathrm{Cov}_*^2(N_i^*, H_k^*) \cdot \frac{k^2}{n^2}\sum_{i=1}^n (v_i - u_i)^2}$$

where $v_i = E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1} = \xi_i] - V_{n,k}$ and $u_i = E_*[H_k^*|\xi_{i_1} = \xi_i] - U_{n,k}$. If we show that $E(V_{n,k} - U_{n,k})^2 = o(1/n)$ and $E(E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1} = \xi_i] - E_*[H_k^*|\xi_{i_1} = \xi_i])^2 = o(1/n)$, then $E[\sum_{i=1}^n (v_i - u_i)^2] = o(1)$ and under the condition $k = O(n^\gamma)$ with $\gamma < 1/2$ we have

$$\sum_{i=1}^n \mathrm{Cov}_{\tilde{*}}^2(N_i^{\tilde{*}}, H_k^{\tilde{*}}) - \sum_{i=1}^n \mathrm{Cov}_*^2(N_i^*, H_k^*) = \frac{k^2}{n^2}o_p(1) + \sqrt{o_p\big(\frac{1}{n}\cdot\frac{k^2}{n^2}\big)} = o_p\big(\frac{1}{n}\big)$$

which concludes the theorem.

The first error $E(V_{n,k} - U_{n,k})^2 = o(1/n)$ can be deduced from (C.24) in the proof of Theorem 4.5.3. We only need to notice that, in the setting of that proof, $E(U_{n,k} - R_{n,l})^2 = O(1)$ due to Assumption 4.4.1 and that each $E(A_{n,s} - U_{n,k})^2 = O(1/k^2)$ for $s \ge k - l$ due to Assumption 4.5.1.

The second error $E(E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1} = \xi_i] - E_*[H_k^*|\xi_{i_1} = \xi_i])^2 = o(1/n)$ needs some further discussion. We study $E(E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1} = \xi_1] - E_*[H_k^*|\xi_{i_1} = \xi_1])^2$ without loss of generality. Given that the first resampled data point ξ_{i_1} is ξ_1, for any fixed integer $l \ge 0$ we obtain the following decomposition of

$E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1}=\xi_1]$ similar to that in the proof of Theorem 4.5.3

$$n^{k-1}E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1}=\xi_1]=\sum_{s=k-1-l}^{k-1}c(n-1,k-1,s)A_s+(n^{k-1}-\sum_{s=k-1-l}^{k-1}c(n-1,k-1,s))R_l$$

where A_s is the average of all $H_k(\xi_1,\xi_{i_2},\ldots,\xi_{i_k})$'s where $\xi_{i_2},\ldots,\xi_{i_k}$ contain exactly s distinct data and none of them is ξ_1, and R_l is the average of all other $H_k(\xi_1,\xi_{i_2},\ldots,\xi_{i_k})$'s. Note that, in particular, $A_{k-1}=E_*[H_k^*|\xi_{i_1}=\xi_1]$. We have the following analog of (C.24)

$$\begin{aligned}&n^{k-1}(E_*[H_k^*|\xi_{i_1}=\xi_1]-E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1}=\xi_1])\\ =\;&(n^{k-1}-\sum_{s=k-1-l}^{k-1}c(n-1,k-1,s))(A_{k-1}-R_l)-\sum_{s=k-1-l}^{k-2}c(n-1,k-1,s)(A_s-A_{k-1}).\end{aligned}$$

Note that the coefficient of the first term does not match the form of (C.24), but we have

$$n^{k-1}-\sum_{s=k-1-l}^{k-1}c(n-1,k-1,s)=n^{k-1}-(n-1)^{k-1}+\sum_{s=1}^{k-l-2}c(n-1,k-1,s).$$

Like in the proof of Theorem 4.5.3

$$\begin{aligned}&\sum_{s=1}^{k-l-2}c(n-1,k-1,s)=O\big((\frac{k^2}{n})^{l+1}(n-1)^{k-1}\big),\ E(A_{k-1}-R_l)^2=O(1)\\ &c(n-1,k-1,s)=O(k^{2(k-1-s)}n^s)\text{ and }E(A_s-A_{k-1})^2=O\big(\frac{1}{k^2}\big)\text{ for }s\geq k-1-l.\end{aligned}$$

Moreover by Bernoulli's inequality $(1+x)^r\geq 1+rx$ for any integer $r\geq 0$ and real $x\geq -1$

$$n^{k-1}-(n-1)^{k-1}=n^{k-1}(1-(1-\frac{1}{n})^{k-1})\leq n^{k-2}(k-1).$$

With all these bounds and Minkowski inequality we get

$$\begin{aligned} & E(E_*[H_k^*|\xi_{i_1}=\xi_1]-E_{\tilde{*}}[H_k^{\tilde{*}}|\xi_{i_1}=\xi_1])^2 \\ = \; & O\left((\frac{k}{n}+(\frac{k^2}{n})^{l+1})^2 E(A_{k-1}-R_l)^2+\sum_{s=k-1-l}^{k-2}(\frac{k^2}{n})^{2(k-1-s)}E(A_s-A_{k-1})^2\right) \\ = \; & O\left((\frac{k}{n}+(\frac{k^2}{n})^{l+1})^2+\frac{k^2}{n^2}\right)=o(\frac{1}{n}) \end{aligned}$$

when l is chosen according to (C.25). □

C.8 Proof of Theorem 4.7.3 and Corollary 4.7.4

Proof of Theorem 4.7.3. We have two tasks. One is that $\tilde{Z}_k^{bag}-U_{n,k}=o_p(1/\sqrt{n})$ when resampling without replacement, or $\tilde{Z}_k^{bag}-V_{n,k}=o_p(1/\sqrt{n})$ with replacement, so that by Slutsky's theorem the CLTs still hold with $U_{n,k}$ or $V_{n,k}$ replaced by their estimate $\tilde{Z}_k^{bag}$. The other thing is that

$$\left|\sum_{i=1}^n \widehat{Cov}_*^2(N_i^*,\hat{Z}_k^*)-\sum_{i=1}^n Cov_*^2(N_i^*,H_k^*)\right|=o_p(1/n)$$

so that the variance estimation is consistent and CLTs remain valid by Slutsky's theorem. Note that in the case of resampling without replacement, the condition $k\le\theta n$ for some $\theta<1$ implies $1\le n^2/(n-k)^2\le 1/(1-\theta)^2$, hence the error remains $o_p(1/n)$ after multiplying the factor $n^2/(n-k)^2$.

The first task is relatively easy. Note that $\tilde{Z}_k^{bag}$ is unbiased (for estimating $U_{n,k}$ and $V_{n,k}$ respectively) in either case, and

$$Var_*(\tilde{Z}_k^{bag})=\frac{1}{B}Var_*(H_k^*)\le\frac{1}{B}E_*H_k^{*2}\le\frac{1}{Bn}\sum_{i=1}^n\sup_{x\in\mathcal{X}}|h(x,\xi_i)|^2 \tag{C.38}$$

where the last inequality follows from the argument used in (4.16). Due to Assumption 4.4.1 and the strong law of large numbers $\sum_{i=1}^n\sup_{x\in\mathcal{X}}|h(x,\xi_i)|^2/n\xrightarrow{p}E\sup_{x\in\mathcal{X}}|h(x,\xi)|^2<\infty$, hence $Var_*(\tilde{Z}_k^{bag})=O_p(1/B)$. If $B/(kn)\to\infty$ we have

$$E_*(\tilde{Z}_k^{bag}-U_{n,k})^2=o_p(\frac{1}{kn})=o_p(\frac{1}{n}),\; E_*(\tilde{Z}_k^{bag}-V_{n,k})^2=o_p(\frac{1}{kn})=o_p(\frac{1}{n})$$

For a non-negative random variable, if its conditional expectation is of order $o_p(1)$, then itself is also $o_p(1)$. Therefore $(\tilde{Z}_k^{bag} - U_{n,k})^2 = o_p(1/n)$ and $(\tilde{Z}_k^{bag} - V_{n,k})^2 = o_p(1/n)$.

For the second task, we first deal with resampling without replacement. By Cauchy Schwartz inequality the Monte Carlo error can be bounded as

$$\left| \sum_{i=1}^n \widehat{Cov}_*^2(N_i^*, \hat{Z}_k^*) - \sum_{i=1}^n Cov_*^2(N_i^*, H_k^*) \right| \leq \sum_{i=1}^n (\widehat{Cov}_i - Cov_i)^2 + 2\sqrt{\sum_{i=1}^n Cov_i^2 \sum_{i=1}^n (\widehat{Cov}_i - Cov_i)^2}$$

where $Cov_i = Cov_*(N_i^*, H_k^*)$ and $\widehat{Cov}_i = \widehat{Cov}_*^2(N_i^*, \hat{Z}_k^*)$ for short. Since $\sum_{i=1}^n Cov_i^2$ is the desired variance of order $1/n$, we only need to show $\sum_{i=1}^n (\widehat{Cov}_i - Cov_i)^2 = o_p(1/n)$. By computing variances of the sample covariances one can get

$$\begin{aligned} & E_*\big[\sum_{i=1}^n (\widehat{Cov}_i - Cov_i)^2\big] \\ \leq & \sum_{i=1}^n \left(\frac{1}{B} E_*[(H_k^* - E_* H_k^*)^2 (N_i^* - \frac{k}{n})^2] + \frac{1}{B^2} Var_*(H_k^*) Var_*(N_i^*) + \frac{2}{B} Cov_i^2 \right) \\ \leq & \frac{1}{B} E_*[(H_k^* - E_* H_k^*)^2 \sum_{i=1}^n (N_i^* - \frac{k}{n})^2] + \frac{1}{B^2} Var_*(H_k^*) \sum_{i=1}^n Var_*(N_i^*) + \frac{2}{B} \sum_{i=1}^n Cov_i^2 . \quad \text{(C.39)} \end{aligned}$$

Note that $\sum_{i=1}^n Cov_i^2 = O_p(1/n)$, $Var_*(H_k^*) = O_p(1)$ as shown in (C.38), and $\sum_{i=1}^n (N_i^* - \frac{k}{n})^2 = k(n-k)/n, Var_*(N_i^*) = k(n-k)/n^2$ since $N_i^* = 0$ or 1 and $\sum_{i=1}^n N_i^* = k$. With all these bounds, we have

$$E_*\big[\sum_{i=1}^n (\widehat{Cov}_i - Cov_i)^2\big] = O_p\big(\frac{k}{B} + \frac{k}{B^2} + \frac{1}{Bn}\big) = O_p\big(\frac{k}{B}\big).$$

If $B/(kn) \to \infty$, then $E_*\big[\sum_{i=1}^n (\widehat{Cov}_i - Cov_i)^2\big] = o_p(1/n)$, which implies $\sum_{i=1}^n (\widehat{Cov}_i - Cov_i)^2 = o_p(1/n)$.

In the case of resampling with replacement, we have the same bound (C.39), where $Var_*(N_i^*) = k(n-1)/n^2$ and $Var_*(H_k^*) = O_p(1)$. However, the first term becomes more complicated. We bound

the first term by a conditioning argument on N_i^*

$$\begin{aligned} E_*\big[H_k^{*2}|N_i^*=s\big] &\leq E_*\big[\big(\frac{1}{k}\sum_{j=1}^{k}\sup_{x\in\mathcal{X}}|h(x,\xi_{i_j})|\big)^2\big|N_i^*=s\big] \\ &= \frac{s}{k}\sup_{x\in\mathcal{X}}|h(x,\xi_i)|^2+\frac{k-s}{k}\frac{1}{n-1}\sum_{j\neq i}\sup_{x\in\mathcal{X}}|h(x,\xi_j)|^2 \text{ by Minkowski inequality} \end{aligned}$$

therefore

$$\begin{aligned} E_*\big[H_k^{*2}(N_i^*-\frac{k}{n})^2\big] &= \sum_{s=0}^{k}E_*\big[H_k^{*2}|N_i^*=s\big](s-\frac{k}{n})^2P(N_i^*=s) \\ &\leq \sum_{s=0}^{k}\big(\frac{s}{k}\sup_{x\in\mathcal{X}}|h(x,\xi_i)|^2+\frac{k-s}{k}\frac{1}{n-1}\sum_{j\neq i}\sup_{x\in\mathcal{X}}|h(x,\xi_j)|^2\big)(s-\frac{k}{n})^2P(N_1^*=s). \end{aligned}$$

Now we have

$$\begin{aligned} E_*\big[H_k^{*2}\sum_{i=1}^{n}(N_i^*-\frac{k}{n})^2\big] &= \sum_{i=1}^{n}E_*\big[H_k^{*2}(N_i^*-\frac{k}{n})^2\big] \\ &\leq \sum_{s=0}^{k}\big(\sum_{i=1}^{n}\sup_{x\in\mathcal{X}}|h(x,\xi_i)|^2\big)(s-\frac{k}{n})^2P(N_1^*=s) \\ &\leq \sum_{i=1}^{n}\sup_{x\in\mathcal{X}}|h(x,\xi_i)|^2Var_*(N_1^*)=O_p(k) \end{aligned}$$

and the first term can be bounded as

$$E_*[(H_k^*-E_*H_k^*)^2\sum_{i=1}^{n}(N_i^*-\frac{k}{n})^2]\leq 2E_*[H_k^{*2}\sum_{i=1}^{n}(N_i^*-\frac{k}{n})^2]+2(E_*H_k^*)^2E_*[\sum_{i=1}^{n}(N_i^*-\frac{k}{n})^2]=O_p(k).$$

With these bounds, $E_*\big[\sum_{i=1}^{n}(\widehat{Cov}_i-Cov_i)^2\big]=O_p(k/B)$ and the conclusion follows. □

Proof of Corollary 4.7.4. From Theorem 4.7.3, we have

$$P\left(\frac{\tilde{Z}_k^{bag}-W_k}{\tilde{\sigma}_{IJ}}\leq z_{1-\alpha}\right)\to 1-\alpha \tag{C.40}$$

Note that

$$P\left(\frac{\tilde{Z}_k^{bag} - W_k}{\tilde{\sigma}_{IJ}} \leq z_{1-\alpha}\right) = P\left(\tilde{Z}_k^{bag} - z_{1-\alpha}\tilde{\sigma}_{IJ} \leq W_k\right) \leq P\left(\tilde{Z}_k^{bag} - z_{1-\alpha}\tilde{\sigma}_{IJ} \leq Z^*\right) \tag{C.41}$$

by (4.6). Combining (C.40) and (C.41) gives the conclusion. □

Appendix D

Technical Proofs for Chapter 5

This chapter presents technical proofs for results in Chapter 5. In all the proofs, for universal constants which are usually denoted C or c, we abuse notation slightly to allow C or c to take a different value in each appearance. For example, consider three quantities x, y, z such that $x \leq Cy$ and $z \leq 2x$. This implies $z \leq 2Cy$, but we would write as $z \leq Cy$ to simplify the notation.

D.1 Existing Central Limit Theorems in High Dimensions

This section reviews some results on high-dimensional central limit theorems that are needed subsequently in our proofs. We start with some notations. Let $\mathbf{X}_i := (X_{i,1}, \ldots, X_{i,p}), i = 1, \ldots, n$ be n i.i.d. copies of the random vector $\mathbf{X} := (X_1, \ldots, X_p) \in \mathbb{R}^p$, and $\mu_j := E[X_j]$ for $j = 1, \ldots, p$. Let $\bar{X}_j = \sum_{i=1}^n X_{i,j}/n$ be the sample mean of the j-th component. We denote by $\mathbf{Z} := (Z_1, \ldots, Z_p)$ a p-dimensional Gaussian random vector with $E[Z_j] = 0$ and covariance structure $Cov(Z_j, Z_{j'}) = \Sigma(j, j') := Cov(X_j, X_{j'})$ for $j, j' = 1, \ldots, p$, and by $\widehat{\mathbf{Z}} := (\widehat{Z}_1, \ldots, \widehat{Z}_p)$ a p-dimensional centered Gaussian random vector with covariance $\widehat{\Sigma}$, where

$$\widehat{\Sigma}(j, j') = \frac{1}{n} \sum_{i=1}^n X_{i,j} X_{i,j'} - \bar{X}_j \bar{X}_{j'}$$

is the sample covariance of all $\mathbf{X}_i$'s. We also denote $\sigma_j^2 = \Sigma(j, j)$ and $\hat{\sigma}_j^2 = \widehat{\Sigma}(j, j)$.

We make the following assumption:

Assumption D.1.1 *There exist constants $b > 0$ and $B \geq 1$ such that*

$$Var[X_j] \geq b \text{ and } E[\exp(|X_j - \mu_j|^2 / B^2)] \leq 2 \text{ for all } j = 1, \ldots, p$$
$$E[|X_j - \mu_j|^{2+k}] \leq B^k \text{ for all } j = 1, \ldots, p \text{ and } k = 1, 2.$$

Note that, since the sub-exponential norm of a random variable is always bounded above by its sub-Gaussian norm up to some universal constant C, the exponential condition in Assumption D.1.1 implies $E[\exp(|X_j - \mu_j| / (CB))] \leq 2$. Chernozhukov et al. (2017) proved the following CLT:

Theorem D.1.1 (Proposition 2.1 in Chernozhukov et al. (2017)) *When Assumption D.1.1 holds, we have*

$$\sup_{a_j \leq b_j, j=1,\ldots,p} \left| P(a_j \leq \sqrt{n}(\bar{X}_j - \mu_j) \leq b_j \text{ for all } j) - P(a_j \leq Z_j \leq b_j \text{ for all } j) \right|$$
$$\leq C_1 \Big(\frac{B^2 \log^7(pn)}{n} \Big)^{\frac{1}{6}}$$

where the constant C_1 depends only on b.

To derive confidence bounds based on the CLT, one needs to properly estimate the quantile of the limit Gaussian vector $\mathbf{Z} \sim N_p(0, \Sigma)$. One common approach is to use the Gaussian vector $\widehat{\mathbf{Z}} \sim N_p(0, \widehat{\Sigma})$, where $\widehat{\Sigma}$ is the sample covariance matrix, to approximate $\mathbf{Z}$. This approach is also called the multiplier bootstrap. Chernozhukov et al. (2017) gave the following result concerning the statistical accuracy of the multiplier bootstrap:

Theorem D.1.2 (Corollary 4.2 in Chernozhukov et al. (2017)) *If Assumption D.1.1 holds, then for any constant $0 < \alpha < \frac{1}{e}$ we have*

$$\sup_{a_j \leq b_j, j=1,\ldots,p} \left| P(a_j \leq \widehat{Z}_j \leq b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(a_j \leq Z_j \leq b_j \text{ for all } j) \right|$$
$$\leq C_2 \Big(\frac{B^2 \log^5(pn) \log^2(1/\alpha)}{n} \Big)^{\frac{1}{6}}$$

with probability at least $1 - \alpha$, where the constant C_2 depends only on b.

D.2 Proofs of Results in Section 5.3

This section proves the performance guarantees of our Gaussian supremum validators. Section D.2.1 adapts the high-dimensional CLTs in Section D.1 to handle small-variance situations that potentially arise in our optimization context. Section D.2.2 extends them to the case where the sample means are normalized by their standard deviations, a key step to justify our normalized Gaussian supremum validator. Section D.2.3 presents results on the consistency of the multiplier bootstrap to approximate the limiting Gaussian distributions. Section D.2.4 puts together all these results to synthesize the main proofs for Section 5.3.

D.2.1 A CLT for Random Vectors with Potentially Small Variances

Note that in both Theorems D.1.1 and D.1.2, the constants C_1, C_2 depend on the minimum variance b. By tracing the proof in Chernozhukov et al. (2017), the constant C_1 is of the form $c_1(b^{-1} \vee c_2)$ where c_1, c_2 are two universal constants. Due to such a dependence on the minimum variance, the bound can deteriorate when the noise levels across different components of $\mathbf{X}$ are not of the same scale, e.g., in the case of CCPs. To resolve this issue, we derive an alternate CLT that applies to normalized random vectors. We assume:

Assumption D.2.1 *$Var[X_j] > 0$ for all $j = 1, \ldots, p$ and there exists some constant $D_1 \geq 1$ such that*

$$E\big[\exp\Big(\frac{|X_j - \mu_j|^2}{D_1^2 Var[X_j]}\Big)\big] \leq 2 \text{ for all } j = 1, \ldots, p \tag{D.1}$$

$$E\big[\Big(\frac{|X_j - \mu_j|}{\sqrt{Var[X_j]}}\Big)^{2+k}\big] \leq D_1^k \text{ for all } j = 1, \ldots, p \text{ and } k = 1, 2. \tag{D.2}$$

Note that rectangles in $\mathbb{R}^p$ are invariant with respect to component-wise rescaling, i.e., for any rectangle $R = \{(x_1, \ldots, x_p) : a_j \leq x_j \leq b_j, j = 1, \ldots, p\}$, the rescaled set $R' := \{(\lambda_1 x_1, \ldots, \lambda_p x_p) : (x_1, \ldots, x_p) \in R\}$ with each $\lambda_j > 0$ is still a rectangle that can be represented as $R' = \{(x_1, \ldots, x_p) : \lambda_j a_j \leq x_j \leq \lambda_j b_j, j = 1, \ldots, p\}$. Hence one can show the following CLT by applying Theorem D.1.1 to the rescaled data:

Theorem D.2.1 *If Assumption D.2.1 holds, then*

$$\sup_{a_j \le b_j, j=1,\dots,p} \Big| P(a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le b_j \text{ for all } j) - P(a_j \le Z_j \le b_j \text{ for all } j) \Big| \le C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}}$$

where C is a universal constant.

Proof of Theorem D.2.1. Consider the rescaled data $Y_{i,j} = (X_{i,j} - \mu_j)/\sqrt{Var[X_j]}$. Due to Assumption D.2.1, $Y_{i,j}$'s satisfy Assumption D.1.1 with $b = 1$ and $B = D_1$, and has covariance structure $\Sigma_Y(j, j') = \Sigma(j, j')/\sqrt{\Sigma(j, j)\Sigma(j', j')}$. Let $\bar{Y}_j = \sum_{i=1}^n Y_{i,j}/n$. By Theorem D.1.1 we have

$$\begin{aligned}
&\sup_{a_j \le b_j, j=1,\dots,p} \Big| P(a_j \le \sqrt{n}\bar{Y}_j \le b_j \text{ for all } j) - P(a_j \le \frac{Z_j}{\sqrt{Var[X_j]}} \le b_j \text{ for all } j) \Big| \\
\le & C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}}.
\end{aligned}$$

The theorem follows from

$$\begin{aligned}
&\sup_{a_j \le b_j, j=1,\dots,p} \Big| P(a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le b_j \text{ for all } j) - P(a_j \le Z_j \le b_j \text{ for all } j) \Big| \\
= &\sup_{a_j \le b_j, j=1,\dots,p} \Big| P(\sqrt{Var[X_j]}a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le \sqrt{Var[X_j]}b_j \text{ for all } j) \\
&\qquad - P(\sqrt{Var[X_j]}a_j \le Z_j \le \sqrt{Var[X_j]}b_j \text{ for all } j) \Big| \\
= &\sup_{a_j \le b_j, j=1,\dots,p} \Big| P(a_j \le \sqrt{n}\bar{Y}_j \le b_j \text{ for all } j) - P(a_j \le \frac{Z_j}{\sqrt{Var[X_j]}} \le b_j \text{ for all } j) \Big|.
\end{aligned}$$

□

Similarly, we have the following result regarding the multiplier bootstrap:

Theorem D.2.2 *If Assumption D.2.1 holds, then for any constant $0 < \alpha < \frac{1}{e}$ we have*

$$\begin{aligned}
&\sup_{a_j \le b_j, j=1,\dots,p} \Big| P(a_j \le \widehat{Z}_j \le b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(a_j \le Z_j \le b_j \text{ for all } j) \Big| \\
\le & C\Big(\frac{D_1^2 \log^5(pn) \log^2(1/\alpha)}{n}\Big)^{\frac{1}{6}}
\end{aligned}$$

with probability at least $1 - \alpha$, where C is a universal constant.

Proof of Theorem D.2.2. Again, consider the rescaled data $Y_{i,j} = (X_{i,j} - \mu_j)/\sqrt{Var[X_j]}$. Note that the sample covariance of $Y_{i,j}$ is the same as the covariance of $(\widehat{Z}_1/\sqrt{Var[X_1]}, \ldots, \widehat{Z}_m/\sqrt{Var[X_p]})$. Theorem D.2.2 entails that

$$\begin{aligned}&\sup_{a_j \leq b_j, j=1,\ldots,p} \Big| P(a_j \leq \frac{\widehat{Z}_j}{\sqrt{Var[X_j]}} \leq b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(a_j \leq \frac{Z_j}{\sqrt{Var[X_j]}} \leq b_j \text{ for all } j) \Big| \\ \leq & C\Big(\frac{D_1^2 \log^5(pn) \log^2(1/\alpha)}{n}\Big)^{\frac{1}{6}}\end{aligned}$$

with probability at least $1 - \alpha$. The desired conclusion then follows by invariance of the class of rectangles under component-wise rescaling. □

Theorems D.2.1 and D.2.2 rely on conditions more pertinent to our context than those in Theorems D.1.1 and D.1.2. The first condition (D.1) of Assumption D.2.1 measures the ratio of the sub-exponential norm to the L_2 norm of each component of the vector, whereas the second condition (D.2) concerns the kurtosis of each component. Therefore, to guarantee a valid CLT, we need the vector to be sufficiently light-tailed after being normalized to have unit variance.

D.2.2 CLTs for Sample Means Normalized by Standard Deviations

We establish CLTs for sample means normalized by sample standard deviations, needed to prove results regarding our normalized supremum validator. Note that when the dimension p is fixed, such CLTs can be established by Slutsky's theorem, but when p is huge or grows with the data size n this is no longer applicable. Instead, we need to develop concentration inequalities for sample variances, which we state below.

Lemma D.2.3 (Concentration of sample variances) *Let $\xi_1, \ldots, \xi_n$ be n i.i.d. copies of the random variable $\xi \in \mathbb{R}$, $\sigma^2 := Var[\xi]$ be the true variance, and $\hat{\sigma}^2 = \sum_{i=1}^n (\xi_i - \bar{\xi})^2/n$, where $\bar{\xi} = \sum_{i=1}^n \xi_i/n$ is the sample mean, be the sample variance. We have the following concentration inequalities:*

1. *if ξ is $[0,1]$-valued, then there exists a universal constant C such that for any $t > 0$*

$$P(|\hat{\sigma}^2 - \sigma^2| > t) \leq 2\exp\big(-\frac{Cnt^2}{\sigma^2 + t}\big). \tag{D.3}$$

2. *if $\xi - E[\xi]$ has a sub-Gaussian norm at most K, i.e. $E[\exp((\xi - E[\xi])^2/K^2)] \leq 2$, then there exists a universal constant C such that for any $t > 0$*

$$P(|\hat{\sigma}^2 - \sigma^2| > t) \leq 4\exp(-\frac{Cnt^2}{K^4 + K^2 t}). \tag{D.4}$$

Proof of Lemma D.2.3. **Case 1:** Since the unbiased sample variance, $(n/(n-1))\hat{\sigma}^2$, is a U-statistic of degree 2, Hoeffding's inequality for U-statistics (see, e.g., Hoeffding (1963), Peel et al. (2010)) entails that with probability at least $1 - \alpha$

$$|\frac{n}{n-1}\hat{\sigma}^2 - \sigma^2| \leq \sqrt{\frac{4Var[(\xi - \xi')^2]}{n}\log\frac{2}{\alpha}} + \frac{4}{3n}\log\frac{2}{\alpha}$$

where ξ, ξ' are i.i.d. copies. Note that $Var[(\xi - \xi')^2] \leq E[(\xi - \xi')^4] \leq E[(\xi - \xi')^2] = 2\sigma^2$ because $|\xi - \xi'| \leq 1$, and $\hat{\sigma}^2 \leq 1$ for the same reason. Therefore with probability at least $1 - \alpha$

$$\begin{aligned} |\hat{\sigma}^2 - \sigma^2| &\leq \sqrt{\frac{8\sigma^2}{n}\log\frac{2}{\alpha}} + \frac{4}{3n}\log\frac{2}{\alpha} + \frac{\hat{\sigma}^2}{n-1} \\ &\leq \sqrt{\frac{8\sigma^2}{n}\log\frac{2}{\alpha}} + \frac{5}{n}\log\frac{2}{\alpha} \end{aligned}$$

and the conclusion easily follows by fixing the right hand side and solving for α.

Case 2: Since the sub-Gaussian norm of $\xi - E[\xi]$ is at most K, $(\xi - E[\xi])^2$ has a sub-exponential norm of at most K^2 by definition. Centering a variable can only inflate its sub-exponential norm by a constant factor (Remark 5.18 in Vershynin (2010)), that is, $(\xi - E[\xi])^2 - \sigma^2$ must have a sub-exponential norm of at most CK^2 where C is a universal constant. By Hoeffding's inequality and Bernstein's inequality for sums of independent variables (Propositions 5.10 and 5.16 in Vershynin

(2010)) we have for some universal constant C and any $t > 0$

$$P\big(|\frac{1}{n}\sum_{i=1}^{n}(\xi_i - E[\xi])^2 - \sigma^2| > t\big) \leq 2\exp(-\frac{Cnt^2}{K^4 + K^2 t})$$
$$P\big(|\frac{1}{n}\sum_{i=1}^{n}\xi_i - E[\xi]| > t\big) \leq 2\exp(-\frac{Cnt^2}{K^2}).$$

Note that the sample variance can be expressed as

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - E[\xi])^2 - (\frac{1}{n}\sum_{i=1}^{n}\xi_i - E[\xi])^2.$$

Hence by a union bound

$$\begin{aligned} P(|\hat{\sigma}^2 - \sigma^2| > t) &\leq P\big(|\frac{1}{n}\sum_{i=1}^{n}(\xi_i - E[\xi])^2 - \sigma^2| > t/2\big) + P\big(|\frac{1}{n}\sum_{i=1}^{n}\xi_i - E[\xi]| > \sqrt{t/2}\big) \\ &\leq 2\exp(-\frac{Cnt^2}{K^4 + K^2 t}) + 2\exp(-\frac{Cnt}{K^2}) \\ &\leq 4\exp(-\frac{Cnt^2}{K^4 + K^2 t}). \end{aligned}$$

This completes the proof. □

Note that inequality (D.3) cannot be deduced from inequality (D.4) as a special case because of the appearance of σ^2 in the bound. In fact (D.3) is a sharper bound than (D.4) when the variable ξ is Bernoulli, e.g., in the case of chance constrained optimization, because the sub-Gaussian norm of $\xi - E[\xi]$ is of order $K^2 = \Theta(\log^{-1}(1/\epsilon)) \gg \Theta(\epsilon) = \sigma^2$ when the success probability ϵ is small.

We also need the following anti-concentration inequality for Gaussian distribution:

Lemma D.2.4 (Nazarov's inequality) *Let $(Y_1, \ldots, Y_p)$ be an p-dimensional centered Gaussian random vector such that $Var[Y_j] \geq b$ for all $j = 1, \ldots, p$ and some constant $b > 0$. Then for every $-\infty \leq a_j \leq b_j \leq +\infty, j = 1, \ldots, p$ and every $\delta > 0$ it holds*

$$P(a_j - \delta \leq Y_j \leq b_j + \delta \text{ for all } j) - P(a_j \leq Y_j \leq b_j \text{ for all } j) \leq C_3\delta\sqrt{\log p}$$

where C_3 is a constant that depends only on b.

A special case of this inequality where $a_j = -\infty$ for all $j = 1, \ldots, p$ has appeared in Chernozhukov et al. (2017). Establishing a similar inequality for the case of possibly finite a_j's involves a routine application of union bound. For completeness we provide a proof here. *Proof of Lemma D.2.4.* Lemma A.1 in Chernozhukov et al. (2017) states that for every $\delta > 0$ and every $b_1, \ldots, b_p$ the following bound holds

$$P(Y_j \leq b_j + \delta \text{ for all } j) - P(Y_j \leq b_j \text{ for all } j) \leq C\delta\sqrt{\log p}$$

where C depends on b only. Applying the same bound to $-Y_j, j = 1, \ldots, p$ and $-a_j, j = 1, \ldots, p$ gives

$$P(a_j - \delta \leq Y_j \text{ for all } j) - P(a_j \leq Y_j \text{ for all } j) \leq C\delta\sqrt{\log p}.$$

Therefore

$$\begin{aligned}
& P(a_j - \delta \leq Y_j \leq b_j + \delta \text{ for all } j) - P(a_j \leq Y_j \leq b_j \text{ for all } j) \\
= \; & P(a_j - \delta \leq Y_j \leq b_j + \delta \text{ for all } j) - P(a_j - \delta \leq Y_j \leq b_j \text{ for all } j) \\
& \quad + P(a_j - \delta \leq Y_j \leq b_j \text{ for all } j) - P(a_j \leq Y_j \leq b_j \text{ for all } j) \\
\leq \; & P(Y_j \leq b_j + \delta \text{ for all } j) - P(Y_j \leq b_j \text{ for all } j) \\
& \quad + P(a_j - \delta \leq Y_j \text{ for all } j) - P(a_j \leq Y_j \text{ for all } j) \\
\leq \; & 2C\delta\sqrt{\log p}.
\end{aligned}$$

This completes the proof. □

We have the following CLT with componentwise normalization. Recall that $\hat{\sigma}_j^2$ is the sample variance of X_j computed from the data $\{X_{1,j}, \ldots, X_{n,j}\}$.

Theorem D.2.5 *Under Assumption D.2.1 we have*

$$\begin{aligned}&\sup_{a_j\le b_j, j=1,\dots,p}\Big|P(\hat\sigma_j a_j \le \sqrt{n}(\bar X_j-\mu_j)\le \hat\sigma_j b_j \text{ for all } j) - P(\sigma_j a_j \le Z_j \le \sigma_j b_j \text{ for all } j)\Big|\\ \le& C\Big(\frac{D_1^2\log^7(pn)}{n}\Big)^{\frac{1}{6}} + Cp\exp\big(-\frac{cn^{2/3}}{D_1^{10/3}}\big)\end{aligned}$$

for some universal constants C, c.

If each component of the random vector is $[0,1]$-valued, we assume:

Assumption D.2.2 *Each X_j is $[0,1]$-valued and $\sigma_j^2 := Var[X_j] \ge \delta$ for all $j = 1,\dots,p$ and some constant $\delta > 0$.*

Then we have an alternate CLT:

Theorem D.2.6 *Under Assumptions D.2.1 and D.2.2 we have*

$$\begin{aligned}&\sup_{a_j\le b_j, j=1,\dots,p}\Big|P(\hat\sigma_j a_j \le \sqrt{n}(\bar X_j-\mu_j)\le \hat\sigma_j b_j \text{ for all } j) - P(\sigma_j a_j \le Z_j \le \sigma_j b_j \text{ for all } j)\Big|\\ \le& C\Big(\frac{D_1^2\log^7(pn)}{n}\Big)^{\frac{1}{6}} + Cp\exp\big(-c\delta D_1^{2/3}n^{2/3}\big)\end{aligned}$$

for some universal constants C, c.

Proof of Theorems D.2.5 and D.2.6. For any $a_j \le b_j, j = 1,\dots,p$ and $0 < \epsilon < 1/2$

$$\begin{aligned}&P(\hat\sigma_j a_j \le \sqrt{n}(\bar X_j-\mu_j)\le \hat\sigma_j b_j \text{ for all } j)\\ =&P(\frac{\hat\sigma_j}{\sigma_j}a_j \le \frac{\sqrt{n}(\bar X_j-\mu_j)}{\sigma_j}\le \frac{\hat\sigma_j}{\sigma_j}b_j \text{ for all } j)\\ =&P(\frac{\hat\sigma_j}{\sigma_j}a_j \le \frac{\sqrt{n}(\bar X_j-\mu_j)}{\sigma_j}\le \frac{\hat\sigma_j}{\sigma_j}b_j, \big|\frac{\hat\sigma_j}{\sigma_j}-1\big|\le\epsilon \text{ for all } j)\\ &+P(\frac{\hat\sigma_j}{\sigma_j}a_j \le \frac{\sqrt{n}(\bar X_j-\mu_j)}{\sigma_j}\le \frac{\hat\sigma_j}{\sigma_j}b_j \text{ for all } j, \big|\frac{\hat\sigma_j}{\sigma_j}-1\big|>\epsilon \text{ for some } j)\\ \le&P(a_j-\epsilon|a_j| \le \frac{\sqrt{n}(\bar X_j-\mu_j)}{\sigma_j}\le b_j+\epsilon|b_j| \text{ for all } j) + P(\big|\frac{\hat\sigma_j}{\sigma_j}-1\big|>\epsilon \text{ for some } j)\\ \le&P(a_j-\epsilon|a_j| \le \frac{Z_j}{\sigma_j}\le b_j+\epsilon|b_j| \text{ for all } j) + C\Big(\frac{D_1^2\log^7(pn)}{n}\Big)^{\frac{1}{6}} + \sum_{j=1}^p P(\big|\hat\sigma_j-\sigma_j\big|>\epsilon\sigma_j)\end{aligned}$$

where in the last inequality we use Theorem D.2.1 for the first probability and apply the union bound to the second probability. Note that $\epsilon < 1/2$ hence for any constant $M > 0$ if we denote by $f_M(x) = -\infty \cdot \mathbf{1}(x < -M) + \infty \cdot \mathbf{1}(x > M)$ then we have

$$\begin{aligned}
&P(a_j - \epsilon |a_j| \le \frac{Z_j}{\sigma_j} \le b_j + \epsilon |b_j| \text{ for all } j) \\
\le &P(f_M(a_j) + a_j - \epsilon |a_j| \le \frac{Z_j}{\sigma_j} \le f_M(b_j) + b_j + \epsilon |b_j| \text{ for all } j) + 2p\exp(-cM^2) \\
\le &P(f_M(a_j) + a_j \le \frac{Z_j}{\sigma_j} \le f_M(b_j) + b_j \text{ for all } j) + C\epsilon M\sqrt{\log p} + 2p\exp(-cM^2) \\
\le &P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j) + 4p\exp(-cM^2) + C\epsilon M\sqrt{\log p}
\end{aligned}$$

where in the second inequality we use Lemma D.2.4 (note that $f_M(a_j) + a_j - \epsilon |a_j|$ is either ∞ or its absolute value $\le \frac{3}{2}M$, so is $f_M(b_j) + b_j + \epsilon |b_j|$), the term $\exp(-cM^2)$ is the tail bound of the univariate standard normal distribution, and C, c are universal constants. Therefore we have derived the following upper bound

$$\begin{aligned}
&P(\hat{\sigma}_j a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le \hat{\sigma}_j b_j \text{ for all } j) \\
\le &P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j) + 4p\exp(-cM^2) + C\epsilon M\sqrt{\log p} + C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} \\
&+ \sum_{j=1}^{p} P(\big|\hat{\sigma}_j - \sigma_j\big| > \epsilon\sigma_j)
\end{aligned}$$

which holds true for $\epsilon < 1/2$ and $M > 0$. Similarly, one can show

$$\begin{aligned}
&P(\hat{\sigma}_j a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le \hat{\sigma}_j b_j \text{ for all } j) \\
\ge &P(a_j + \epsilon |a_j| \le \frac{Z_j}{\sigma_j} \le b_j - \epsilon |b_j| \text{ for all } j) - C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} - \sum_{j=1}^{p} P(\big|\hat{\sigma}_j - \sigma_j\big| > \epsilon\sigma_j)
\end{aligned}$$

along with

$$\begin{aligned}
&P(a_j + \epsilon\,|a_j| \le \frac{Z_j}{\sigma_j} \le b_j - \epsilon\,|b_j| \text{ for all } j)\\
\ge &P(f_M(a_j) + a_j + \epsilon\,|a_j| \le \frac{Z_j}{\sigma_j} \le f_M(b_j) + b_j - \epsilon\,|b_j| \text{ for all } j) - 2p\exp(-cM^2)\\
\ge &P(f_M(a_j) + a_j \le \frac{Z_j}{\sigma_j} \le f_M(b_j) + b_j \text{ for all } j) - C\epsilon M\sqrt{\log p} - 2p\exp(-cM^2)\\
\ge &P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j) - C\epsilon M\sqrt{\log p} - 4p\exp(-cM^2).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\big|P(\hat{\sigma}_j a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le \hat{\sigma}_j b_j \text{ for all } j) - P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j)\big|\\
\le &4p\exp(-cM^2) + C\epsilon M\sqrt{\log p} + C\Big(\frac{D_1^2\log^7(pn)}{n}\Big)^{\frac{1}{6}} + \sum_{j=1}^{p} P(\big|\hat{\sigma}_j - \sigma_j\big| > \epsilon\sigma_j)\\
\le &4p\exp(-cM^2) + C\epsilon M\sqrt{\log p} + C\Big(\frac{D_1^2\log^7(pn)}{n}\Big)^{\frac{1}{6}} + \sum_{j=1}^{p} P(\big|\hat{\sigma}_j^2 - \sigma_j^2\big| > \epsilon\sigma_j^2)\\
\le &4p\exp(-cM^2) + C\epsilon M\sqrt{\log p} + C\Big(\frac{D_1^2\log^7(pn)}{n}\Big)^{\frac{1}{6}} + 4p\exp(-\frac{cn\epsilon^2}{D_1^4 + D_1^2\epsilon})
\end{aligned} \tag{D.5}$$

where the last inequality holds because Assumption D.2.1 guarantees that the sub-Gaussian norm of $X_j - \mu_j$ is at most $D_1\sigma_j$ and one then applies Lemma D.2.3. Now set

$$M = \sqrt{\frac{1}{c}\log(pn)},\ \epsilon = \Big(\frac{D_1^2}{n}\Big)^{\frac{1}{6}}$$

and note that $\epsilon < 1/2$ can be assumed since otherwise the first term in the desired bound is already greater than 1 (by enlarging the universal constant if necessary) and the bound is trivial. We get

the uniform bound

$$\begin{aligned}
&\big|P(\hat{\sigma}_j a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le \hat{\sigma}_j b_j \text{ for all } j) - P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j)\big| \\
\le& C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + Cp\exp\Big(-\frac{cD_1^{2/3}n^{2/3}}{D_1^4 + D_1^2\epsilon}\Big) \\
\le& C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + Cp\exp\Big(-\frac{cn^{2/3}}{D_1^{10/3}}\Big)
\end{aligned}$$

where the second inequality holds because $D_2 \ge 1$ and $\epsilon < 1/2$. In particular, if X_j's are $[0,1]$-valued, we use the concentration inequality (D.3) instead of (D.4) to refine the bound (D.5) to be

$$\begin{aligned}
&\big|P(\hat{\sigma}_j a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le \hat{\sigma}_j b_j \text{ for all } j) - P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j)\big| \\
\le& 4p\exp(-cM^2) + C\epsilon M\sqrt{\log p} + C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + 4p\exp(-\frac{cn\delta\epsilon^2}{1+\epsilon}).
\end{aligned}$$

Likewise, letting M and ϵ take the same values as before, we obtain

$$\begin{aligned}
&\big|P(\hat{\sigma}_j a_j \le \sqrt{n}(\bar{X}_j - \mu_j) \le \hat{\sigma}_j b_j \text{ for all } j) - P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j)\big| \\
\le& C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + Cp\exp\Big(-c\delta D_1^{2/3}n^{2/3}\Big).
\end{aligned}$$

This completes the proof of Theorems D.2.5 and D.2.6. □

The following are corresponding results for the multiplier bootstrap:

Theorem D.2.7 *If Assumption D.2.1 holds, for any constant $0 < \alpha < \frac{1}{e}$ we have*

$$\begin{aligned}
&\sup_{a_j \le b_j, j=1,\dots,p} \Big|P(\hat{\sigma}_j a_j \le \widehat{Z}_j \le \hat{\sigma}_j b_j \textit{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(\sigma_j a_j \le Z_j \le \sigma_j b_j \textit{ for all } j)\Big| \\
\le& C\Big(\frac{D_1^2 \log^5(pn)\log^2(1/\alpha)}{n}\Big)^{\frac{1}{6}} + \frac{CD_1^2 \log(pn)\log(p/\alpha)}{\sqrt{n}}
\end{aligned}$$

with probability at least $1-\alpha$, where C is a universal constant.

Theorem D.2.8 *If Assumptions D.2.1 and D.2.2 hold, for any constant* $0 < \alpha < \frac{1}{e}$ *we have*

$$\begin{aligned}&\sup_{a_j \le b_j, j=1,\dots,p} \Big| P(\hat{\sigma}_j a_j \le \widehat{Z}_j \le \hat{\sigma}_j b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(\sigma_j a_j \le Z_j \le \sigma_j b_j \text{ for all } j) \Big| \\ \le & C\Big(\frac{D_1^2 \log^5(pn) \log^2(1/\alpha)}{n}\Big)^{\frac{1}{6}} + \frac{C \log(pn) \log(p/\alpha)}{\sqrt{n\delta}}\end{aligned}$$

with probability at least $1 - \alpha$*, where* C *is a universal constant.*

Proof of Theorems D.2.7 and D.2.8. For any $a_j \le b_j, j = 1, \dots, p$ and $0 < \epsilon < 1/2$

$$\begin{aligned}&P(\hat{\sigma}_j a_j \le \widehat{Z}_j \le \hat{\sigma}_j b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) \\ =&P(\frac{\hat{\sigma}_j}{\sigma_j} a_j \le \frac{\widehat{Z}_j}{\sigma_j} \le \frac{\hat{\sigma}_j}{\sigma_j} b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) \\ =&P(\frac{\hat{\sigma}_j}{\sigma_j} a_j \le \frac{\widehat{Z}_j}{\sigma_j} \le \frac{\hat{\sigma}_j}{\sigma_j} b_j, \Big|\frac{\hat{\sigma}_j}{\sigma_j} - 1\Big| \le \epsilon \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) \\ &+ P(\frac{\hat{\sigma}_j}{\sigma_j} a_j \le \frac{\widehat{Z}_j}{\sigma_j} \le \frac{\hat{\sigma}_j}{\sigma_j} b_j \text{ for all } j, \Big|\frac{\hat{\sigma}_j}{\sigma_j} - 1\Big| > \epsilon \text{ for some } j | \{\mathbf{X}_i\}_{i=1}^n) \\ \le &P(a_j - \epsilon |a_j| \le \frac{\widehat{Z}_j}{\sigma_j} \le b_j + \epsilon |b_j|, \Big|\frac{\hat{\sigma}_j}{\sigma_j} - 1\Big| \le \epsilon \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) \\ &\text{with probability at least } 1 - \sum_{j=1}^p P(\big|\hat{\sigma}_j - \sigma_j\big| > \epsilon \sigma_j) \\ \le &P(a_j - \epsilon |a_j| \le \frac{\widehat{Z}_j}{\sigma_j} \le b_j + \epsilon |b_j| \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) \\ \le &P(a_j - \epsilon |a_j| \le \frac{Z_j}{\sigma_j} \le b_j + \epsilon |b_j| \text{ for all } j) + C\Big(\frac{D_1^2 \log^5(pn) \log^2(4/\alpha)}{n}\Big)^{\frac{1}{6}} \\ &\text{with probability at least } 1 - \frac{\alpha}{4}.\end{aligned}$$

In the proof of Theorems D.2.5 and D.2.6 we show

$$\begin{aligned}&P(a_j - \epsilon |a_j| \le \frac{Z_j}{\sigma_j} \le b_j + \epsilon |b_j| \text{ for all } j) \\ \le &P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j) + 4p \exp(-cM^2) + C\epsilon M \sqrt{\log p}.\end{aligned}$$

Similarly we can show the other direction

$$
\begin{aligned}
&P(\hat{\sigma}_j a_j \leq \widehat{Z}_j \leq \hat{\sigma}_j b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) \\
\geq & P(a_j + \epsilon |a_j| \leq \frac{Z_j}{\sigma_j} \leq b_j - \epsilon |b_j| \text{ for all } j) - C\Big(\frac{D_1^2 \log^5(pn) \log^2(4/\alpha)}{n}\Big)^{\frac{1}{6}} \\
&\text{with probability at least } 1 - \frac{\alpha}{4} - \sum_{j=1}^p P(\big|\hat{\sigma}_j - \sigma_j\big| > \epsilon \sigma_j) \\
\geq & P(a_j \leq \frac{Z_j}{\sigma_j} \leq b_j \text{ for all } j) - C\epsilon M\sqrt{\log p} - 4p\exp(-cM^2) - C\Big(\frac{D_1^2 \log^5(pn) \log^2(4/\alpha)}{n}\Big)^{\frac{1}{6}}.
\end{aligned}
$$

Therefore the following uniform bound holds with probability at least $1 - \frac{\alpha}{2} - 2\sum_{j=1}^p P(\big|\hat{\sigma}_j - \sigma_j\big| > \epsilon\sigma_j)$

$$
\begin{aligned}
&\big|P(\hat{\sigma}_j a_j \leq \widehat{Z}_j \leq \hat{\sigma}_j b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(a_j \leq \frac{Z_j}{\sigma_j} \leq b_j \text{ for all } j)\big| \\
\leq & C\Big(\frac{D_1^2 \log^5(pn) \log^2(4/\alpha)}{n}\Big)^{\frac{1}{6}} + C\epsilon M\sqrt{\log p} + 4p\exp(-cM^2).
\end{aligned}
$$

Note that

$$
2\sum_{j=1}^p P(\big|\hat{\sigma}_j - \sigma_j\big| > \epsilon\sigma_j) \leq 8p\exp(-\frac{cn\epsilon^2}{D_1^4 + D_1^2\epsilon}).
$$

By setting the right hand side of the above inequality to be $\alpha/2$ and $M = \sqrt{\frac{1}{c}\log(pn)}$ we get

$$
\begin{aligned}
&\big|P(\hat{\sigma}_j a_j \leq \widehat{Z}_j \leq \hat{\sigma}_j b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(a_j \leq \frac{Z_j}{\sigma_j} \leq b_j \text{ for all } j)\big| \\
\leq & C\Big(\frac{D_1^2 \log^5(pn) \log^2(4/\alpha)}{n}\Big)^{\frac{1}{6}} + C\Big(\sqrt{\frac{D_1^4}{n}\log\frac{16p}{\alpha}} + \frac{D_1^2}{n}\log\frac{16p}{\alpha}\Big)\log(pn) \\
\leq & C\Big(\frac{D_1^2 \log^5(pn) \log^2(4/\alpha)}{n}\Big)^{\frac{1}{6}} + C\frac{D_1^2}{\sqrt{n}}\log\frac{16p}{\alpha}\log(pn)
\end{aligned}
$$

with probability at least $1 - \alpha$.

In case of $[0,1]$-valued variables, we use (D.3) instead of (D.4) to get

$$\begin{aligned}
&\left|P(\hat{\sigma}_j a_j \le \widehat{Z}_j \le \hat{\sigma}_j b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(a_j \le \frac{Z_j}{\sigma_j} \le b_j \text{ for all } j)\right| \\
\le & C\Big(\frac{D_1^2 \log^5(pn) \log^2(4/\alpha)}{n}\Big)^{\frac{1}{6}} + C\Big(\sqrt{\frac{1}{n\delta}\log\frac{16p}{\alpha}} + \frac{1}{n\delta}\log\frac{16p}{\alpha}\Big)\log(pn) \\
\le & C\Big(\frac{D_1^2 \log^5(pn) \log^2(4/\alpha)}{n}\Big)^{\frac{1}{6}} + C\Big(\frac{1}{\sqrt{n\delta}} + \frac{1}{n\delta}\Big)\log\frac{16p}{\alpha}\log(pn)
\end{aligned}$$

with probability at least $1-\alpha$. We can assume that $n\delta \ge 1$ to get the desired bound since otherwise the second term in the bound is already greater than 1. □

D.2.3 Coverage Probability through Multiplier Bootstrap

Theorem D.2.9 (Coverage probability for unnormalized supremum) *If Assumption D.2.1 holds, for every $0 < \beta < 1$ we have*

$$|P(\sqrt{n}(\bar{X}_j - \mu_j) \le \hat{q}_{1-\beta} \text{ for all } j) - (1-\beta)| \le C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}}$$

where $\hat{q}_{1-\beta}$ is such that

$$P(\widehat{Z}_j \le \hat{q}_{1-\beta} \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) = 1-\beta$$

and C is a universal constant.

Proof of Theorem D.2.9. Denote by $\epsilon = C\Big(\frac{D_1^2 \log^5(pn)\log^2(1/\alpha)}{n}\Big)^{1/6}$ and by A_α the event that

$$\sup_{a_j \le b_j, j=1,\dots,p} \Big|P(a_j \le \widehat{Z}_j \le b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(a_j \le Z_j \le b_j \text{ for all } j)\Big| \le \epsilon.$$

Then we can rewrite

$$\begin{aligned}
&P(\sqrt{n}(\bar{X}_j - \mu_j) \le \hat{q}_{1-\beta} \text{ for all } j) \\
=&P(\sqrt{n}(\bar{X}_j - \mu_j) \le \hat{q}_{1-\beta} \text{ for all } j, \text{ and } A_\alpha) + P(\sqrt{n}(\bar{X}_j - \mu_j) \le \hat{q}_{1-\beta} \text{ for all } j, \text{ and } A_\alpha^c)
\end{aligned}$$

The second term is bounded by α because of Theorem D.2.2. To study the first term, denote by $q_{1-\beta}$ the true $(1-\beta)$-level quantile of the limit distribution, i.e., $q_{1-\beta}$ is such that $P(Z_j \leq q_{1-\beta} \text{ for all } j) = 1-\beta$. On event A_α we have $q_{1-\beta-\epsilon} \leq \hat{q}_{1-\beta} \leq q_{1-\beta+\epsilon}$, therefore

$$\begin{aligned} P(\sqrt{n}(\bar{X}_j - \mu_j) \leq q_{1-\beta-\epsilon} \text{ for all } j, \text{ and } A_\alpha) &\leq P(\sqrt{n}(\bar{X}_j - \mu_j) \leq \hat{q}_{1-\beta} \text{ for all } j, \text{ and } A_\alpha) \\ &\leq P(\sqrt{n}(\bar{X}_j - \mu_j) \leq q_{1-\beta+\epsilon} \text{ for all } j, \text{ and } A_\alpha). \end{aligned}$$

From this two-sided bound we get

$$\begin{aligned} & P(\sqrt{n}(\bar{X}_j - \mu_j) \leq \hat{q}_{1-\beta} \text{ for all } j) \\ \leq & P(\sqrt{n}(\bar{X}_j - \mu_j) \leq q_{1-\beta+\epsilon} \text{ for all } j) + \alpha \\ \leq & P(Z_j \leq q_{1-\beta+\epsilon} \text{ for all } j) + C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{1/6} + \alpha \\ = & 1 - \beta + \epsilon + C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{1/6} + \alpha. \end{aligned}$$

Similarly the lower bound can be derived as

$$\begin{aligned} & P(\sqrt{n}(\bar{X}_j - \mu_j) \leq \hat{q}_{1-\beta} \text{ for all } j) \\ \geq & P(\sqrt{n}(\bar{X}_j - \mu_j) \leq q_{1-\beta-\epsilon} \text{ for all } j, \text{ and } A_\alpha) \\ = & P(\sqrt{n}(\bar{X}_j - \mu_j) \leq q_{1-\beta+\epsilon} \text{ for all } j) - P(\sqrt{n}(\bar{X}_j - \mu_j) \leq q_{1-\beta-\epsilon} \text{ for all } j, \text{ and } A_\alpha^c) \\ \geq & P(Z_j \leq q_{1-\beta-\epsilon} \text{ for all } j) - C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{1/6} - \alpha \\ = & 1 - \beta - \epsilon - C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{1/6} - \alpha. \end{aligned}$$

This gives the following bound for any $\alpha < 1/e$

$$|P(\sqrt{n}(\bar{X}_j - \mu_j) \leq \hat{q}_{1-\beta} \text{ for all } j) - (1-\beta)| \leq C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + \epsilon + \alpha.$$

Set $\alpha = 1/n$ and note that $1/n$ is less than the leading term, thus we have shown the desired conclusion. □

Theorem D.2.10 (Coverage probability for normalized supremum) *If Assumptions D.2.1 holds, for every $0 < \beta < 1$ we have*

$$\begin{aligned}&|P(\sqrt{n}(\bar{X}_j - \mu_j) \leq \hat{\sigma}_j \hat{q}_{1-\beta} \text{ for all } j) - (1-\beta)| \\ \leq& C\Big(\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + \frac{D_1^2 \log^2(pn)}{\sqrt{n}} + p\exp\big(-\frac{cn^{2/3}}{D_1^{10/3}}\big)\Big).\end{aligned}$$

If Assumption D.2.2 also holds, then

$$\begin{aligned}&|P(\sqrt{n}(\bar{X}_j - \mu_j) \leq \hat{\sigma}_j \hat{q}_{1-\beta} \text{ for all } j) - (1-\beta)| \\ \leq& C\Big(\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + \frac{\log^2(pn)}{\sqrt{n\delta}} + p\exp\big(-c\delta D_1^{2/3} n^{2/3}\big)\Big).\end{aligned}$$

Here $\hat{q}_{1-\beta}$ is such that

$$P(\widehat{Z}_j \leq \hat{\sigma}_j \hat{q}_{1-\beta} \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) = 1 - \beta$$

and C, c are universal constants.

Proof of Theorem D.2.10. Let $\epsilon = C\Big(\frac{D_1^2 \log^5(pn)\log^2(1/\alpha)}{n}\Big)^{1/6} + \frac{CD_1^2 \log(pn)\log(p/\alpha)}{\sqrt{n}}$ and A_α be the event that

$$\sup_{a_j \leq b_j, j=1,\ldots,p} \Big| P(\hat{\sigma}_j a_j \leq \widehat{Z}_j \leq \hat{\sigma}_j b_j \text{ for all } j | \{\mathbf{X}_i\}_{i=1}^n) - P(\sigma_j a_j \leq Z_j \leq \sigma_j b_j \text{ for all } j)\Big| \leq \epsilon.$$

We know that $P(A_\alpha) \leq \alpha$ from Theorem D.2.7. Following the same line of the proof for Theorem D.2.9 and using the CLT in Theorem D.2.5 we can derive that

$$|P(\sqrt{n}(\bar{X}_j - \mu_j) \leq \hat{\sigma}_j \hat{q}_{1-\beta} \text{ for all } j) - (1-\beta)| \leq C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + Cp\exp\big(-\frac{cn^{2/3}}{D_1^{10/3}}\big) + \epsilon + \alpha.$$

Again setting $\alpha = \frac{1}{n}$ leads to the first bound.

The second bound can be derived similarly. Let $\epsilon = C\Big(\frac{D_1^2 \log^5(pn)\log^2(1/\alpha)}{n}\Big)^{1/6} + \frac{C\log(pn)\log(p/\alpha)}{\sqrt{n\delta}}$,

and now Theorem D.2.8 entails that $P(A_\alpha) \leq \alpha$ again. Using the CLT in Theorem D.2.6 gives

$$\begin{aligned}&|P(\sqrt{n}(\bar{X}_j - \mu_j) \leq \hat{\sigma}_j \hat{q}_{1-\beta} \text{ for all } j) - (1-\beta)| \\ \leq& C\Big(\frac{D_1^2 \log^7(pn)}{n}\Big)^{\frac{1}{6}} + Cp\exp\big(-c\delta D_1^{2/3} n^{2/3}\big) + \epsilon + \alpha.\end{aligned}$$

The second bound follows from setting $\alpha = \frac{1}{n}$. □

D.2.4 Proofs of Main Statistical Guarantees

We now put together all the previous results to prove the statistical guarantees of our validators. For convenience, we suppress the subscript $\boldsymbol{\xi}_{1:n_2}$ in the probability notation.

Proof of Theorem 5.3.1. We bound the probability as follows

$$\begin{aligned}&P(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \\ \geq& P(\hat{H}_j \geq \gamma + \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for some } j = 1, \ldots, p \text{ in (5.8) and } H(x^*(s_j)) \geq \hat{H}_j - \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) \\ \geq& P(H(x^*(s_j)) \geq \hat{H}_j - \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) - P(\hat{H}_j < \gamma + \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) \\ \geq& 1 - \beta - C\Big(\frac{D_1^2 \log^7(pn_2)}{n_2}\Big)^{\frac{1}{6}} - P(\hat{H}_j < \gamma + \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) \qquad \text{(D.6)}\end{aligned}$$

where we use Theorem D.2.9 for the first probability by letting $X_{i,j} = h(x^*(s_j), \xi_i)$. To bound the second probability, we recall that $\hat{\sigma}_j^2$ is the sample variance of $\{h(x^*(s_j), \xi_1), \ldots, h(x^*(s_j), \xi_{n_2})\}$ and

write

$$
\begin{aligned}
& P(\hat{H}_j < \gamma + \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) \\
\leq\ & P(\hat{H}_j < \gamma + \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ and } \hat{\sigma}_j^2 \leq 2\sigma^2(x^*(s_j)) \text{ for all } j = 1, \ldots, p) \\
& \quad + P(\hat{\sigma}_j^2 > 2\sigma^2(x^*(s_j)) \text{ for some } j = 1, \ldots, p) \\
\leq\ & P(\hat{H}_j < \gamma + \frac{C\bar{\sigma}\sqrt{\log(p/\beta)}}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) + P(\hat{\sigma}_j^2 > 2\sigma^2(x^*(s_j)) \text{ for some } j = 1, \ldots, p) \\
& \quad \text{because of the fact that } q_{1-\beta} \leq C \max_j \hat{\sigma}_j \sqrt{\log(p/\beta)} \text{ for some universal constant } C \\
\leq\ & P(\hat{H}_{\bar{j}} - \overline{H} < \gamma + \frac{C\bar{\sigma}\sqrt{\log(p/\beta)}}{\sqrt{n_2}} - \overline{H}) + Cp\exp(-\frac{cn_2}{D_1^4}) \\
& \quad \text{where } \bar{j} \text{ is the index such that } H(x^*(s_{\bar{j}})) = \overline{H} \text{ and the concentration (D.4) is used} \\
\leq\ & C\exp\big(-\frac{cn_2\epsilon^2}{D_1^2\bar{\sigma}^2}\big) + Cp\exp(-\frac{cn_2}{D_1^4}) \\
& \quad \text{because the sub-Gaussian norm of } h(x^*(s_{\bar{j}}), \xi) \text{ is at most } D_1\bar{\sigma}.
\end{aligned}
$$

Substituting this bound into (D.6) gives the desired conclusion. □

Proof of Theorem 5.3.2. Similar to the proof of Theorem 5.3.1, we have the bound

$$
\begin{aligned}
& P(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \\
\geq\ & P(H(x^*(s_j)) \geq \hat{H}_j - \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) - P(\hat{H}_j < \gamma + \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) \\
\geq\ & 1 - \beta - C\left(\Big(\frac{D_1^2\log^7(pn_2)}{n_2}\Big)^{\frac{1}{6}} + \frac{D_1^2\log^2(pn_2)}{\sqrt{n_2}} + p\exp\big(-\frac{cn_2^{2/3}}{D_1^{10/3}}\big)\right) - \\
& \quad P(\hat{H}_j < \gamma + \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) \\
& \quad \text{where the first bound is due to Theorem D.2.10.}
\end{aligned}
$$

For the second probability we write

$$\begin{aligned}
& P(\hat{H}_j < \gamma + \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) \\
\leq\ & P(\hat{H}(x^*(\bar{s})) < \gamma + \frac{q_{1-\beta}\hat{\sigma}(x^*(\bar{s}))}{\sqrt{n_2}}) \\
\leq\ & P(\hat{H}(x^*(\bar{s})) < \gamma + \frac{q_{1-\beta}\sqrt{2}\sigma(x^*(\bar{s}))}{\sqrt{n_2}}) + P(\hat{\sigma}^2(x^*(\bar{s})) > 2\sigma^2(x^*(\bar{s}))) \\
\leq\ & P(\hat{H}(x^*(\bar{s})) < \gamma + \frac{C\sqrt{\log(p/\beta)}\sigma(x^*(\bar{s}))}{\sqrt{n_2}}) + P(\hat{\sigma}^2(x^*(\bar{s})) > 2\sigma^2(x^*(\bar{s}))) \\
& \text{since } q_{1-\beta} \leq C\sqrt{\log(p/\beta)} \\
\leq\ & C\exp\big(-\frac{cn_2\epsilon^2}{D_1^2\sigma^2(x^*(\bar{s}))}\big) + C\exp(-\frac{cn_2}{D_1^4})
\end{aligned}$$

Combining the two probability bounds and noting that $p\exp\big(-\frac{cn_2^{2/3}}{D_1^{10/3}}\big)$ dominates $\exp(-\frac{cn_2}{D_1^4})$ (because $\frac{n_2}{D_1^4} = \frac{n_2^{2/3}}{D_1^{10/3}} \cdot \big(\frac{n_2}{D_1^2}\big)^{1/3}$ and $\frac{n_2}{D_1^2} \geq 1$ can be assumed), we obtain the desired conclusion. □

Proof of Corollary 5.3.3. When $\overline{H} > \gamma$, we have $\epsilon \to \overline{H} - \gamma > 0$ in Theorems 5.3.1 and 5.3.2, therefore the exponential error term with ϵ vanishes as $n_2 \to \infty$. Under the condition that $p\exp(-n_2^{1/7}) \to 0$ it is straightforward to check that other error terms also vanish. □

Proof of Theorem 5.3.4. Unlike the proof of Theorem 5.3.1, we use the Bernoulli structure to derive the error bound. Note that in this case $\gamma = 1 - \alpha$. Define events

$$\begin{aligned}
E_1 &= \big\{\hat{H}_j \geq 1 - \alpha + \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for some } j = 1, \ldots, p \text{ in (5.8)}\big\} \\
E_2 &= \big\{H(x^*(s_j)) \geq \hat{H}_j - \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j \text{ such that } H(x^*(s_j)) \in (\alpha, 1-\alpha)\big\} \\
E_3 &= \big\{\hat{H}_j < 1 - \alpha + \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j \text{ such that } H(x^*(s_j)) \leq \alpha\big\}.
\end{aligned}$$

Then we have

$$\begin{aligned}
P(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) &\geq P(E_1 \cap E_2 \cap E_3) \\
&\geq 1 - P(E_1^c) - P(E_2^c) - P(E_3^c) \\
&= P(E_2) - P(E_1^c) - P(E_3^c). \qquad \text{(D.7)}
\end{aligned}$$

We bound the three probabilities. Let $q^{\alpha}_{1-\beta}$ be the $1-\beta$ quantile of $\max\{Z_j : H(x^*(s_j)) \in (\alpha, 1-\alpha), 1 \le j \le p\}$ where $(Z_1, \ldots, Z_p) \sim N_p(0, \hat{\Sigma})$. By stochastic dominance it is clear that $q^{\alpha}_{1-\beta} \le q_{1-\beta}$ almost surely, therefore

$$\begin{aligned} P(E_2) \ge P\big(H(x^*(s_j)) &\ge \hat{H}_j - \frac{q^{\alpha}_{1-\beta}}{\sqrt{n_2}} \text{ for all } j \text{ such that } H(x^*(s_j)) \in (\alpha, 1-\alpha)\big) \\ &\ge 1 - \beta - C\Big(\frac{\log^7(pn_2)}{n_2\alpha}\Big)^{\frac{1}{6}} \end{aligned}$$

by applying Theorem D.2.9 to $\{h(x^*(s_j), \xi) : H(x^*(s_j)) \in (\alpha, 1-\alpha), 1 \le j \le p\}$ and noticing that Assumption D.2.1 is satisfied with $D_1 = \frac{C}{\sqrt{\alpha}}$ for some universal constant C.

We then bound the second probability

$$\begin{aligned} P(E_1^c) &= P(\hat{H}_j < 1 - \alpha + \frac{q_{1-\beta}}{\sqrt{n_2}} \text{ for all } j = 1, \ldots, p) \\ &\le P(\hat{H}_{\bar{j}} < 1 - \alpha + \frac{q_{1-\beta}}{\sqrt{n_2}}) \text{ where } \bar{j} \text{ is the index such that } H(x^*(s_{\bar{j}})) = 1 - \bar{\alpha} \\ &\le P(\hat{H}_{\bar{j}} < 1 - \alpha + \frac{C\sqrt{\log(p/\beta)}}{\sqrt{n_2}}) \text{ because } q_{1-\beta} \le C \max_j \hat{\sigma}_j \sqrt{\log(p/\beta)} \le C\sqrt{\log(p/\beta)} \\ &\le \exp\big(-\frac{n_2\epsilon^2}{2(\bar{\alpha}(1-\bar{\alpha}) + \epsilon/3)}\big) \end{aligned}$$

where in the last line we use a Bernstein's inequality for sums of bounded random variables (see equation (2.10) in Boucheron et al. (2013)). Note that this is further bounded by $\exp\big(-cn_2 \min\{\epsilon, \frac{\epsilon^2}{\bar{\alpha}}\}\big)$ if $\bar{\alpha} \le 1/2$.

The third probability can be bounded as

$$\begin{aligned} P(E_3^c) &\le P\big(\hat{H}_j \ge 1 - \alpha \text{ for some } j \text{ such that } H(x^*(s_j)) \le \alpha\big) \\ &\le \sum_{j: H(x^*(s_j)) \le \alpha} P(\hat{H}_j \ge 1 - \alpha) \\ &\le p\exp(-2n_2(1-2\alpha)^2) \le p\exp(-cn_2) \text{ by Hoeffding's inequality.} \end{aligned}$$

Substituting the bounds into (D.7) leads to

$$P(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \leq 1-\beta-C\Big(\frac{\log^7(pn_2)}{n_2\alpha}\Big)^{\frac{1}{6}} - \exp\big(-cn_2\min\{\epsilon, \frac{\epsilon^2}{\bar{\alpha}}\}\big) - p\exp(-cn_2).$$

It remains to show that $p\exp(-cn_2)$ is negligible relative to other error terms. Since $\alpha < 1$ it is clear that $\big(\frac{1}{n_2}\big)^{1/6} \leq \Big(\frac{\log^7(pn_2)}{n_2\alpha}\Big)^{1/6}$, and we argue that $\big(\frac{1}{n_2}\big)^{1/6} \geq p\exp(-cn_2)$ can be assumed so that $p\exp(-cn_2) \leq \Big(\frac{\log^7(pn_2)}{n_2\alpha}\Big)^{1/6}$. If $\big(\frac{1}{n_2}\big)^{1/6} < p\exp(-cn_2)$, then $p > \exp(cn_2)n_2^{-1/6}$, and $\frac{\log^7(pn_2)}{n_2\alpha} \geq \frac{(cn_2)^7}{n_2\alpha} \geq c^7 n_2^6$, hence the first error term already exceeds 1 (enlarge the universal constant C if necessary) and the error bound holds true trivially. □

Proof of Theorem 5.3.5. The proof follows the one for Theorem 5.3.4, and we focus on the modifications. The events are now defined as

$$\begin{aligned}
E_1 &= \big\{\hat{H}_j \geq 1-\alpha+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} \text{ for some } j=1,\ldots,p \text{ in (5.9)}\big\}\\
E_2 &= \big\{H(x^*(s_j)) \geq \hat{H}_j - \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} \text{ for all } j \text{ such that } H(x^*(s_j)) \in (\alpha, 1-\alpha)\big\}\\
E_3 &= \big\{\hat{H}_j < 1-\alpha+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} \text{ for all } j \text{ such that } H(x^*(s_j)) \leq \alpha\big\}.
\end{aligned}$$

Again we have $P(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \geq P(E_2) - P(E_1^c) - P(E_3^c)$.

The first probability bound becomes

$$P(E_2) \geq 1-\beta-C\Big(\Big(\frac{\log^7(pn_2)}{n_2\alpha}\Big)^{\frac{1}{6}} + \frac{\log^2(pn_2)}{\sqrt{n_2\alpha}} + p\exp\big(-c(n_2\alpha)^{2/3}\big)\Big)$$

by using the second half of Theorem D.2.10 and noting that $\delta = \alpha(1-\alpha) \geq \frac{1}{2}\alpha$ if $\alpha < \frac{1}{2}$ and

$D_1 = \frac{C}{\sqrt{\alpha}}$. For the second probability we have

$$\begin{aligned}
P(E_1^c) &\leq P(\hat{H}_{\bar{j}} < 1-\alpha + \frac{q_{1-\beta}\hat{\sigma}_{\bar{j}}}{\sqrt{n_2}}) \text{ where } \bar{j} \text{ is the index such that } H(x^*(s_{\bar{j}})) = 1-\bar{\alpha} \\
&\leq P(\hat{H}_{\bar{j}} < 1-\alpha + \frac{q_{1-\beta}t}{\sqrt{n_2}}) + P(\hat{\sigma}_{\bar{j}} > t) \text{ where } t = \sqrt{\bar{\alpha}(1-\bar{\alpha})} + \sqrt{2\log(n_2\alpha)/n_2} \\
&\leq P(\hat{H}_{\bar{j}} < 1-\alpha + \frac{q_{1-\beta}t}{\sqrt{n_2}}) + \frac{1}{n_2\alpha} \\
&\quad \text{where the bound } \frac{1}{n_2\alpha} \text{ is by (D.3) (see Theorem 10 in Maurer and Pontil (2009))} \\
&\leq P(\hat{H}_{\bar{j}} < 1-\alpha + \frac{C\sqrt{(\bar{\alpha} + \log(n_2\alpha)/n_2)\log(p/\beta)}}{\sqrt{n_2}}) + \frac{1}{n_2\alpha} \\
&\quad \text{because } q_{1-\beta} \leq C\sqrt{\log(p/\beta)} \\
&\leq \exp\big(-\frac{n_2\epsilon^2}{2(\bar{\alpha}(1-\bar{\alpha}) + \epsilon/3)}\big) + \frac{1}{n_2\alpha} \text{ by Bernstein's inequality.}
\end{aligned}$$

Whereas for the third probability we still have $P(E_3^c) \leq p\exp(-cn_2)$.

Finally, using a similar argument in the proof of Theorem 5.3.4, we can show that $\frac{1}{n_2\alpha}$, $p\exp(-cn_2)$, and $p\exp\big(-c(n_2\alpha)^{2/3}\big)$ are all dominated by $\Big(\frac{\log^7(pn_2)}{n_2\alpha}\Big)^{1/6}$ when $\Big(\frac{\log^7(pn_2)}{n_2\alpha}\Big)^{1/6} < 1$, therefore the desired conclusion follows from combining the three probability bounds. □

Proof of Corollary 5.3.6. Like Corollary 5.3.3, this is a direct consequence of the finite sample result, Theorem 5.3.4 or 5.3.5. □

D.3 Proofs of Results in Section 5.4

Proof of Proposition 5.4.1. Case (i): Assumption 5.4.2 follows from the Jain-Marcus theorem (see Example 2.11.13 in Van der Vaart and Wellner (1996)). Assumption 5.4.3 holds because $\sup_{x\in\mathcal{X}}|h(x,\xi)| \leq |h(\tilde{x},\xi)| + \text{diam}(\mathcal{X})M(\xi)$, where $\text{diam}(X)$ denotes the (finite) diameter of $\mathcal{X}$, and finiteness of second moments of $h(\tilde{x},\xi)$ and $M(\xi)$. Assumption 5.4.4 then follows from the Lipschitz continuity of $h(x,\xi)$ in x and an application of the dominated convergence theorem.

Case (ii): We need two results from empirical process theory to verify Assumption 5.4.2:

Theorem D.3.1 (Theorem 2.6.8 and its proof in Van der Vaart and Wellner (1996)) *If a class $\mathcal{F}$ of measurable functions satisfies:*

i. there exists a countable subset $\mathcal{F}_c \subseteq \mathcal{F}$ such that for every $f \in \mathcal{F}$ there exists a sequence $f_n \in \mathcal{F}_c$ such that $\lim_{n\to\infty} f_n(\xi) = f(\xi)$ for all ξ;

ii. the envelope $\mathcal{E}(\xi) := \sup_{f\in\mathcal{F}}|f(\xi)|$ satisfies $\mathbb{E}_F[(\mathcal{E}(\xi))^2] < \infty$;

iii. $\mathcal{F}$ is a VC-subgraph class (see Section 2.6.2 of Van der Vaart and Wellner (1996)),

then $\mathcal{F}$ is F-Donsker.

Theorem D.3.2 (Theorem 2.10.1 in Van der Vaart and Wellner (1996)) *If a class $\mathcal{F}$ is F-Donsker, then any subclass $\mathcal{G} \subseteq \mathcal{F}$ is also F-Donsker.*

Theorem D.3.3 (Example 2.10.8 in Van der Vaart and Wellner (1996)) *If $\mathcal{F}$ and $\mathcal{G}$ both are uniformly bounded F-Donsker classes, then $\mathcal{F}\cdot\mathcal{G} := \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$ is also F-Donsker.*

In order to show F-Donskerness of the class of constraint functions, it suffices to show F-Donskerness for the larger function class $\tilde{\mathcal{F}} := \{\mathbf{1}(a_k'x_k \le b_k y_k + z_k \text{ for } k = 1,\ldots,K) : x_k \in \mathbb{R}^{m_k}, y_k, z_k \in \mathbb{R}, \text{ for } k = 1,\ldots,K\}$ according to Theorem D.3.2. Moreover, note that $\tilde{\mathcal{F}} = \tilde{\mathcal{F}}_1 \cdot \tilde{\mathcal{F}}_2 \cdots \tilde{\mathcal{F}}_K$ where each $\tilde{\mathcal{F}}_k := \{\mathbf{1}(a_k'x \le b_k y + z) : x \in \mathbb{R}^{m_k}, y, z \in \mathbb{R}\}$, therefore by applying Theorem D.3.3 recursively we see that F-Donskerness for all $\tilde{\mathcal{F}}_k$'s implies F-Donskerness of $\tilde{\mathcal{F}}$. It remains to prove F-Donskerness of each $\tilde{\mathcal{F}}_k$ using Theorem D.3.1. Among the conditions of Theorem D.3.1, (ii) is trivially satisfied since the family of indicator functions is uniformly bounded by 1. By writing $a_k'x \le b_k y + z$ as $(a_k', -b_k, -1)(x', y, z)' \le 0$ and noting that the collection of all half-spaces on $\mathbb{R}^{m_k+2}$ has a VC dimension $m_k + 4$ (Problem 14 in Section 2.6 in Van der Vaart and Wellner (1996)), we have that $\tilde{\mathcal{F}}_k$ is a VC-subgraph class (Problem 9 in Section 2.6 in Van der Vaart and Wellner (1996)) therefore (iii) holds. To verify condition (i), consider the countable subclass $\tilde{\mathcal{F}}_k^c = \{\mathbf{1}(a_k'x \le b_k y + z) : x \in \mathbb{Q}^{m_k}, y, z \in \mathbb{Q}\}$ where $\mathbb{Q}$ denotes the set of all rationals. Given $x_o \in \mathbb{R}^{m_k}, y_o, z_o \in \mathbb{R}$, one can pick a sequence $x_o^i \in \mathbb{Q}^{m_k}, y_o^i, z_o^i \in \mathbb{Q}$ such that $z_o^i > z_o$ for all i, $\lim_{i\to\infty} x_o^i \to x_o, \lim_{i\to\infty} y_o^i \to y_o, \lim_{i\to\infty} z_o^i \to z_o$ and

$$\lim_{i\to\infty} \frac{\|x_o^i - x_o\|_2 + |y_o^i - y_o|}{z_o^i - z_o} = 0. \tag{D.8}$$

For every fixed a_k, b_k such that $a_k' x_o < b_k y_o + z_o$ ($a_k' x_o > b_k y_o + z_o$) we have $a_k' x_o^i < b_k y_o^i + z_o^i$ ($a_k' x_o^i > b_k y_o^i + z_o^i$) for sufficiently large i because of the convergence of x_o^i, y_o^i, z_o^i to x_o, y_o, z_o. For a_k, b_k such that $a_k' x_o = b_k y_o + z_o$ we have $a_k' x_o^i \le b_k y_o^i + z_o^i$ for sufficiently large i thanks to (D.8). Therefore $\mathbf{1}(a_k' x_o^i \le b_k y_o^i + z_o^i)$ converges to $\mathbf{1}(a_k' x_o \le b_k y_o + z_o)$ pointwise as $i \to \infty$, giving rise to condition (i). Theorem D.3.1 then implies that each $\tilde{\mathcal{F}}_k$ is F-Donsker.

Assumption 5.4.3 trivially holds since indicator functions are uniformly bounded by 1. It remains to prove Assumption 5.4.4. For any x, x' we write

$$\begin{aligned} & |\mathbf{1}(a_k' A_k(x) \le b_k \text{ for } k = 1, \ldots, K) - \mathbf{1}(a_k' A_k(x') \le b_k \text{ for } k = 1, \ldots, K)| \\ \le & \sum_{k=1}^K \mathbf{1}(a_k' A_k(x) \le b_k < a_k' A_k(x') \text{ or } a_k' A_k(x') \le b_k < a_k' A_k(x)) \\ \le & \sum_{k=1}^K \mathbf{1}(|a_k' A_k(x) - b_k| \le \|a_k\|_2 \|A_k(x) - A_k(x')\|_2). \end{aligned}$$

Therefore

$$\begin{aligned} & \mathbb{E}_F[|\mathbf{1}(a_k' A_k(x) \le b_k \text{ for } k = 1, \ldots, K) - \mathbf{1}(a_k' A_k(x') \le b_k \text{ for } k = 1, \ldots, K)|^2] \\ \le & \big(\sum_{k=1}^K \mathbb{P}_F(|a_k' A_k(x) - b_k| \le \|a_k\|_2 \|A_k(x) - A_k(x')\|_2)\big)^2 \end{aligned}$$

hence it suffices to show each $\mathbb{P}_F(|a_k' A_k(x) - b_k| \le \|a_k\|_2 \|A_k(x) - A_k(x')\|_2) \to 0$ as $x' \to x$. We use the bound $\mathbb{P}_F(|a_k' A_k(x) - b_k| \le \|a_k\|_2 \|A_k(x) - A_k(x')\|_2) \le \mathbb{P}_F(|a_k' A_k(x) - b_k| \le \epsilon) + \mathbb{P}_F(\|a_k\|_2 \|A_k(x) - A_k(x')\|_2 > \epsilon)$ for any $\epsilon > 0$. On one hand we have $\mathbb{P}_F(|a_k' A_k(x) - b_k| \le \epsilon) \to 0$ as $\epsilon \to 0$. To explain, if a_k has a density and $b_k \ne 0$, then $a_k' A_k(x) - b_k$ either has a density on $\mathbb{R}$ or is a point mass at b_k (when $A_k(x)$ is the zero vector), either of which implies $\mathbb{P}_F(|a_k' A_k(x) - b_k| \le \epsilon) \to 0$. Otherwise if (a_k, b_k) has a joint density, $a_k' A_k(x) - b_k$ has a density hence $\mathbb{P}_F(|a_k' A_k(x) - b_k| \le \epsilon) \to 0$ again. On the other hand, by the continuity of A_k it holds $A_k(x') \to A_k(x)$ hence $\|a_k\|_2 \|A_k(x) - A_k(x')\|_2 = o_p(1)$, leading to $\mathbb{P}_F(\|a_k\|_2 \|A_k(x) - A_k(x')\|_2 > \epsilon) \to 0$ as $x' \to x$ for each fixed ϵ. By sending ϵ to 0, we show $\mathbb{P}_F(|a_k' A_k(x) - b_k| \le \|a_k\|_2 \|A_k(x) - A_k(x')\|_2) \to 0$. □

Proof of Proposition 5.4.2. For any $s_1 < s_2$, $v(s_1) \le v(s_2)$ follows trivially from the mono-

tonicity property $\text{Sol}(s_2) \subseteq \text{Sol}(s_1)$. Furthermore, if $x^*(s_1)$ and $x^*(s_2)$ are the unique optimal solutions for $OPT(s_1)$ and $OPT(s_2)$ respectively and they are distinct, then we have $v(s_1) = f(x^*(s_1)) < f(x^*(s_2)) = v(s_2)$ because $x^*(s_2)$ is feasible but not optimal for $OPT(s_1)$. Otherwise if $x^*(s_1) = x^*(s_2)$ then obviously $v(s_1) = f(x^*(s_1)) = f(x^*(s_2)) = v(s_2)$. □

Proof of Proposition 5.4.3. A consequence of Assumption 5.4.4 is the continuity of $H(x)$ on $\mathcal{X}$ because for every x', x it holds $|H(x') - H(x)| \leq \mathbb{E}_F[|h(x', \xi) - h(x, \xi)|] \leq \sqrt{\mathbb{E}_F[|h(x', \xi) - h(x, \xi)|^2]}$.

We prove the uniqueness of optimal solution by contradiction. Suppose there are $x_1 \neq x_2$ and both $x_1, x_2 \in \mathcal{X}_S^*$. Case (i): there are $s_1, s_2 \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ such that $x_1 = x^*(s_1), x_2 = x^*(s_2)$. In this case we must have $s_1 \neq s_2$ hence $f(x_1) \neq f(x_2)$ by Proposition 5.4.2, contradicting with the fact that both x_1, x_2 are optimal. Case (ii): there exists some $s_1 \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ such that $x_1 = x^*(s_1)$, and $x_2 \in x^*(\tilde{s}_{i^*})$ for some $1 \leq i^* \leq M-1$ but $x_2 \neq x^*(s)$ for all $s \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$. Since x_2 is feasible we have $H(x_2) \geq \gamma$ on one hand. On the other hand, $H(x_2) \neq \gamma$ due to Assumption 5.4.7, therefore we must have $H(x_2) > \gamma$. We argue that it must be the case that $x_2 = x^*(\tilde{s}_{i^*}+)$. If $x_2 = x^*(\tilde{s}_{i^*}-)$ then as $s \to \tilde{s}_{i^*}-$ we must have $x^*(s) \neq x_2$ and $x^*(s) \to x_2$, therefore by the continuity of $H(x)$ there exist $s_1' < s_2' < \tilde{s}_{i^*}$ such that $H(x^*(s_1')) > \gamma, H(x^*(s_2')) > \gamma$ and $x^*(s_1') \neq x^*(s_2')$. For such s_1', s_2' we have $f(x^*(s_1')) < f(x^*(s_2')) \leq f(x_2)$ from Proposition 5.4.2, i.e., $x^*(s_1')$ is a feasible solution with strictly less objective value than x_2, contradicting with the optimality of x_2. Hence $x_2 = x^*(\tilde{s}_{i^*}+)$ must hold. If $s_1 < \tilde{s}_{i^*}$, we argue that $f(x_1) < f(x_2)$ hence arrive at a contradiction. Note that the feasible set $\text{Sol}(s_1)$ is closed, that $\text{Sol}(s) \subseteq \text{Sol}(s_1)$ for all $s > s_1$, and that $x_2 = \lim_{s \to \tilde{s}_{i^*}} x^*(s)$ with each $x^*(s) \in \text{Sol}(s_1)$, hence $x_2 \in \text{Sol}(s_1)$. Since $x_2 \neq x^*(s_1) = x_1$ we must have $f(x_1) < f(x_2)$ by the uniqueness of $x^*(s_1)$ for $OPT(s_1)$. Otherwise if $s_1 > \tilde{s}_{i^*}$, we take an $s \in (\tilde{s}_{i^*}, s_1)$ sufficiently close to $\tilde{s}_{i^*}$ so that $x^*(s)$ is sufficiently close to x_2 and $x^*(s) \neq x_1 = x^*(s_1)$, then from Proposition 5.4.2 we have $f(x_2) \leq f(x^*(s)) < f(x_1)$, a contradiction again. Case (iii): there are $\tilde{s}_{i_1^*}, \tilde{s}_{i_2^*}$ such that $x_1 \in x^*(\tilde{s}_{i_1^*})$ and $x_2 \in x^*(\tilde{s}_{i_2^*})$, but there is no $s \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ such that $x_1 = x^*(s)$ or $x_2 = x^*(s)$. By the same argument in Case (ii), we can show that it must be the case that $x_1 = x^*(\tilde{s}_{i_1^*}+)$ and $x_2 = x^*(\tilde{s}_{i_2^*}+)$, therefore $H(x_1), H(x_2) > \gamma$. Assume $\tilde{s}_{i_1^*} < \tilde{s}_{i_2^*}$ without loss of generality, and consider an $s \in (\tilde{s}_{i_1^*}, \tilde{s}_{i_2^*})$ that is sufficiently close to $\tilde{s}_{i_1^*}$ so that $H(x^*(s)) > \gamma$, then by Proposition 5.4.2 we have $f(x_1) = f(x^*(s)) = f(x_2)$ hence

$x^*(s) \in \mathcal{X}_S^*$, and we are in Case (ii) again. This proves that $\mathcal{X}_S^*$ must be a singleton.

To show that the optimal parameter set S^* must be a closed interval, we first observe that S^* must be a closed set due to the continuity of the solution path. Let $s_l^* = \min\{s : s \in S^*\}$ and $s_u^* = \max\{s : s \in S^*\}$, then we have $S^* \subseteq [s_l^*, s_u^*]$. Case (i): both $s_l^*, s_u^* \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$. Note that $x^*(s_l^*) = x^*(s_u^*) = x_S^*$ and $v(s_l^*) = v(s_u^*)$, hence Proposition 5.4.2 then forces $x^*(s) = x_S^*$ hence $s \in S^*$ for all $s \in [s_l^*, s_u^*] - \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ because otherwise $v(s_l^*) < v(s) < v(s_u^*)$. This further implies $x^*(\tilde{s}_i-) = x^*(\tilde{s}_i+) = \{x_S^*\}$ and subsequently $\tilde{s}_i \in S^*$ for any $\tilde{s}_i \in [s_l^*, s_u^*]$. Altogether we have $[s_l^*, s_u^*] = S^*$. Case (ii): $s_l^* = \tilde{s}_{i^*}$ for some $1 \le i^* \le M-1$ and $s_u^* \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$. Using Proposition 5.4.2 as in Case (i), one can show that for every $s \in [s_l^*, s_u^*] - \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ we have $v(s_l^*) = v(s) = v(s_u^*)$ and $x^*(s) = x_S^*$, therefore $S^* = [s_l^*, s_u^*]$ again. Case (iii): $s_u^* = \tilde{s}_{i^*}$ for some $1 \le i^* \le M-1$ and $s_l^* \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$. This case resembles Case (ii) and $S^* = [s_l^*, s_u^*]$ can be shown using the same argument. Case (iv): $s_l^* = \tilde{s}_{i_1^*}, s_u^* = \tilde{s}_{i_2^*}$ for some $1 \le i_1^* \le i_2^* \le M-1$. If $\tilde{s}_{i_1^*} = \tilde{s}_{i_2^*}$ then x_S^* is a singleton and the interval representation trivially holds, so we focus on the case $\tilde{s}_{i_1^*} < \tilde{s}_{i_2^*}$. We argue that $x^*(\tilde{s}_{i_1^*}+) = x_S^*$. Otherwise if $x^*(\tilde{s}_{i_1^*}-) = x_S^*$, then Assumption 5.4.7 forces $H(x^*(\tilde{s}_{i_1^*}-)) > \gamma$, and the continuity of the solution path and the constraint H imply that $H(x^*(s')) > \gamma$ for some s' sufficiently close to $\tilde{s}_{i_1^*}$ but $s' < \tilde{s}_{i_1^*}$. Note that such an s' corresponds to an objective value $f(x^*(s')) = v(s') \le \lim_{s \to \tilde{s}_{i_1^*}-} v(s) = f(x_S^*)$ by Proposition 5.4.2, therefore $x^*(s') = x_S^*$ by the uniqueness of $\mathcal{X}_S^*$, contradicting with the definition of s_l^*. Therefore it must be the case that $x^*(\tilde{s}_{i_1^*}+) = x_S^*$. Because $H(x^*(\tilde{s}_{i_1^*}+)) > \gamma$, there exists a $\delta > 0$ so that $\tilde{s}_{i_1^*} + \delta < \tilde{s}_{i_2^*}$, $\tilde{s}_{i_1^*} + \delta \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$, and $H(x^*(s)) > \gamma$ for all $s \in (\tilde{s}_{i_1^*}, \tilde{s}_{i_1^*} + \delta]$. Since Proposition 5.4.2 implies $v(s) = f(x_S^*)$ for such s, we must have $x^*(s) = x_S^*$ hence $[s_l^*, s_l^* + \delta] \subseteq S^*$. The rest part $[s_l^* + \delta, s_u^*] \subseteq S^*$ can be shown by treating $s_l^* + \delta$ as the s_l^* in Case (iii). Altogether we still have $S^* = [s_l^*, s_u^*]$. In particular, when $v(s)$ is strictly monotonic, it is clear that there can be at most one optimal parameter hence S^* becomes a singleton. □

Proof of Theorem 5.4.4. For any function class $\mathcal{G}$ and $g \in \mathcal{G}$, we write $P(g) = \mathbb{E}_F[g(\xi)]$ (or just Pg) and $P_{n_2}(g) = \frac{1}{n_2}\sum_{i=1}^{n_2} g(\xi_i)$ (or just $P_{n_2}g$), as functions from $\mathcal{G} \to \mathbb{R}$. For any function $\phi : \mathcal{G} \to \mathbb{R}$, define $\|\phi\|_{\mathcal{G}} = \sup_{g \in G}|\phi(g)|$. For example $\|P_{n_2} - P\|_{\mathcal{G}}$ denotes the maximum deviation of the sample mean.

First we show uniform convergence of the standard-deviation-adjusted sample mean to the expected constraint value. The function class $\{h(x,\cdot)|x \in \mathcal{X}\}$ is F-Donsker by Assumption 5.4.2, hence is F-Glivenko-Cantelli(GC). By Lemma 2.10.14 from Van der Vaart and Wellner (1996), the squared class $\{h^2(x,\cdot)|x \in \mathcal{X}\}$ is also F-GC under Assumptions 5.4.2 and 5.4.3. Define $\mathcal{X}_S = \{x^*(s) : s \in S\backslash\{\tilde{s}_1,\ldots,\tilde{s}_{M-1}\}\} \cup \big(\cup_{i=1}^{M-1} x^*(\tilde{s}_i)\big)$. As sub-classes, $\mathcal{F}_{\mathcal{X}_S} := \{h(x,\cdot)|x \in \mathcal{X}_S\}$ and $\mathcal{F}^2_{\mathcal{X}_S} := \{h^2(x,\cdot)|x \in \mathcal{X}_S\}$ are both F-GC, i.e.,

$$\|P_{n_2} - P\|_{\mathcal{F}_{\mathcal{X}_S}} \to 0 \text{ a.s.}$$
$$\|P_{n_2} - P\|_{\mathcal{F}^2_{\mathcal{X}_S}} \to 0 \text{ a.s..}$$

Letting $\hat{\sigma}^2(h) = P_{n_2}(h^2) - (P_{n_2}(h))^2$ and $\sigma^2(h) = P(h^2) - (P(h))^2$ be the sample and true variances, we have

$$\begin{aligned}
\left\|\hat{\sigma}^2 - \sigma^2\right\|_{\mathcal{F}_{\mathcal{X}_S}} &\leq \|P_{n_2} - P\|_{\mathcal{F}^2_{\mathcal{X}_S}} + \|P_{n_2} - P\|^2_{\mathcal{F}_{\mathcal{X}_S}} + 2\|P\|_{\mathcal{F}_{\mathcal{X}_S}} \|P_{n_2} - P\|_{\mathcal{F}_{\mathcal{X}_S}} \\
&= \|P_{n_2} - P\|_{\mathcal{F}^2_{\mathcal{X}_S}} + \|P_{n_2} - P\|^2_{\mathcal{F}_{\mathcal{X}_S}} + 2\sup_{x\in\mathcal{X}_S}|H(x)| \|P_{n_2} - P\|_{\mathcal{F}_{\mathcal{X}_S}} \\
&\to 0 \text{ a.s.} \qquad \text{(D.9)}
\end{aligned}$$

where the limit comes from the fact that $\sup_{x\in\mathcal{X}_S}|H(x)| < \infty$ because H is continuous (implied by Assumption 5.4.4) and $\mathcal{X}_S$ is compact (implied by the piecewise uniform continuity condition, i.e., Assumption 5.4.6). By Assumption 5.4.3 we have $\left\|\sigma^2\right\|_{\mathcal{F}_{\mathcal{X}_S}} = \sup_{x\in\mathcal{X}_S}\mathrm{Var}(h(x,\xi)) \leq \mathbb{E}[\sup_{x\in\mathcal{X}} h^2(x,\xi)] < \infty$, and arrive at

$$\left\|P_{n_2} - \frac{z_{1-\beta}}{\sqrt{n_2}}\hat{\sigma} - P\right\|_{\mathcal{F}_{\mathcal{X}_S}} \leq \|P_{n_2} - P\|_{\mathcal{F}_{\mathcal{X}_S}} + \frac{z_{1-\beta}}{\sqrt{n_2}}\sqrt{\|\hat{\sigma}^2\|_{\mathcal{F}_{\mathcal{X}_S}}} \to 0 \text{ a.s..}$$

When we use a discrete mesh $\{s_1,\ldots,s_p\}$, it is clear that, using the notations from Algorithm 11 and $H_j := H(x^*(s_j))$

$$\max_{1\leq j\leq p}\left|\hat{H}_j - \frac{z_{1-\beta}}{\sqrt{n_2}}\hat{\sigma}_j - H_j\right| \leq \left\|P_{n_2} - \frac{z_{1-\beta}}{\sqrt{n_2}}\hat{\sigma} - P\right\|_{\mathcal{F}_{\mathcal{X}_S}} \to 0 \text{ a.s..} \qquad \text{(D.10)}$$

Secondly, we prove convergence of the estimated solution $x^*(\hat{s}^*)$ to the optimum x_S^*. Fixing any $\epsilon > 0$, we argue that almost surely there exists a finite N and $\overline{\epsilon_S} > 0$ such that for all $n_2 \geq N$ and $\epsilon_S \leq \overline{\epsilon_S}$ it holds $\|x^*(\hat{s}^*) - x_S^*\| < \epsilon$. To proceed, define

$$\delta := \min_{x \in \mathcal{X}_S} \{f(x) - f(x_S^*) | H(x) \geq \gamma, \|x - x_S^*\| \geq \epsilon\}. \tag{D.11}$$

Since the objective f is continuous and $\{x | x \in \mathcal{X}_S, H(x) \geq \gamma, \|x - x_S^*\| \geq \epsilon\}$ is a compact set, by a compactness argument we must have $\delta > 0$. By Assumption 5.4.9, for any $\epsilon' \leq \epsilon$ there exists some $s' \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ such that $H(x^*(s')) > \gamma$ and $\|x^*(s') - x_S^*\| < \epsilon'$. By continuity of f, one can set ϵ' small enough so that $f(x^*(s')) - f(x_S^*) < \frac{\delta}{2}$. Moreover, due to the continuity of $x^*(s)$ at s' and the continuity of f and H, there exists an $\overline{\epsilon_S} > 0$ such that $\min_{s \text{ s.t. } |s-s'| \leq \overline{\epsilon_S}} H(x^*(s)) > \gamma$ and $\max_{s \text{ s.t. } |s-s'| \leq \overline{\epsilon_S}} f(x^*(s)) < f(x_S^*) + \frac{\delta}{2}$. Therefore, when the mesh size $\epsilon_S \leq \overline{\epsilon_S}$, there must exist some $s_{j'} \in \{s_1, \ldots, s_p\}$ such that

$$H(x^*(s_{j'})) \geq \min_{s \text{ s.t. } |s-s'| \leq \overline{\epsilon_S}} H(x^*(s)) > \gamma \tag{D.12}$$

$$f(x^*(s_{j'})) \leq \max_{s \text{ s.t. } |s-s'| \leq \overline{\epsilon_S}} f(x^*(s)) < f(x_S^*) + \frac{\delta}{2}. \tag{D.13}$$

For the given ϵ define for $\Delta \geq 0$

$$\delta_\Delta := \min_{x \in \mathcal{X}_S} \{f(x) - f(x_S^*) | H(x) \geq \gamma - \Delta, \|x - x_S^*\| \geq \epsilon\}. \tag{D.14}$$

We argue that $\lim_{\Delta \to 0+} \delta_\Delta \to \delta$ by contradiction. Clearly δ_Δ is non-increasing in Δ and $\delta_\Delta \leq \delta$, hence the limit $\lim_{\Delta \to 0+} \delta_\Delta$ must exists and is finite. Suppose $\lim_{\Delta \to 0+} \delta_\Delta < \delta$, then there exist $\tilde{\delta} < \delta$ and a sequence $\{x_k\}_{k=1}^\infty \subset \mathcal{X}_S$ such that $\|x_k - x_S^*\| \geq \epsilon$, $H(x_k) \geq \gamma - \Delta_k$ with $\Delta_k \to 0+$, and $f(x_k) - f(x_S^*) \leq \tilde{\delta}$. By the compactness of $\mathcal{X}_S$, there must exist a subsequence $\{x_{k_s}\}_{s=1}^\infty$ converging to some $x_\infty \in \mathcal{X}_S$, and by continuity x_∞ must satisfy $\|x_\infty - x_S^*\| \geq \epsilon$, $H(x_\infty) \geq \gamma$ and $f(x_\infty) - f(x_S^*) \leq \tilde{\delta}$. From the definition (D.11) of δ this implies $\delta \leq \tilde{\delta}$, a contradiction. Now pick a small enough Δ' so that $\delta_{\Delta'} > \frac{\delta}{2}$. From the uniform convergence (D.10) we know that almost

surely there exists some N such that for any $n_2 \geq N$

$$\max_{1\leq j\leq p}\left|\hat{H}_j - \frac{z_{1-\beta}}{\sqrt{n_2}}\hat{\sigma}_j - H_j\right| < \min\{\Delta', \min_{s \text{ s.t. } |s-s'|\leq\overline{\epsilon_S}} H(x^*(s)) - \gamma\} \leq \min\{\Delta', H(x^*(s_{j'})) - \gamma\}. \tag{D.15}$$

where the second inequality is due to (D.12). In particular, (D.15) implies that for such n_2 we have $\hat{H}(x^*(s_{j'})) - \frac{z_{1-\beta}}{\sqrt{n_2}}\hat{\sigma}(x^*(s_{j'})) > H(x^*(s_{j'})) - \min\{\Delta', H(x^*(s_{j'})) - \gamma\} \geq \gamma$, therefore on one hand we must have

$$f(x^*(\hat{s}^*)) \leq f(x^*(s_{j'})) < f(x_S^*) + \frac{\delta}{2} \tag{D.16}$$

where the first inequality holds due to the way $\hat{s}^*$ is chosen and the second results from (D.13). On the other hand it also follows from (D.15) that

$$\begin{aligned} H(x^*(\hat{s}^*)) &> \hat{H}(x^*(\hat{s}^*)) - \frac{z_{1-\beta}}{\sqrt{n_2}}\hat{\sigma}(x^*(\hat{s}^*)) - \min\{\Delta', H(x^*(s_{j'})) - \gamma\} \\ &\geq \gamma - \min\{\Delta', H(x^*(s_{j'})) - \gamma\} \\ &\geq \gamma - \Delta'. \end{aligned} \tag{D.17}$$

The bounds (D.16) and (D.17) on the objective value and the constraint value at the estimated solution $x^*(\hat{s}^*)$, together with the fact that $\delta_{\Delta'} > \frac{\delta}{2}$ due to the way Δ' is chosen, imply that $\|x^*(\hat{s}^*) - x_S^*\| < \epsilon$ by the definition (D.14) of δ_Δ. Since ϵ can be arbitrarily small, we have $\lim_{n_2\to\infty} x^*(\hat{s}^*) = x_S^*$ a.s.. Convergence of $\hat{s}^*$ to the optimal parameter set $\mathcal{S}^*$ is then a consequence of the convergence of $x^*(\hat{s}^*)$ to x_S^*. Suppose $\hat{s}^*$ does not converge to S^*, then by compactness of S there exists a subsequence $\hat{s}_k^*$ converging to some $s_\infty \notin S^*$. Since the corresponding $x^*(\hat{s}_k^*) \to x_S^*$ we have either $s_\infty \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ with $x^*(s_\infty) = x_S^*$ or $s_\infty \in \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ with $x_S^* \in x^*(s_\infty)$, however in either case $s_\infty \in S^*$, a contradiction.

Then we prove the feasibility guarantees. The case $H(x_S^*) > \gamma$ is relatively straightforward. By the continuity of H and that a.s. $x^*(\hat{s}^*) \to x_S^*$ we have $H(x^*(\hat{s}^*)) \to H(x_S^*) > \gamma$ a.s.. Almost surely convergence implies convergence in probability, thus $H(x^*(\hat{s}^*)) \to H(x_S^*)$ in probability and,

in particular, $P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \to 1$. If $H(x_S^*) = \gamma$ we denote by

$$\mathbb{G}_{n_2}(x) := \sqrt{n_2}(P_{n_2}(h(x,\cdot)) - P(h(x,\cdot)))$$

the empirical process indexed by the decision x and let $\hat{\sigma}^2(x), \sigma^2(x)$ represent the sample and true variance of $h(x,\xi)$, and then write

$$\begin{aligned} H(x^*(\hat{s}^*)) &\geq H(x^*(\hat{s}^*)) - \big(\hat{H}(x^*(\hat{s}^*)) - z_{1-\beta}\frac{\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}} - \gamma\big) \qquad \text{(D.18)} \\ &= \gamma + (H(x^*(\hat{s}^*)) - \hat{H}(x^*(\hat{s}^*))) + z_{1-\beta}\frac{\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}} \\ &= \gamma - \frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x^*(\hat{s}^*)) + z_{1-\beta}\frac{\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}} \\ &= \gamma - \frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + z_{1-\beta}\frac{\hat{\sigma}(x_S^*)}{\sqrt{n_2}} + E_1 + E_2 \qquad \text{(D.19)} \end{aligned}$$

where the errors

$$E_1 = \frac{1}{\sqrt{n_2}}\big(\mathbb{G}_{n_2}(x_S^*) - \mathbb{G}_{n_2}(x^*(\hat{s}^*))\big),\ E_2 = \frac{z_{1-\beta}}{\sqrt{n_2}}\big(\hat{\sigma}(x^*(\hat{s}^*)) - \hat{\sigma}(x_S^*)\big).$$

We need to show that $E_1 = o_p\big(\frac{1}{\sqrt{n_2}}\big), E_2 = o_p\big(\frac{1}{\sqrt{n_2}}\big)$. We deal with E_2 first. E_2 can be bounded as

$$|E_2| \leq \frac{z_{1-\beta}}{\sqrt{n_2}}\big(2\,\|\hat{\sigma} - \sigma\|_{\mathcal{F}_{\mathcal{X}_S}} + |\sigma(x^*(\hat{s}^*)) - \sigma(x_S^*)|\,\big).$$

On one hand we have already shown that $x^*(\hat{s}^*) \to x_S^*$ a.s.. On the other hand, $\sigma^2(x)$ is continuous in x. Therefore $|\sigma(x^*(\hat{s}^*)) - \sigma(x_S^*)| \to 0$ a.s. as $n_2 \to \infty$. By uniform convergence (D.9) and the relation $\|\hat{\sigma} - \sigma\|_{\mathcal{F}_{\mathcal{X}_S}} \leq \sqrt{\|\hat{\sigma}^2 - \sigma^2\|_{\mathcal{F}_{\mathcal{X}_S}}}$, we have $\|\hat{\sigma} - \sigma\|_{\mathcal{F}_{\mathcal{X}_S}} \to 0$ a.s.. Consequently it holds $\sqrt{n_2}\,|E_2| = o(1)$ a.s. and, in particular, $E_2 = o_p\big(\frac{1}{\sqrt{n_2}}\big)$.

To bound the error E_1, let $\rho(x,x') = \sqrt{\mathrm{Var}(h(x,\xi) - h(x',\xi))}$ denote the intrinsic semimetric of the tight Gaussian process $\mathbb{G}$ indexed by $x \in \mathcal{X}_S$ with mean zero and covariance structure $\mathrm{Cov}(\mathbb{G}(x), \mathbb{G}(x')) = \mathrm{Cov}_F(h(x,\xi), h(x',\xi))$, and for any $\epsilon > 0$ let $\delta(\epsilon) = \sup\{\rho(x, x_S^*) | x \in$

$\mathcal{X}_S, \|x - x_S^*\| < \epsilon\}$. Note that Assumption 5.4.4 entails $\delta(\epsilon) \to 0$ as $\epsilon \to 0$. We have for any $\epsilon > 0$

$$\begin{aligned}
|E_1| &= |E_1|\,\mathbf{1}\{\|x^*(\hat{s}^*) - x_S^*\| < \epsilon\} + |E_1|\,\mathbf{1}\{\|x^*(\hat{s}^*) - x_S^*\| \geq \epsilon\} \\
&\leq \sup_{x \in \mathcal{X}_S \text{ s.t. } \|x - x_S^*\| < \epsilon} \frac{1}{\sqrt{n_2}} |\mathbb{G}_{n_2}(x_S^*) - \mathbb{G}_{n_2}(x)|\,\mathbf{1}\{\|x^*(\hat{s}^*) - x_S^*\| < \epsilon\} + \infty \cdot \mathbf{1}\{\|x^*(\hat{s}^*) - x_S^*\| \geq \epsilon\} \\
&\quad \text{where } \infty \cdot 0 = 0 \\
&\leq \sup_{x \in \mathcal{X}_S \text{ s.t. } \|x - x_S^*\| < \epsilon} \frac{1}{\sqrt{n_2}} |\mathbb{G}_{n_2}(x_S^*) - \mathbb{G}_{n_2}(x)| + \infty \cdot \mathbf{1}\{\|x^*(\hat{s}^*) - x_S^*\| \geq \epsilon\} \\
&\leq \sup_{x, x' \in \mathcal{X}_S \text{ s.t. } \rho(x, x') \leq \delta(\epsilon)} \frac{1}{\sqrt{n_2}} \left|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\right| + \infty \cdot \mathbf{1}\{\|x^*(\hat{s}^*) - x_S^*\| \geq \epsilon\}.
\end{aligned}$$

We have already shown that $\|x^*(\hat{s}^*) - x_S^*\| \to 0$ a.s., hence $P_{\boldsymbol{\xi}_{1:n_2}}(\|x^*(\hat{s}^*) - x_S^*\| \geq \epsilon) \to 0$ for any fixed $\epsilon > 0$. Therefore we can choose an n_2-dependent $\epsilon := \epsilon_{n_2}$ so that both $P_{\boldsymbol{\xi}_{1:n_2}}(\|x^*(\hat{s}^*) - x_S^*\| \geq \epsilon_{n_2}) \to 0$ and $\epsilon_{n_2} \to 0$ as $n_2 \to \infty$, and get

$$|E_1| \leq \sup_{x, x' \in \mathcal{X}_S \text{ s.t. } \rho(x, x') \leq \delta(\epsilon_{n_2})} \frac{1}{\sqrt{n_2}} \left|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\right| + \infty \cdot \mathbf{1}\{\|x^*(\hat{s}^*) - x_S^*\| \geq \epsilon_{n_2}\}. \tag{D.20}$$

By the way ϵ_{n_2} is chosen, the second term on the right hand side of (D.20) is of arbitrarily small order, in particular, $o_p\big(\frac{1}{\sqrt{n_2}}\big)$. To control the first term, note that $\delta(\epsilon_{n_2}) \to 0$ as $n_2 \to \infty$. Since the function class $\mathcal{F}_{\mathcal{X}_S}$ is F-Donsker, the empirical process $\mathbb{G}_{n_2}$ on $\mathcal{F}_{\mathcal{X}_S}$ is asymptotically tight, hence by Theorem 1.5.7 and Addendum 1.5.8 from Van der Vaart and Wellner (1996) $\mathbb{G}_{n_2}$ is asymptotically uniformly equicontinuous in probability with respect to the intrinsic semimetric ρ of the limit Gaussian process $\mathbb{G}$, i.e., for any $\epsilon > 0$

$$\lim_{\delta \to 0} \limsup_{n_2 \to \infty} P_{\boldsymbol{\xi}_{1:n_2}}\Big(\sup_{x, x' \in \mathcal{X}_S \text{ s.t. } \rho(x, x') \leq \delta} \left|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\right| > \epsilon \Big) = 0. \tag{D.21}$$

Note that $\sup_{x, x' \in \mathcal{X}_S \text{ s.t. } \rho(x, x') \leq \delta} |\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')|$ is monotonically increasing in δ a.s. and that $\delta(\epsilon_{n_2}) \to 0$, therefore it must hold that for any fixed $\delta > 0$

$$\sup_{x, x' \in \mathcal{X}_S \text{ s.t. } \rho(x, x') \leq \delta(\epsilon_{n_2})} \left|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\right| \leq \sup_{x, x' \in \mathcal{X}_S \text{ s.t. } \rho(x, x') \leq \delta} \left|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\right| \quad \text{a.s.}$$

when n_2 is sufficiently large, therefore for any $\epsilon > 0$ the first term in (D.20) can be controlled as

$$\begin{aligned} & \limsup_{n_2\to\infty} P_{\boldsymbol{\xi}_{1:n_2}}\Big(\sup_{x,x'\in\mathcal{X}_S \text{ s.t. } \rho(x,x')\le\delta(\epsilon_{n_2})} \big|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\big| > \epsilon\Big) \\ \le\ & \limsup_{n_2\to\infty} P_{\boldsymbol{\xi}_{1:n_2}}\Big(\sup_{x,x'\in\mathcal{X}_S \text{ s.t. } \rho(x,x')\le\delta} \big|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\big| > \epsilon\Big). \end{aligned} \tag{D.22}$$

Due to (D.21) the right hand side of (D.22) can be made arbitrarily small by sending $\delta \to 0$, hence the left hand side of (D.22) must be identical to zero. Since ϵ is arbitrary, by definition $\sup_{x,x'\in\mathcal{X}_S \text{ s.t. } \rho(x,x')\le\delta(\epsilon_{n_2})} |\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')| = o_p(1)$, which in turn leads to $E_1 = o_p\big(\frac{1}{\sqrt{n_2}}\big)$.

We now go back to the representation (D.19) of $H(x^*(\hat{s}^*))$ to conclude the coverage guarantee. From (D.19) we see that

$$\begin{aligned} & \liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \ge \gamma) \\ \ge\ & \liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(\gamma - \frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + z_{1-\beta}\frac{\hat{\sigma}(x_S^*)}{\sqrt{n_2}} + E_1 + E_2 \ge \gamma\Big) \\ =\ & \liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(- \frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + z_{1-\beta}\frac{\hat{\sigma}(x_S^*)}{\sqrt{n_2}} + o_p\big(\frac{1}{\sqrt{n_2}}\big) \ge 0\Big) \\ =\ & \liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(\frac{\mathbb{G}_{n_2}(x_S^*)}{\hat{\sigma}(x_S^*)} + o_p(1) \le z_{1-\beta}\Big) \\ & \quad \text{since } \hat{\sigma}(x_S^*) \to \sigma(x_S^*) \text{ a.s. and } \sigma^2(x_S^*) > 0 \text{ (Assumption 5.4.5)} \\ =\ & 1-\beta \end{aligned}$$

where in the last equality we use Slutsky's theorem to justify that $\frac{\mathbb{G}_{n_2}(x_S^*)}{\hat{\sigma}(x_S^*)} + o_p(1)$ weakly converges to the standard normal. □

Proof of Theorem 5.4.5. Following the proof of Theorem 5.4.4, we see that in order to conclude the tight feasibility confidence level it suffices to show that the inequality gap of (D.18) is of order $o_p\big(\frac{1}{\sqrt{n_2}}\big)$, i.e.,

$$\gamma \le \hat{H}(x^*(\hat{s}^*)) - z_{1-\beta}\frac{\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}} \le \gamma + o_p\big(\frac{1}{\sqrt{n_2}}\big). \tag{D.23}$$

Indeed, once the second inequality in (D.23) is shown, we can use the representation (D.19) and

apply Slutsky's theorem, like in the proof of Theorem 5.4.4, to get

$$\begin{aligned}
& \lim P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \\
= \; & \lim P_{\boldsymbol{\xi}_{1:n_2}}\Big(\gamma - \frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + z_{1-\beta}\frac{\hat{\sigma}(x_S^*)}{\sqrt{n_2}} + E_1 + E_2 + o_p\big(\frac{1}{\sqrt{n_2}}\big) \geq \gamma\Big) \\
= \; & \lim P_{\boldsymbol{\xi}_{1:n_2}}\Big(-\frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + z_{1-\beta}\frac{\hat{\sigma}(x_S^*)}{\sqrt{n_2}} + o_p\big(\frac{1}{\sqrt{n_2}}\big) \geq 0\Big) \;\; \text{since } E_1, E_2 = o_p\big(\frac{1}{\sqrt{n_2}}\big) \\
= \; & \lim P_{\boldsymbol{\xi}_{1:n_2}}\Big(\frac{\mathbb{G}_{n_2}(x_S^*)}{\hat{\sigma}(x_S^*)} + o_p(1) \leq z_{1-\beta}\Big) \\
= \; & 1 - \beta.
\end{aligned}$$

Now we prove the second inequality in (D.23). By Proposition 5.4.3 the optimal parameter set is a singleton $S^* = \{s^*\}$. Moreover, in the case $H(x_S^*) = \gamma$ Assumption 5.4.7 forces that $s^* \notin \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\} \cup \{s_l, s_u\}$. Suppose $s^* \in (\tilde{s}_{i^*}, \tilde{s}_{i^*+1})$ for some $0 \leq i^* \leq M-1$ (note that $\tilde{s}_0 = s_l, \tilde{s}_M = s_u$). Assumption 5.4.6 then ensures that the parameter-to-solution mapping $x^*(\cdot)$ is uniformly continuous in some neighborhood $\mathcal{N}(s^*) \subseteq (\tilde{s}_{i^*}, \tilde{s}_{i^*+1})$ of s^*. Since $\mathcal{N}(s^*)$ is contained in a compact set, the standard deviation function $\sigma(x^*(\cdot))$ is uniformly continuous in $\mathcal{N}(s^*)$. Moreover, the semimetric $\rho(x^*(\cdot), x^*(\cdot))$ between two solutions is also uniformly continuous in $\mathcal{N}(s^*) \times \mathcal{N}(s^*)$. Therefore as $\epsilon_S \to 0$ the following holds

$$\begin{aligned}
\omega_\sigma(2\epsilon_S) :&= \sup_{s,s' \in \mathcal{N}(s^*) \text{ s.t. } \|s-s'\|<2\epsilon_S} \big|\sigma(x^*(s)) - \sigma(x^*(s'))\big| = o(1) \\
\omega_\rho(2\epsilon_S) :&= \sup_{s,s' \in \mathcal{N}(s^*) \text{ s.t. } \|s-s'\|<2\epsilon_S} \rho(x^*(s), x^*(s')) = o(1).
\end{aligned}$$

According to the criterion of choosing $\hat{s}^*$ we must have for every parameter value s_j that either $f(x^*(s_j)) \geq f(x^*(\hat{s}^*))$ or $\hat{H}(x^*(s_j)) - z_{1-\beta}\frac{\hat{\sigma}(x^*(s_j))}{\sqrt{n_2}} < \gamma$. Therefore if $\hat{s}^* \in \mathcal{N}(s^*)$, say $\hat{s}^* = s_{j^*}^{i^*}$, and $s_{j^*-1}^{i^*} \in \mathcal{N}(s^*)$ as well, then because $s_{j^*-1}^{i^*} < s_{j^*}^{i^*}$ and the parameter-to-objective mapping $f(x^*(\cdot))$ is strictly increasing in s it must hold that

$$\hat{H}(x^*(s_{j^*-1}^{i^*})) - z_{1-\beta}\frac{\hat{\sigma}(x^*(s_{j^*-1}^{i^*}))}{\sqrt{n_2}} < \gamma. \tag{D.24}$$

We shall use this fact to derive (D.23). For convenience, we denote by $B(s, \epsilon) := \{s' \in \mathcal{S} | \, \|s' - s\| <$

$\epsilon\}$ the ball of radius $\epsilon > 0$ centered at s. Because $\hat{s}^* \to s^*$ a.s. and $\epsilon_S \to 0$, it is implied that $P_{\boldsymbol{\xi}_{1:n_2}}(B(\hat{s}^*, 2\epsilon_S) \subseteq \mathcal{N}(s^*)) \to 1$ as $n_2 \to \infty$. Thus we can write

$$
\begin{aligned}
& \hat{H}(x^*(\hat{s}^*)) - z_{1-\beta}\frac{\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}} \\
\leq\ & \infty \cdot \mathbf{1}\{B(\hat{s}^*, 2\epsilon_S) \not\subseteq \mathcal{N}(s^*)\} + \big(\hat{H}(x^*(s_{j^*}^{i^*})) - z_{1-\beta}\frac{\hat{\sigma}(x^*(s_{j^*}^{i^*}))}{\sqrt{n_2}}\big)\mathbf{1}\{B(\hat{s}^*, 2\epsilon_S) \subseteq \mathcal{N}(s^*)\} \\
& \text{where } s_{j^*}^{i^*} = \hat{s}^* \\
\leq\ & o_p\big(\frac{1}{\sqrt{n_2}}\big) + \big(\hat{H}(x^*(s_{j^*-1}^{i^*})) - z_{1-\beta}\frac{\hat{\sigma}(x^*(s_{j^*-1}^{i^*}))}{\sqrt{n_2}}\big)\mathbf{1}\{B(\hat{s}^*, 2\epsilon_S) \subseteq \mathcal{N}(s^*)\} + \\
& \Big(\Big|\hat{H}(x^*(s_{j^*-1}^{i^*})) - \hat{H}(x^*(s_{j^*}^{i^*}))\Big| + \frac{z_{1-\beta}}{\sqrt{n_2}}\Big|\hat{\sigma}(x^*(s_{j^*-1}^{i^*})) - \hat{\sigma}(x^*(s_{j^*}^{i^*}))\Big|\Big)\mathbf{1}\{B(\hat{s}^*, 2\epsilon_S) \subseteq \mathcal{N}(s^*)\} \\
\leq\ & o_p\big(\frac{1}{\sqrt{n_2}}\big) + \gamma + \Big(\Big|H(x^*(s_{j^*-1}^{i^*})) - H(x^*(s_{j^*}^{i^*}))\Big| + \frac{1}{\sqrt{n_2}}\Big|\mathbb{G}_{n_2}(x^*(s_{j^*-1}^{i^*})) - \mathbb{G}_{n_2}(x^*(s_{j^*}^{i^*}))\Big| + \\
& \frac{z_{1-\beta}}{\sqrt{n_2}}\big(2\left\|\hat{\sigma} - \sigma\right\|_{\mathcal{F}_{\mathcal{X}_S}} + \Big|\sigma(x^*(s_{j^*-1}^{i^*})) - \sigma(x^*(s_{j^*}^{i^*}))\Big|\big)\Big)\mathbf{1}\{B(\hat{s}^*, 2\epsilon_S) \subseteq \mathcal{N}(s^*)\} \\
& \text{where the } \gamma \text{ term comes from (D.24)} \\
\leq\ & o_p\big(\frac{1}{\sqrt{n_2}}\big) + \gamma + \Big(o\big(\frac{1}{\sqrt{n_2}}\big) + \frac{1}{\sqrt{n_2}}\sup_{x,x' \in \mathcal{X}_S \text{ s.t. } \rho(x,x') \leq \omega_\rho(2\epsilon_S)}\big|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\big| + \\
& \frac{z_{1-\beta}}{\sqrt{n_2}}\big(2\left\|\hat{\sigma} - \sigma\right\|_{\mathcal{F}_{\mathcal{X}_S}} + \omega_\sigma(2\epsilon_S)\big)\Big)\mathbf{1}\{B(\hat{s}^*, 2\epsilon_S) \subset \mathcal{N}(s^*)\} \\
& \text{where the } o\big(\frac{1}{\sqrt{n_2}}\big) \text{ terms is due to condition (5.21)} \\
\leq\ & o_p\big(\frac{1}{\sqrt{n_2}}\big) + \gamma + o\big(\frac{1}{\sqrt{n_2}}\big) + \frac{1}{\sqrt{n_2}}\sup_{x,x' \in \mathcal{X}_S \text{ s.t. } \rho(x,x') \leq \omega_\rho(2\epsilon_S)}\big|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\big| \\
& + \frac{z_{1-\beta}}{\sqrt{n_2}}\big(o_p(1) + o(1)\big) \\
=\ & \gamma + o_p\big(\frac{1}{\sqrt{n_2}}\big) + \frac{1}{\sqrt{n_2}}\sup_{x,x' \in \mathcal{X}_S \text{ s.t. } \rho(x,x') \leq \omega_\rho(2\epsilon_S)}\big|\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')\big|.
\end{aligned}
$$

Since $\omega_\rho(2\epsilon_S) \to 0$, through an argument similar to (D.22) the asymptotically uniform equicontinuity of $\mathbb{G}_{n_2}$ results in $\sup_{x,x' \in \mathcal{X}_S \text{ s.t. } \rho(x,x') \leq \omega_\rho(2\epsilon_S)} |\mathbb{G}_{n_2}(x) - \mathbb{G}_{n_2}(x')| = o_p(1)$. This finally leads to the upper bound in (D.23), hence concludes the theorem. □

Proof of Theorem 5.4.6. We first treat the unnormalized validator (Algorithm 9). First, we introduce some notations and a few auxiliary Gaussian processes. Let $S^o := S \backslash \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$ be

the parameter space after excluding the pathological points $\{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\}$. We denote by

$$\{\overline{\mathbb{G}}(s) : s \in S^o\}$$

the Gaussian process with mean zero and covariance $\mathrm{Cov}(s, s') = \mathrm{Cov}_F(h(x^*(s), \xi), h(x^*(s'), \xi))$, and by

$$\{\overline{\mathbb{G}}'(s) : s \in S^o\}$$

the Gaussian process with mean zero and covariance structure $\mathrm{Cov}(s, s') = \frac{1}{n_2}\sum_{i=1}^{n_2}(h(x^*(s), \xi_i) - \hat{H}(x^*(s)))(h(x^*(s'), \xi_i) - \hat{H}(x^*(s')))$ where $\hat{H}(x^*(s)) = (1/n_2)\sum_{i=1}^{n_2} h(x^*(s), \xi_i)$ is the sample mean at $x^*(s)$ and $\hat{H}(x^*(s'))$ is the sample mean at $x^*(s')$. For a generic stochastic process $\{Y(\theta) : \theta \in \Theta\}$ over some set Θ, we denote by

$$\psi_{1-\beta}(\{Y(\theta) : \theta \in \Theta\})$$

the $1-\beta$ quantile of $\sup_{\theta\in\Theta} Y(\theta)$. We can formally express the critical value calibrated in Algorithm 9 as $q_{1-\beta} = \psi_{1-\beta}(\{\overline{\mathbb{G}}'(s) : s \in \{s_1, \ldots, s_p\}\})$, and $\bar{q}_{1-\beta} = \psi_{1-\beta}(\{\overline{\mathbb{G}}(s) : s \in S^o\})$. Under Assumption 5.4.2, the Gaussian process $\overline{\mathbb{G}}$ as the weak limit of the empirical process $\{\sqrt{n_2}((1/n_2)\sum_{i=1}^{n_2} h(x^*(s), \xi_i) - H(x^*(s))) : s \in S^o\}$ is a tight Borel measurable element in the space $l^\infty(S^o) := \{f : f \text{ is a function } S^o \to \mathbb{R} \text{ such that } \sup_{s\in S^o}|f(s)| < \infty\}$, so the sample path of $\overline{\mathbb{G}}$ is uniformly continuous with respect to the semimetric $\rho(s, s') := \sqrt{\mathrm{Var}_F(h(x^*(s), \xi) - h(x^*(s'), \xi))}$ almost surely (Example 1.5.10 in Van der Vaart and Wellner (1996)). Note that, under Assumptions 5.4.4 and 5.4.6, on each continuous piece of the solution path this semimetric is continuous in the pair s, s' with respect to the Euclidean metric on S. In other words, almost surely the sample path of $\overline{\mathbb{G}}$ is continuous with respect to the Euclidean metric $d(s, s') := |s - s'|$ on each piece $(\tilde{s}_i, \tilde{s}_{i+1})$. Therefore, by continuity, every countable dense (w.r.t. the Euclidean metric) subset $S_c^o \subset S^o$, e.g., the set of all rational s, renders

$$\sup_{s\in S^o} \overline{\mathbb{G}}(s) = \sup_{s\in S_c^o} \overline{\mathbb{G}}(s) \text{ almost surely.}$$

Suppose $S_c^o = \{s_j^o\}_{j=1}^\infty$, then $\max_{1\le j\le k} \overline{\mathbb{G}}(s_j^o)$ monotonically increases in k towards $\sup_{s\in S^o} \overline{\mathbb{G}}(s)$ al-

most surely, and almost sure convergence implies convergence in distribution therefore $\psi_{1-\beta}(\{\overline{\mathbb{G}}(s) : s \in \{s_1^o, \ldots, s_k^o\}\})$ monotonically increases in k towards the limit $\bar{q}_{1-\beta}$, i.e.,

$$\psi_{1-\beta}(\{\overline{\mathbb{G}}(s) : s \in \{s_1^o, \ldots, s_k^o\}\}) \leq \bar{q}_{1-\beta}, \text{ and } \lim_{k\to\infty} \psi_{1-\beta}(\{\overline{\mathbb{G}}(s) : s \in \{s_1^o, \ldots, s_k^o\}\}) = \bar{q}_{1-\beta}. \quad \text{(D.25)}$$

As the second step, we want to show that $q_{1-\beta}$ converges to $\bar{q}_{1-\beta}$ almost surely. Under Assumptions 5.4.2 and 5.4.3, Theorem 10.6 in Kosorok (2008) states that, for almost every realization of the data sequence $\{\xi_i\}_{i=1}^{\infty}$, the Gaussian process $\overline{\mathbb{G}}'$ weakly converges to $\overline{\mathbb{G}}$ as $n_2 \to \infty$. By the continuous mapping theorem, $\sup_{s\in S^o} \overline{\mathbb{G}}'(s)$ also weakly converges to $\sup_{s\in S^o} \overline{\mathbb{G}}(s)$ almost surely, therefore $\lim_{n_2\to\infty} \psi_{1-\beta}(\{\overline{\mathbb{G}}'(s) : s \in S^o\}) = \bar{q}_{1-\beta}$ almost surely. It is obvious that $q_{1-\beta} \leq \psi_{1-\beta}(\{\overline{\mathbb{G}}'(s) : s \in S^o\})$, hence we have established that $\limsup_{n_2\to\infty,\epsilon_S\to 0} q_{1-\beta} \leq \bar{q}_{1-\beta}$. To show the other direction, we exploit the separability (D.25) of $\overline{\mathbb{G}}$. For each i, let $s_{j_i} \in \{s_1, \ldots, s_p\}$ be such that $s_{j_i} \to s_i^o$ as $\epsilon_S \to 0$. Consider two more Gaussian processes $\{\overline{\mathbb{G}}'(s) : s \in \{s_{j_1}, \ldots, s_{j_k}\}\}$ and $\{\overline{\mathbb{G}}(s) : s \in \{s_{j_1}, \ldots, s_{j_k}\}\}$. For a fixed k, let $\hat{\Sigma}', \hat{\Sigma}$ be the covariance matrices of $\{\overline{\mathbb{G}}'(s) : s \in \{s_{j_1}, \ldots, s_{j_k}\}\}$ and $\{\overline{\mathbb{G}}(s) : s \in \{s_{j_1}, \ldots, s_{j_k}\}\}$ respectively, and let Σ be the covariance matrix of $\{\overline{\mathbb{G}}(s) : s \in \{s_1^o, \ldots, s_k^o\}\}$. Assumption 5.4.4 and the convergence of each s_{j_i} to s_i^o ensure that $\hat{\Sigma} \to \Sigma$ as $\epsilon_S \to 0$. To argue that $\hat{\Sigma}' - \hat{\Sigma} \to \mathbf{0} \in \mathbb{R}^{k\times k}$, where $\mathbf{0}$ denotes the $k \times k$ matrix with zero entries, we need the F-Glivenko-Contelli property of the product class $\mathcal{F} \cdot \mathcal{F} := \{f(\cdot) = h(x, \cdot)h(x', \cdot) : x, x' \in \mathcal{X}\}$. F-Donskerness implies F-Glivenko-Contelli, therefore $\mathcal{F} = \{h(x, \cdot) : x \in \mathcal{X}\}$ is F-Glivenko-Contelli under Assumption 5.4.2, which together with Assumption 5.4.3 forces the product class $\mathcal{F} \cdot \mathcal{F}$ to be F-Glivenko-Contelli by statement (ii) of Corollary 9.27 from Kosorok (2008). As a result, we have

$$\sup_{x,x'\in\mathcal{X}} \Big|\frac{1}{n_2}\sum_{i=1}^{n_2}(h(x, \xi_i) - \hat{H}(x))(h(x', \xi_i) - \hat{H}(x')) - \mathrm{Cov}_F(h(x, \xi), h(x', \xi))\Big| \to 0 \text{ as } n_2 \to \infty$$

almost surely, where $\hat{H}(x) = (1/n_2)\sum_{i=1}^{n_2} h(x, \xi_i)$ and $\hat{H}(x')$ is similar. In particular $\hat{\Sigma}' - \hat{\Sigma} \to \mathbf{0} \in \mathbb{R}^{k\times k}$ as desired. Combining this with the convergence of $\hat{\Sigma}$ to Σ, we conclude $\lim_{n_2\to\infty,\epsilon_S\to 0} \hat{\Sigma}' = \Sigma$ almost surely. Since the distribution of a zero mean multivariate Gaussian is uniquely determined by its covariance matrix, we must have $\{\overline{\mathbb{G}}'(s) : s \in \{s_{j_1}, \ldots, s_{j_k}\}\}$ weakly converges to $\{\overline{\mathbb{G}}(s) :$

$s \in \{s_1^o, \ldots, s_k^o\}\}$, and hence $\psi_{1-\beta}(\{\overline{\mathbb{G}}'(s) : s \in \{s_{j_1}, \ldots, s_{j_k}\}\}) \to \psi_{1-\beta}(\{\overline{\mathbb{G}}(s) : s \in \{s_1^o, \ldots, s_k^o\}\})$ almost surely. Note that $\psi_{1-\beta}(\{\overline{\mathbb{G}}'(s) : s \in \{s_{j_1}, \ldots, s_{j_k}\}\}) \leq \psi_{1-\beta}(\{\overline{\mathbb{G}}'(s) : s \in \{s_1, \ldots, s_p\}\}) = q_{1-\beta}$, hence

$$\liminf_{n_2 \to \infty, \epsilon_S \to 0} q_{1-\beta} \geq \psi_{1-\beta}(\{\overline{\mathbb{G}}(s) : s \in \{s_1^o, \ldots, s_k^o\}\}) \text{ for each } k.$$

This together with (D.25) gives $\liminf_{n_2 \to \infty, \epsilon_S \to 0} q_{1-\beta} \geq \bar{q}_{1-\beta}$. Putting all these together we have $\lim_{n_2 \to \infty, \epsilon_S \to 0} q_{1-\beta} = \bar{q}_{1-\beta}$ almost surely.

The rest of the proof closely follows that of Theorem 5.4.4. We only highlight some modifications. First, each occurrence of $\frac{z_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}}$, $\frac{z_{1-\beta}\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}}$ and $\frac{z_{1-\beta}\hat{\sigma}(x_S^*)}{\sqrt{n_2}}$ shall be replaced by $\frac{q_{1-\beta}}{\sqrt{n_2}}$. Second, the second error E_2 in (D.19) is no longer present, and the series of inequalities in the last paragraph become

$$\begin{aligned}
& \liminf_{n_2 \to \infty, \epsilon_S \to 0} P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \\
\geq & \liminf_{n_2 \to \infty, \epsilon_S \to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(-\frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + \frac{q_{1-\beta}}{\sqrt{n_2}} + o_p\big(\frac{1}{\sqrt{n_2}}\big) \geq 0\Big) \\
= & \liminf_{n_2 \to \infty, \epsilon_S \to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(-\frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + \frac{\bar{q}_{1-\beta}}{\sqrt{n_2}} + o_p\big(\frac{1}{\sqrt{n_2}}\big) \geq 0\Big) \quad \text{since } q_{1-\beta} \to \bar{q}_{1-\beta} \text{ a.s.} \\
= & \liminf_{n_2 \to \infty, \epsilon_S \to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(\frac{\mathbb{G}_{n_2}(x_S^*)}{\sigma(x_S^*)} + o_p(1) \leq \frac{\bar{q}_{1-\beta}}{\sigma(x_S^*)}\Big) \\
= & \Phi\big(\frac{\bar{q}_{1-\beta}}{\sigma(x_S^*)}\big) \quad \text{by Slutsky's theorem.}
\end{aligned}$$

This completes the proof for Algorithm 9.

Now we prove the results for Algorithm 10 by a similar argument. Consider the Gaussian process

$$\{\tilde{\mathbb{G}}(s) : s \in S^o\}$$

with mean zero and covariance $\mathrm{Cov}(s, s') = \mathrm{Cov}_F(h(x^*(s), \xi), h(x^*(s'), \xi))/(\sigma(x^*(s))\sigma(x^*(s')))$, and the Gaussian process

$$\{\tilde{\mathbb{G}}'(s) : s \in S^o\}$$

with mean zero and covariance

$$\mathrm{Cov}(s,s') = \big[\frac{1}{n_2}\sum_{i=1}^{n_2}(h(x^*(s),\xi_i) - \hat{H}(x^*(s)))(h(x^*(s'),\xi_i) - \hat{H}(x^*(s')))\big]/(\hat{\sigma}(x^*(s))\hat{\sigma}(x^*(s')))$$

where $\hat{H}(x^*(s)) = \frac{1}{n_2}\sum_{i=1}^{n_2} h(x^*(s),\xi_i)$, $\hat{\sigma}^2(x^*(s)) = \frac{1}{n_2}\sum_{i=1}^{n_2}(h(x^*(s),\xi_i) - \hat{H}(x^*(s)))^2$. $\hat{H}(x^*(s'))$ and $\hat{\sigma}^2(x^*(s'))$ are similarly defined. We have $\tilde{q}_{1-\beta} = \psi_{1-\beta}(\{\tilde{\mathbb{G}}(s) : s \in S^o\})$, and $q_{1-\beta} = \psi_{1-\beta}(\{\tilde{\mathbb{G}}'(s) : s \in \{s_1,\ldots,s_p\}\}) \le \psi_{1-\beta}(\{\tilde{\mathbb{G}}'(s) : s \in S^o\})$. Under the depicted conditions, Lemma 3 from Lam (2019) states that, for almost every realization of the data sequence $\{\xi_i\}_{i=1}^{\infty}$, the Gaussian process $\tilde{\mathbb{G}}'$ weakly converges to $\tilde{\mathbb{G}}$, so we have

$$\limsup_{n_2\to\infty,\epsilon_S\to 0} q_{1-\beta} \le \limsup_{n_2\to\infty,\epsilon_S\to 0} \psi_{1-\beta}(\{\tilde{\mathbb{G}}'(s) : s \in S^o\}) = \tilde{q}_{1-\beta}$$

almost surely. By a similar argument based on the separability of $\tilde{\mathbb{G}}$ and the uniform convergence of covariance as in the case of Algorithm 9, we can show the other direction $\liminf_{n_2\to\infty,\epsilon_S\to 0} q_{1-\beta} \ge \tilde{q}_{1-\beta}$ and thereby conclude that $\lim_{n_2\to\infty,\epsilon_S\to 0} q_{1-\beta} = \tilde{q}_{1-\beta}$ almost surely. The rest of the proof for Algorithm 10 also follows that of Theorem 5.4.4, but with each occurrence of $z_{1-\beta}$ replaced by $q_{1-\beta}$. The display in the last paragraph should be modified to be

$$\begin{aligned}
& \liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \ge \gamma) \\
\ge\ & \liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(-\frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + \frac{q_{1-\beta}\hat{\sigma}(x_S^*)}{\sqrt{n_2}} + o_p\big(\frac{1}{\sqrt{n_2}}\big) \ge 0\Big) \\
=\ & \liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(-\frac{1}{\sqrt{n_2}}\mathbb{G}_{n_2}(x_S^*) + \frac{\tilde{q}_{1-\beta}\sigma(x_S^*)}{\sqrt{n_2}} + o_p\big(\frac{1}{\sqrt{n_2}}\big) \ge 0\Big) \\
& \quad \text{since } q_{1-\beta}\to\tilde{q}_{1-\beta} \text{ and } \hat{\sigma}(x_S^*)\to\sigma(x_S^*) > 0 \text{ a.s.} \\
=\ & \liminf_{n_2\to\infty,\epsilon_S\to 0} P_{\boldsymbol{\xi}_{1:n_2}}\Big(\frac{\mathbb{G}_{n_2}(x_S^*)}{\sigma(x_S^*)} + o_p(1) \le \tilde{q}_{1-\beta}\Big) \\
=\ & \Phi(\tilde{q}_{1-\beta}) \ \text{ by Slutsky's theorem.}
\end{aligned}$$

Lastly, by stochastic dominance of the supremum of the Gaussian process $\overline{\mathbb{G}}$ or $\tilde{\mathbb{G}}$ over each of its marginal Gaussian component, it is straightforward that $\bar{q}_{1-\beta} \ge z_{1-\beta}\sigma(x_S^*)$ and that $\tilde{q}_{1-\beta} \ge z_{1-\beta}$, therefore both $\Phi\big(\frac{\bar{q}_{1-\beta}}{\sigma(x_S^*)}\big)$ and $\Phi(\tilde{q}_{1-\beta})$ are at least $1-\beta$. □

D.4 Proofs of Results in Section 5.5

We first provide a lemma on the continuity of the solution path $x^*(s)$:

Lemma D.4.1 *Suppose the formulation $OPT(s)$ satisfies Assumptions 5.4.8 and 5.5.1-5.5.3. If $\hat{\mathcal{F}}(s) = \{x : g_t(x,s) \leq 0, t = 1,\ldots,T\}$ for some finite T where each g_t is jointly continuous in x, s and convex in x for every fixed s, the objective $f(x)$ is continuous, and $OPT(s) := \min\{f(x) : x \in \mathcal{X} \cap \hat{\mathcal{F}}(s)\}$ has a unique solution $x^*(s)$ for all $s \in [\underline{s}, \overline{s}]$, then the solution path $x^*(s)$ is continuous on $[\underline{s}, \overline{s}]$.*

Proof of Lemma D.4.1. The lemma is an application of Proposition 4.4 from Bonnans and Shapiro (2013). Based on the discussion following Proposition 4.4 in Bonnans and Shapiro (2013), we argue one by one that $OPT(s)$ satisfies assumptions (i)-(iv) of Proposition 4.4. Assumption (i): The objective $f(x)$ is continuous and independent of s hence it's jointly continuous in x, s. Assumption (ii): The constraints of $OPT(s)$ can be formulated as $(f_1(x,s),\ldots,f_R(x,s), w_1'x - z_1,\ldots,w_L'x - z_L, g_1(x,s),\ldots,g_T(x,s)) \in [0,+\infty)^{R+L+T}$, where the left hand side is a vector of continuous functions and the right hand size is a closed convex cone. Assumption (iii) is implied by our Assumptions 5.4.8 and 5.5.3. Assumption (iv): Since $OPT(s)$ is convex, our Assumptions 5.4.8 and 5.5.2 ensure Slater's condition for $OPT(s)$ for all $s \in [s_l, s_u]$, and Slater's condition implies Robinson's constraint qualification, a sufficient condition for assumption (iv). Therefore the set-valued mapping $\mathcal{X}^*(s) := \{x \in \mathcal{X} \cup \hat{\mathcal{F}}(s) : f(x) = \min_{x \in \mathcal{X} \cup \hat{\mathcal{F}}(s)} f(x)\}$ is upper semicontinuous at every $s \in [\underline{s}, \overline{s}]$. When the optimal solution $x^*(s)$ for $OPT(s)$ is unique, upper semicontinuity implies continuity, hence $x^*(s)$ is continuous on $[\underline{s}, \overline{s}]$. □

The second lemma we present concerns the uniqueness of $x^*(s)$ for linear objectives:

Lemma D.4.2 *Consider an optimization problem in the form of $\min c'x$ subject to $f_r(x) \leq 0$ for $r = 1,\ldots,R$ and $Ax \leq b$ for $A = [a_1,\ldots,a_L]' \in \mathbb{R}^{L\times d}$ and $b \in \mathbb{R}^L$ where each f_r is continuous and convex and c is a non-zero vector. For each f_r and any two solutions $x_1 \neq x_2$ such that $f_r(x_1) = f_r(x_2) = 0$, assume $f_r(\theta x_1 + (1-\theta)x_2) < 0$ for any $\theta \in (0,1)$. If any $k \leq d-1$ rows of A does not satisfy the SCI condition, then the optimal solution must be unique whenever one exists.*

Proof of Lemma D.4.2. Suppose there are two optimal solutions x_1, x_2. By convexity any solution in the form $\theta x_1 + (1-\theta)x_2$ for $\theta \in [0,1]$ is also optimal, and because of the condition on f_r we can assume that $f_r(x_1) < 0, f_r(x_2) < 0$ for all $r = 1, \ldots, R$. Therefore, only the linear constraints can be binding on the line segment $\theta x_1 + (1-\theta)x_2, \theta \in [0,1]$. Let $A_o x \leq b_o$ be the binding linear constraints on the segment where A_o consists of rows of A and b_o contains the corresponding components of b, then it is clear that solution of the form $\theta x_1 + (1-\theta)x_2$ is optimal for the linear program $\min c'x$ subject to $A_o x \leq b_o$. Since $A_o x_1 = A_o x_2 = b_o$, we have $A_o(x_2 - x_1) = 0$ hence the rank of A_o is at most $d-1$. Now consider the dual $\min b_o' y$ subject to $A_o' y = -c, y \geq 0$. Since the rank of A_o is at most $d-1$, by removing linearly dependent rows, the constraint $A_o' y = -c$ can be simplified to $\tilde{A}_o' y = -\tilde{c}$ where $\tilde{A}_o'$ has at most $d-1$ linearly independent rows. Let y^* be an optimal basic feasible solution of the dual with the simplified constraint $\tilde{A}_o' y = -\tilde{c}$, then y^* has at most $d-1$ non-zero (positive) components. However as a feasible solution y^* has to satisfy $A_o' y^* = -c$ therefore the SCI condition holds for the rows of A corresponding to the positive components of y^*, leading to a contradiction. □

Proof of Theorem 5.5.2. We only need to verify the conditions of Lemma D.4.1. In both cases (i) and (ii), $\hat{\mathcal{F}}(s) = \{x : \gamma + s - \frac{1}{n}\sum_{i=1}^n h(x, \xi_i) \leq 0\}$ and $\gamma + s - \frac{1}{n}\sum_{i=1}^n h(x, \xi_i)$ is obviously jointly continuous in x, s and convex in x, and also $f(x)$ is continuous. Therefore, it only remains to check uniqueness of $x^*(s)$ in order to apply Lemma D.4.1.

In case (i), the strict convexity of $f(x)$ forces the solution $x^*(s)$ to be unique. In case (ii), we first treat the case when h is linear in x. We first note that for such h the SAA takes the form $-\big(\frac{1}{n}\sum_{i=1}^n A(\xi_i)\big)' x \leq \frac{1}{n}\sum_{i=1}^n b(\xi_i) - \gamma - s$. Therefore each constraint of $OPT(s)$ is either linear or strictly convex, and thanks to Lemma D.4.2 it remains to show that the SCI condition is not satisfied for each $s \in S$. For any $k \leq d-2$ rows $\{w_{l(1)}, \ldots, w_{l(k)}\}$ of W and the coefficient vector $-\frac{1}{n}\sum_{i=1}^n A(\xi_i)$, we want to show that the SCI condition does not hold for $\{w_{l(1)}, \ldots, w_{l(k)}, -\frac{1}{n}\sum_{i=1}^n A(\xi_i)\}$. Suppose SCI does hold, then we have the representation $-\frac{1}{n}\sum_{i=1}^n A(\xi_i) = \sum_{j=1}^k \lambda_j w_{l(j)} + \lambda_c c$, i.e., $-\frac{1}{n}\sum_{i=1}^n A(\xi_i)$ lies in the subspace of dimension spanned by $\{w_{l(1)}, \ldots, w_{l(k)}, c\}$. However, $-\frac{1}{n}\sum_{i=1}^n A(\xi_i)$ has a density hence lies in any given subspace of dimension $\leq d-1$ with probability zero. Therefore almost surely SCI does not hold for

$\{w_{l(1)}, \ldots, w_{l(k)}, -\frac{1}{n}\sum_{i=1}^{n} A(\xi_i)\}$. If only linear coefficients from $Wx \leq z$ are considered, SCI condition is again not satisfied by the condition imposed. Therefore almost surely SCI is not satisfied for $OPT(s)$. By noting that the SCI condition is independent of s since s is on the right hand side, we conclude that almost surely SCI is not satisfied for all $s \in S$. When $h(x, \xi)$ is strictly concave in x, Lemma D.4.2 can be directly applied to show the uniqueness of $x^*(s)$. □

Proof of Theorem 5.5.4. We first argue that the constraint function

$$g(x, s) = \inf \Big\{ \sum_{i=1}^{n} w_i h(x, \xi_i) : \sum_{i=1}^{n} \frac{1}{n}\phi(nw_i) \leq s, \sum_{i=1}^{n} w_i = 1, w_i \geq 0 \text{ for all } i \Big\}$$

is jointly continuous in x, s. Viewing both x, s as parameters of the optimization problem defining $g(x, s)$, one can easily check that the assumptions of Proposition 4.4 from Bonnans and Shapiro (2013) are satisfied, hence $g(x, s)$ as the optimal value of the optimization problem is continuous in the parameters x, s. It is also obvious that $g(x, s)$ is concave in x for every s because of its representation as the minimum of a family of concave functions. By Lemma D.4.1 it remains to show the uniqueness of $x^*(s)$.

Case (i) follows from the strict convexity of f as in Theorem 5.5.2. In case (ii), we would like to show that the constraint function $g(x, s)$ is strictly concave in x. Indeed, due to compactness an optimal weight vector $\mathbf{w}^*$ must exist for the minimization problem defining $g(x, s)$. Consider $x_1 \neq x_2$ and $\theta \in (0, 1)$, and let $\mathbf{w}^*$ be the minimizing weight vector that gives the worst-case value $g(\theta x_1 + (1-\theta)x_2, s)$ at the solution $\theta x_1 + (1-\theta)x_2$. Then because of the strict concavity of $h(x, \xi)$ in x, we have

$$\begin{aligned} g(\theta x_1 + (1-\theta)x_2, s) &= \sum_{i=1}^{n} w_i^* h(\theta x_1 + (1-\theta)x_2, \xi_i) \\ &> \sum_{i=1}^{n} w_i^* (\theta h(x_1, \xi_i) + (1-\theta) h(x_2, \xi_i)) \\ &\geq \theta g(x_1, s) + (1-\theta) g(x_2, s). \end{aligned}$$

Therefore $g(x, s)$ is strictly concave in x, and uniqueness of $x^*(s)$ follows from the SCI condition not being satisfied and applying Lemma D.4.2. In case (iii), the strict concavity of $g(x, s)$ can be shown

as follows. Due to the strict convexity of ϕ, for each decision x the minimizing weight vector $\mathbf{w}^*$ not only exists but also is unique. Let $x_1 \neq x_2$, then by the condition there must be some $\theta' \in [0,1)$ such that $\widehat{\text{Corr}}(x_1, \theta' x_1 + (1-\theta')x_2) \neq 1$, and let $\mathbf{w}^{1*}, \mathbf{w}^{\theta'*}$ be the respective minimizing weight vectors. We argue that $\mathbf{w}^{1*} \neq \mathbf{w}^{\theta'*}$. The condition $\lim_{x\to 0+} \phi(x) = +\infty$ ensures positive components of the minimizing weight vector, hence the optimality condition entails $h(x_1, \xi_i) = \lambda_1 \phi'(n w_i^{1*}) - \beta_1$ and $h(\theta' x_1 + (1-\theta')x_2, \xi_i) = \lambda_{\theta'} \phi'(n w_i^{\theta'*}) - \beta_{\theta'}$ for all $i = 1, \ldots, n$ and some constants $\lambda_1, \beta_1, \lambda_{\theta'}, \beta_{\theta'}$. Therefore the empirical correlation between $h(x_1, \xi)$ and $h(\theta' x_1 + (1-\theta')x_2, \xi)$ takes the form

$$\widehat{\text{Corr}}(x_1, \theta' x_1 + (1-\theta')x_2) = \frac{(1/n)\sum_{i=1}^n (\phi'(n w_i^{1*}) - \bar{\phi}_1')(\phi'(n w_i^{\theta'*}) - \bar{\phi}_{\theta'}')}{\sqrt{(1/n)\sum_{i=1}^n (\phi'(n w_i^{1*}) - \bar{\phi}_1')^2 (1/n)\sum_{i=1}^n (\phi'(n w_i^{\theta'*}) - \bar{\phi}_{\theta'}')^2}}$$

where $\bar{\phi}_1' = (1/n)\sum_{i=1}^n \phi'(n w_i^{1*}), \bar{\phi}_{\theta'}' = (1/n)\sum_{i=1}^n \phi'(n w_i^{\theta'*})$. If $\mathbf{w}^{1*} = \mathbf{w}^{\theta'*}$, we must have $\widehat{\text{Corr}}(x_1, \theta' x_1 + (1-\theta')x_2) = 1$, a contradiction. Therefore, if $\theta' > 0$, we have $g(\theta' x_1 + (1-\theta')x_2, s) = \sum_{i=1}^n w_1^{\theta'*} h(\theta' x_1 + (1-\theta')x_2, \xi_i) \geq \theta' \sum_{i=1}^n w_1^{\theta'*} h(x_1, \xi_i) + (1-\theta')\sum_{i=1}^n w_1^{\theta'*} h(x_2, \xi_i) > \theta' g(x_1, s) + (1-\theta') g(x_2, s)$, hence $g(\theta x_1 + (1-\theta)x_2, s) > \theta g(x_1, s) + (1-\theta) g(x_2, s)$ for all $\theta \in (0,1)$ by the (non-strict) concavity of $g(x, s)$ in x. Otherwise, if $\theta' = 0$, i.e., $\widehat{\text{Corr}}(x_1, x_2) \neq 1$, then by continuity there exists a small enough $\theta > 0$ such that $\widehat{\text{Corr}}(x_1, \theta x_1 + (1-\theta)x_2) \neq 1$, hence things reduce to the previous case. □

Proof of Theorem 5.5.5. Strong duality results from Gao and Kleywegt (2016) or Blanchet et al. (2019) show that the constraint function takes the form

$$g(x, s) := \inf\left\{\mathbb{E}_G[h(x,\xi)] : d_p(G, F_n) \leq s\right\} = \sup_{\lambda \geq 0}\left\{\frac{1}{n}\sum_{i=1}^n \inf_{\xi \in \Xi}(h(x,\xi) + \lambda \left\|\xi - \xi_i\right\|^p) - \lambda s^p\right\}.$$

We would like to show that $g(x, s)$ is jointly continuous in x, s. Let $a(x, \lambda) = \frac{1}{n}\sum_{i=1}^n \inf_{\xi\in\Xi}(h(x,\xi) + \lambda \left\|\xi - \xi_i\right\|^p)$. It is clear that $a(x,\lambda) - \lambda s^p \leq \frac{1}{n}\sum_{i=1}^n h(x, \xi_i) - \lambda s^p \leq \frac{1}{n}\sum_{i=1}^n h(x,\xi_i) - \lambda s_l^p$ by taking $\xi = \xi_i$ in each infimum, and that $a(x, 0) = \inf_{\xi\in\Xi} h(x,\xi)$. For each x choose $\overline{\lambda}(x)$ so that $\frac{1}{n}\sum_{i=1}^n h(x,\xi_i) - \overline{\lambda}(x) s_l^p = \inf_{\xi\in\Xi} h(x,\xi)$. Since $h(x,\xi)$ is jointly continuous in x, ξ and Ξ is compact, $h(x,\xi)$ is uniformly continuous in x, ξ on $[x_o - \delta, x_o + \delta] \times \Xi$ for given x_o and $0 < \delta < \infty$. This uniform continuity implies that the infimum $\inf_{\xi\in\Xi} h(x,\xi)$ is continuous in x because

$|\inf_{\xi\in\Xi} h(x,\xi) - \inf_{\xi\in\Xi} h(x_o,\xi)| \leq \sup_{\xi\in\Xi}|h(x,\xi) - h(x_o,\xi)| \to 0$ as $x \to x_o$. Therefore $\overline{\lambda}(x)$ is continuous in x and

$$g(x,s) = \sup_{0\leq\lambda\leq\overline{\lambda}(x)} (a(x,\lambda) - \lambda s^p) = \sup_{0\leq\lambda\leq\overline{\lambda}(x)} \Big\{\frac{1}{n}\sum_{i=1}^{n} \inf_{\xi\in\Xi}(h(x,\xi) + \lambda\,\|\xi - \xi_i\|^p) - \lambda s^p\Big\}.$$

Since Ξ is compact, by an argument similar to the one used to prove the continuity of $\inf_{\xi\in\Xi} h(x,\xi)$ we see that $a(x,\lambda)$ is jointly continuous in x,λ. Continuity of $a(x,\lambda)$ and $\overline{\lambda}(x)$ leads to the joint continuity of $g(x,s)$ in x,s. To explain, for a fixed x and some $\delta > 0$, define $\overline{\lambda}_\delta := \sup_{x' \text{ s.t. } \|x'-x\|_2\leq\delta} \overline{\lambda}(x')$, so for all x', s' such that $\|x' - x\|_2 \leq \delta$ and $|s' - s| \leq \delta$ we have

$$\begin{aligned} |g(x',s') - g(x,s)| &= |\sup_{0\leq\lambda\leq\overline{\lambda}_\delta} (a(x',\lambda) - \lambda s'^p) - \sup_{0\leq\lambda\leq\overline{\lambda}_\delta} (a(x,\lambda) - \lambda s^p)| \\ &\leq \sup_{0\leq\lambda\leq\overline{\lambda}_\delta} |a(x',\lambda) - a(x,\lambda)| + \overline{\lambda}_\delta |s'^p - s^p| \to 0 \text{ as } x' \to x, s' \to s \end{aligned}$$

where the limit holds because $a(x,\lambda)$ is uniformly continuous on the compact set $\{x' : \|x' - x\| \leq \delta\} \times [0, \overline{\lambda}_\delta]$. Concavity of $g(x,s)$ in x holds because for any probability measure G the expectation $\mathbb{E}_G[h(x,\xi)]$ is concave in x and the infimum operation perserves concavity.

In order to utilize Lemma D.4.1, it remains to prove uniqueness of $x^*(s)$ for all $s \in S$. In case (i) uniqueness trivially follows from strict convexity of f. In case (ii), we first establish a result concerning the existence of the worst-case distribution:

Lemma D.4.3 *Under the same conditions of Theorem 5.5.5, if $g(x,s) = a(x,\lambda^*) - \lambda^* s^p$ for some $\lambda^* > 0$, then there exists a distribution G^* that belongs to the Wasserstein ball and that achieves the worst-case expectation, i.e., $g(x,s) = \mathbb{E}_{G^*}[h(x,\xi)]$.*

Proof of Lemma D.4.3. This is a direct consequence of Corollary 1 in Gao and Kleywegt (2016). Note that, since Ξ is compact and $h(x,\xi)$ is continuous, for each decision x the quantity $a(x,\lambda)$ is finite for all $\lambda \geq 0$. Corollary 1 from Gao and Kleywegt (2016) then entails the existence of the worst-case distribution if there exists a dual maximizer $\lambda^* > 0$. □

Consider $x_1 \neq x_2$. If there exists some $\theta' \in (0,1)$ such that $g(\theta' x_1 + (1-\theta')x_2, s) = a(\theta' x_1 + (1-\theta')x_2, \lambda') - \lambda' s^p$ for some $\lambda' > 0$, then there exists some distribution $G_{\theta'}$ in the Wasserstein ball

generating the worst-case expectation $g(\theta' x_1 + (1-\theta')x_2, s)$. The strict concavity of h then implies

$$\begin{aligned} g(\theta' x_1 + (1-\theta')x_2, s) &= \mathbb{E}_{G_{\theta'}}[h(\theta' x_1 + (1-\theta')x_2, \xi)] \\ &> \mathbb{E}_{G_{\theta'}}[\theta' h(x_1, \xi) + (1-\theta')h(x_2, \xi)] \\ &\geq \theta' g(x_1, s) + (1-\theta')g(x_2, s). \end{aligned}$$

Since $g(x, s)$ is (non-strictly) concave in x, the above strict inequality at a certain θ' extends to all other θ, i.e., $g(\theta x_1 + (1-\theta)x_2, s) > \theta g(x_1, s) + (1-\theta)g(x_2, s)$ for all $\theta \in (0, 1)$. Otherwise if $g(\theta x_1 + (1-\theta)x_2, s) = a(\theta x_1 + (1-\theta)x_2, 0) > a(\theta x_1 + (1-\theta)x_2, \lambda) - \lambda s^p$ for all $\theta \in (0, 1)$ and $\lambda > 0$, since $a(x, 0) = \inf_{\xi \in \Xi} h(x, \xi)$ we still have the strict concavity of $g(\theta x_1 + (1-\theta)x_2, s)$ in θ. Therefore, according to Lemma D.4.2, the solution $x^*(s)$ is unique. □

Proof of Theorem 5.5.6. We first show that each $g_i(x, s) := \sup_{(\mu,\Sigma)\in\mathcal{U}_i(s)} \mu' x + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x'\Sigma x}$ is jointly continuous in x and s. For a fixed pair x_o, s_o and an arbitrary pair x, s, we write

$$\begin{aligned} & |g_i(x, s) - g_i(x_o, s_o)| \\ \leq\ & |g_i(x_o, s) - g_i(x_o, s_o)| + |g_i(x, s) - g_i(x_o, s)| \\ \leq\ & |g_i(x_o, s) - g_i(x_o, s_o)| + \sup_{(\mu,\Sigma)\in\mathcal{U}_i(s)} |\mu' x + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x'\Sigma x} - (\mu' x_o + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_o'\Sigma x_o})| \\ \leq\ & |g_i(x_o, s) - g_i(x_o, s_o)| + \sup_{(\mu,\Sigma)\in\mathcal{U}_i(s_u)} |\mu' x + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x'\Sigma x} - (\mu' x_o + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_o'\Sigma x_o})| \\ & \text{where } s_u \text{ is the maximum value for } s. \end{aligned} \tag{D.26}$$

Note that $\mu' x + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x'\Sigma x}$ as a function jointly in μ, Σ, x is continuous, and hence by the compactness of $\mathcal{U}_i(s_u)$ is uniformly continuous for $(\mu, \Sigma) \in \mathcal{U}_i(s_u)$ and x in some neighborhood of x_o. Uniform continuity implies that the second term in (D.26) vanishes as $x \to x_o$. It remains to show that the first term in (D.26) also vanishes, i.e., $g_i(x_o, s) \to g_i(x_o, s_o)$, as $s \to s_o$. We first show

that as $s \to s_o$

$$\sup_{(\mu_s,\Sigma_s)\in\mathcal{U}_i(s)} \inf_{(\mu_{s_o},\Sigma_{s_o})\in\mathcal{U}_i(s_o)} (\|\mu_s - \mu_{s_o}\|_2 + \|\Sigma_s - \Sigma_{s_o}\|_2) \to 0, \tag{D.27}$$

$$\sup_{(\mu_{s_o},\Sigma_{s_o})\in\mathcal{U}_i(s_o)} \inf_{(\mu_s,\Sigma_s)\in\mathcal{U}_i(s)} (\|\mu_s - \mu_{s_o}\|_2 + \|\Sigma_s - \Sigma_{s_o}\|_2) \to 0. \tag{D.28}$$

We prove (D.27) by contradiction. Suppose there exists $(\mu_{s_k}, \Sigma_{s_k}) \in \mathcal{U}_i(s_k)$ and $s_k \to s_o$ such that $\inf_{(\mu_{s_o},\Sigma_{s_o})\in\mathcal{U}_i(s_o)}(\|\mu_{s_k} - \mu_{s_o}\|_2 + \|\Sigma_{s_k} - \Sigma_{s_o}\|_2) > \epsilon$ for some $\epsilon > 0$. Note that all $(\mu_{s_k}, \Sigma_{s_k})$ lie in the compact set $\mathcal{U}_i(s_u)$, hence there is a subsequence converging to some $(\mu_\infty, \Sigma_\infty)$ such that $\inf_{(\mu_{s_o},\Sigma_{s_o})\in\mathcal{U}_i(s_o)}(\|\mu_\infty - \mu_{s_o}\|_2 + \|\Sigma_\infty - \Sigma_{s_o}\|_2) \geq \epsilon$, i.e., $(\mu_\infty, \Sigma_\infty) \notin \mathcal{U}_i(s_o)$. Since $\cap_{s>s_o}\mathcal{U}_i(s) = \mathcal{U}_i(s_o)$ and $\mathcal{U}_i(s)$ is non-decreasing in s, there exists some $\delta > 0$ such that $(\mu_\infty, \Sigma_\infty) \notin \mathcal{U}_i(s)$ for all $s \leq s_o + \delta$, a contradiction with the convergence to $(\mu_\infty, \Sigma_\infty)$. To show (D.28), suppose there exists $(\mu_k, \Sigma_k) \in \mathcal{U}_i(s_o)$ and $s_k \to s_o$ such that

$$\inf_{(\mu_{s_k},\Sigma_{s_k})\in\mathcal{U}_i(s_k)} (\|\mu_k - \mu_{s_k}\|_2 + \|\Sigma_k - \Sigma_{s_k}\|_2) > \epsilon \tag{D.29}$$

for some $\epsilon > 0$. By compactness, assume without loss of generality that (μ_k, Σ_k) converges to some limit $(\mu_\infty, \Sigma_\infty) \in \mathcal{U}_i(s_o)$. However, the condition $\overline{\cup_{s<s_o}\mathcal{U}_i(s)} = \mathcal{U}_i(s_o)$ ensures that $\inf_{(\mu_{s_k},\Sigma_{s_k})\in\mathcal{U}_i(s_k)}(\|\mu_\infty - \mu_{s_k}\|_2 + \|\Sigma_\infty - \Sigma_{s_k}\|_2) \to 0$ as $s_k \to s_o$, which further entails that

$$\begin{aligned}
& \inf_{(\mu_{s_k},\Sigma_{s_k})\in\mathcal{U}_i(s_k)} (\|\mu_k - \mu_{s_k}\|_2 + \|\Sigma_k - \Sigma_{s_k}\|_2) \\
\leq\ & \inf_{(\mu_{s_k},\Sigma_{s_k})\in\mathcal{U}_i(s_k)} (\|\mu_\infty - \mu_{s_k}\|_2 + \|\Sigma_\infty - \Sigma_{s_k}\|_2 + \|\mu_\infty - \mu_k\|_2 + \|\Sigma_\infty - \Sigma_k\|_2) \\
\leq\ & \inf_{(\mu_{s_k},\Sigma_{s_k})\in\mathcal{U}_i(s_k)} (\|\mu_\infty - \mu_{s_k}\|_2 + \|\Sigma_\infty - \Sigma_{s_k}\|_2) + \|\mu_\infty - \mu_k\|_2 + \|\Sigma_\infty - \Sigma_k\|_2 \\
\to\ & 0
\end{aligned}$$

a contradiction with (D.29). This proves (D.28). Now we use (D.27) and (D.28) to conclude $g_i(x_o, s) \to g_i(x_o, s_o)$ as $s \to s_o$. Since $\mathcal{U}_i(s_o)$ is compact, there exists an $(\mu^*_{s_0}, \Sigma^*_{s_o}) \in \mathcal{U}_i(s_o)$ such that $g_i(x_o, s_o) = \mu^{*\prime}_{s_o} x_o + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_o' \Sigma^*_{s_o} x_o}$. (D.28) entails that there exists some $(\mu_s, \Sigma_s) \in \mathcal{U}_i(s)$ for each s such that $(\mu_s, \Sigma_s) \to (\mu^*_{s_0}, \Sigma^*_{s_o})$, therefore $\liminf_{s\to s_o} g_i(x_o, s) \geq \liminf_{s\to s_o} \mu_s' x_o +$

$\sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_o'\Sigma_s x_o} = g_i(x_o, s_o)$. On the other hand, for each s, compactness of $\mathcal{U}_i(s)$ implies the existence of some $(\mu_s^*, \Sigma_s^*) \in \mathcal{U}_i(s)$ such that $g_i(x_o, s) = \mu_s^{*\prime} x_o + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_o'\Sigma_s^* x_o}$. (D.27) then implies that there exists corresponding $(\mu_s^o, \Sigma_s^o) \in \mathcal{U}_i(s_o)$ such that $\|\mu_s^o - \mu_s^*\| \to 0$ and $\|\Sigma_s^o - \Sigma_s^*\| \to 0$ as $s \to s_o$. Since $\mu' x_o + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_o'\Sigma x_o}$ as a function of (μ, Σ) is uniformly continuous on $\mathcal{U}_i(s_u)$, we have $\limsup_{s\to s_o} g_i(x_o, s) = \limsup_{s\to s_o} \mu_s^{*\prime} x_o + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_o'\Sigma_s^* x_o} = \limsup_{s\to s_o} \mu_s^{o\prime} x_o + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_o'\Sigma_s^o x_o} \le g_i(x_o, s_o)$. Altogether we have shown that $g_i(x_o, s) \to g_i(x_o, s_o)$, hence g_i is jointly continuous in x, s.

Secondly, we show the uniqueness of $x^*(s)$ so that the desired result follows from applying Lemma D.4.1. Note that the supremum of a family of convex functions is still convex, therefore each g_i is convex in x. In case (i), strictly convexity of f automatically forces uniqueness of $x^*(s)$. In case (ii), we prove uniqueness by either condition (3) or condition (4). Consider $x_1 \neq x_2$ such that $g_i(x_1, s) = g_i(x_2, s) = b_i$, and $x_\theta := (1-\theta)x_1 + \theta x_2$ for some $\theta \in (0, 1)$. Note that it is impossible that $x_1 = cx_2$ or $x_2 = cx_1$ for some $c \ge 0$ and $c \neq 1$, because otherwise $g_i(x_1, s) = cg_i(x_2, s) = cb_i \neq b_i$ or $g_i(x_2, s) = cg_i(x_1, s) = cb_i \neq b_i$. Let $(\mu_\theta, \Sigma_\theta) \in \mathcal{U}_i(s)$ be such that $g_i(x_\theta, s) = \mu_\theta' x_\theta + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_\theta'\Sigma_\theta x_\theta}$. Under condition (3), Σ_θ is automatically positive definite. Under condition (4), Σ_θ can be taken to be Σ_s because $x_\theta'\Sigma_s x_\theta - x_\theta'\Sigma_\theta x_\theta = x_\theta'(\Sigma_s - \Sigma_\theta)x_\theta \ge 0$, where the last inequality is due to $\Sigma_\theta \preceq \Sigma_s$. That is, in either case, Σ_θ can be taken to be positive definite. We then follow the proof of Theorem 5.5.8 to show that

$$\begin{aligned} g_i(x_\theta, s) &< (1-\theta)\big(\mu_\theta' x_1 + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_1'\Sigma_\theta x_1}\big) + \theta\big(\mu_\theta' x_2 + \sqrt{\frac{1-\alpha_i}{\alpha_i}}\sqrt{x_2'\Sigma_\theta x_2}\big) \\ &\le (1-\theta)g_i(x_1, s) + \theta g_i(x_2, s) \end{aligned}$$

and to conclude uniqueness of $x^*(s)$ for each s using Lemma D.4.2. □

Proof of Theorem 5.5.7. We first transform the infinitely constrained robust counterpart into finitely many constraints. Note that, since each uncertainty set $\mathcal{U}_i(s)$ is a bounded polytope, in each robust constraint $\max_{a_i \in \mathcal{U}_i(s)} a_i' x \le b_i$ the maximum is attained at a vertex of $\mathcal{U}_i(s)$. The set

of vertices of $\mathcal{U}_i(s)$ takes the form

$$\mathcal{V}_i(s) = \left\{ \widetilde{\mathcal{W}}_i^{-1} z_i + s\widetilde{\mathcal{W}}_i^{-1} e_i : \begin{array}{l} \widetilde{\mathcal{W}}_i \in \mathbb{R}^{d\times d} \text{ is an invertible submatrix of } \mathcal{W}_i \\ \mathcal{W}_i\widetilde{\mathcal{W}}_i^{-1} z_i - z_i \leq s(e_i - \mathcal{W}_i\widetilde{\mathcal{W}}_i^{-1} e_i) \end{array} \right\}$$

where the second condition ensures that $\widetilde{\mathcal{W}}_i^{-1} z_i + s\widetilde{\mathcal{W}}_i^{-1} e_i \in \mathcal{U}_i(s)$. The robust counterpart then becomes $v_i' x \leq b_i, v_i \in \mathcal{V}_i(s)$ for all $i = 1, \ldots, K$. We make two important observations for $\mathcal{V}_i(s)$. First, the number of elements in $\mathcal{V}_i(s)$ is no more than the number of square submatrices of $\mathcal{W}_i$ which is finite. Second, the right hand side of $\mathcal{W}_i\widetilde{\mathcal{W}}_i^{-1} z_i - z_i \leq s(e_i - \mathcal{W}_i\widetilde{\mathcal{W}}_i^{-1} e_i)$ is linear in s hence the system of inequalities are valid for s in some interval of the form $(-\infty, u], [l, -\infty)$ or $[l, u]$, therefore the set of bases corresponding to vertices in $\mathcal{V}_i(s)$ changes at only finitely many s values. That is, there are $s_l = s_0' \leq s_1' < \cdots < s_{q-1}' < s_q' = s_u$ such that, for each $1 \leq j \leq q$, there exist submatrices $\widetilde{\mathcal{W}}_{i,1}, \widetilde{\mathcal{W}}_{i,2}, \ldots, \widetilde{\mathcal{W}}_{i,k_{i,j}}$ of each $\mathcal{W}_i$ such that the polyhedral RO $OPT(s)$ for all $s \in [s_{j-1}', s_j']$ can be simply expressed as

$$\begin{array}{ll} \min\limits_{x\in\mathcal{X}} & c'x \\ \text{subject to} & (\widetilde{\mathcal{W}}_{i,l}^{-1} z_i + s\widetilde{\mathcal{W}}_{i,l}^{-1} e_i)' x \leq b_i \text{ for all } i = 1, \ldots, K, l = 1, \ldots, k_{i,j}. \end{array}$$

The uniqueness of $x^*(s)$ is relatively straightforward to justify. The above representation of the RO and an application of Lemma D.4.2 suggest that, under the imposed conditions regarding satisfaction of the SCI condition, the solution $x^*(s)$ can be non-unique at only finitely many s values.

We now prove piecewise uniform continuity of the solution path. If at parameter value $\tilde{s}$ the solution $x^*(\tilde{s})$ is not unique, we call it a *non-unique point*. Between every two consecutive non-unique points $\tilde{s}_j < \tilde{s}_{j+1}$, $x^*(s)$ is unique hence is continuous in $(\tilde{s}_j, \tilde{s}_{j+1})$ due to Lemma D.4.1. To show that $x^*(s)$ is actually uniformly continuous, it is sufficient and necessary to demonstrate that, as s approaches some non-unique point $\tilde{s}$, left and right limits $\lim_{s\to\tilde{s}-} x^*(s), \lim_{s\to\tilde{s}+} x^*(s)$ exist. Without loss of generality, we focus on left limit. Toward this goal, we first derive a convenient formula of the optimal solution $x^*(s)$ for s in a sufficiently small neighborhood $[\tilde{s} - \delta, \tilde{s})$ of $\tilde{s}$. As

shown in the first step, for sufficiently small δ the reformulation $OPT(s)$ takes the form

$$\begin{aligned} \min_{x} \quad & c'x \\ \text{subject to} \quad & (\widetilde{\mathcal{W}}_{i,l}^{-1} z_i + s\widetilde{\mathcal{W}}_{i,l}^{-1} e_i)'x \leq b_i \text{ for all } i = 1, \ldots, K, l = 1, \ldots, k_i \\ & Wx \leq z \end{aligned}$$

for all $s \in [\tilde{s} - \delta, \tilde{s})$. For convenience, we rewrite the above parametric program in a more compact form

$$\begin{aligned} \min_{x} \quad & c'x \\ \text{subject to} \quad & (A + s\Delta)x \leq b \end{aligned} \tag{D.30}$$

where the matrix A contains all $\widetilde{\mathcal{W}}_{i,l}^{-1} z_i$'s and W as its rows, and the right hand side b has all the corresponding b_i's and z as its components, whereas the perturbation matrix Δ consists of all the $\widetilde{\mathcal{W}}_{i,l}^{-1} e_i$'s (and zero entries for the W part of A). Note again that $x^*(s)$ is the unique optimal solution of (D.30) for all $s \in [\tilde{s} - \delta, \tilde{s})$. The dual of (D.30) takes the form

$$\begin{aligned} \max_{y} \quad & b'y \\ \text{subject to} \quad & (A + s\Delta)'y = c \\ & y \leq 0 \end{aligned} \tag{D.31}$$

By the theory of simplex method, for the dual (D.31) there exists some basis $A_\beta + s\Delta_\beta$, where β is a subset of size d of $\{1, 2, \ldots, \sum_{i=1}^{K} k_i + L\}$ and A_β, Δ_β denote the submatrices formed by the corresponding rows of A, Δ, that gives rise to the optimal solution $y_\beta^*(s) = (A_\beta' + s\Delta_\beta')^{-1} c$ to (D.31) (other components of $y^*(s)$ are all zero). Moreover, the corresponding primal optimal solution to (D.30) is $x^*(s) = (A_\beta + s\Delta_\beta)^{-1} b_\beta$. By statement (ii) in Lemma 1 from Freund (1985), this optimal basis β for (D.31) can change for only finitely many times as the parameter s varies, therefore by choosing a small enough δ this basis β remain the same one for all $s \in [\tilde{s} - \delta, \tilde{s})$. That is, the unique optimal solution $x^*(s) = (A_\beta + s\Delta_\beta)^{-1} b_\beta$ for all $s \in [\tilde{s} - \delta, \tilde{s})$ and some basis β. Given this convenient formula, we now establish existence of the left limit. Case (i): $A_\beta + \tilde{s}\Delta_\beta$

is invertible. In this case the inverse $(A_\beta + s\Delta_\beta)^{-1}$ must be continuous in s at the non-unique point $\tilde{s}$, hence the left limit $\lim_{s\to\tilde{s}-} x^*(s) = (A_\beta + \tilde{s}\Delta_\beta)^{-1}b_\beta$. Case (ii): b_β is the zero vector. This case is trivial because $x^*(s)$ is also the zero vector hence the left limit exists and is the zero vector. Case (iii): $A_\beta + \tilde{s}\Delta_\beta$ is singular and b_β is a non-zero vector. Note that Assumption 5.5.3 implies that the solution path $\{x^*(s) : s \in \mathcal{S}\}$ is confined within a bounded region, and we shall use this key information to conclude this case. For convenience we reparametrize the solution path as $s_o = (s - \tilde{s} + \delta)^{-1}$ and $x_o^*(s_o) := x^*(s)$ for $s \in (\tilde{s} - \delta, \tilde{s})$. Letting $A_\beta^o = A_\beta + (\tilde{s} - \delta)\Delta_\beta$, we can express the reparametrization $x_o^*(s_o)$ as

$$x_o^*(s_o) = s_o\big((A_\beta^o)^{-1}\Delta_\beta + s_o I_d\big)^{-1}(A_\beta^o)^{-1}b_\beta \text{ for } s_o \in (\delta^{-1}, +\infty)$$

and our goal is to show $\lim_{s_o\to\delta^{-1}+} x_o^*(s_o)$ exists. The matrix $(A_\beta^o)^{-1}\Delta_\beta$ admits a Jordan decomposition $(A_\beta^o)^{-1}\Delta_\beta = P^{-1}JP$, where P is an invertible matrix with complex entries and J is the Jordan normal form with the diagonal structure

$$J = \begin{pmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_T \end{pmatrix}, \text{ with each diagonal block } J_t = \begin{pmatrix} \lambda_t & 1 & & & \\ & \lambda_t & 1 & & \\ & & \ddots & & \\ & & & \lambda_t & 1 \\ & & & & \lambda_t \end{pmatrix}$$

where each λ_t is an eigenvalue of $(A_\beta^o)^{-1}\Delta_\beta$. With the Jordan decomposition, the reparametrized solution path takes the form

$$x_o^*(s_o) = s_o P^{-1}\big(J + s_o I_d\big)^{-1}P(A_\beta^o)^{-1}b_\beta.$$

Moreover, the inverse $\big(J + s_o I_d\big)^{-1}$ can be expressed as

$$\big(J + s_o I_d\big)^{-1} = \begin{pmatrix} (J_1 + s_o I)^{-1} & & & \\ & (J_2 + s_o I)^{-1} & & \\ & & \ddots & \\ & & & (J_T + s_o I)^{-1} \end{pmatrix}$$

where each diagonal block, if $J_t \in \mathbb{R}^{d_t \times d_t}$, has the form

$$(J_t + s_o I)^{-1} = \begin{pmatrix} (\lambda_t + s_o)^{-1} & -(\lambda_t + s_o)^{-2} & \cdots & (-1)^{d_t - 1}(\lambda_t + s_o)^{-d_t} \\ & (\lambda_t + s_o)^{-1} & -(\lambda_t + s_o)^{-2} & \cdots & \\ & & \ddots & \ddots & \vdots \\ & & & (\lambda_t + s_o)^{-1} & -(\lambda_t + s_o)^{-2} \\ & & & & (\lambda_t + s_o)^{-1} \end{pmatrix}. \tag{D.32}$$

If we let $(P(A^o_\beta)^{-1} b_\beta)_t$ be the vector of length d_t consisting of the $(1 + \sum_{i=1}^{t-1} d_i)$-th to $(\sum_{i=1}^{t} d_i)$-th components of $P(A^o_\beta)^{-1} b_\beta$, then

$$x^*_o(s_o) = s_o P^{-1} \begin{pmatrix} (J_1 + s_o I)^{-1}(P(A^o_\beta)^{-1} b_\beta)_1 \\ (J_2 + s_o I)^{-1}(P(A^o_\beta)^{-1} b_\beta)_2 \\ \vdots \\ (J_T + s_o I)^{-1}(P(A^o_\beta)^{-1} b_\beta)_T \end{pmatrix}.$$

We argue that $(P(A^o_\beta)^{-1} b_\beta)_t$ must be the zero vector for all t such that $\lambda_t = -\delta^{-1}$. Note that, since $A_\beta + \tilde{s}\Delta_\beta$ is singular, some λ_t must be $-\delta^{-1}$. Consider a Jordan block J_t with $\lambda_t = -\delta^{-1}$. From the form (D.32) of the inverse, one can check that $\big\|(J_t + s_o I)^{-1} v\big\|_2 \to \infty$ as $s_o \to \delta^{-1}$ for any given non-zero vector v. However, the solution $x^*(s)$, hence each $(J_t + s_o I)^{-1}(P(A^o_\beta)^{-1} b_\beta)_t$, is confined to a bounded region, therefore $(P(A^o_\beta)^{-1} b_\beta)_t$ must be zero if $\lambda_t = -\delta^{-1}$. For those blocks with $\lambda_t \neq -\delta^{-1}$, the inverse $(J_t + s_o I)^{-1}$ is continuous in s_o at $s_o = \delta^{-1}$. Altogether, each block $(J_t + s_o I)^{-1}(P(A^o_\beta)^{-1} b_\beta)_t$ is either constantly zero or continuous at $s_o = \delta^{-1}$, therefore $x^*_o(s_o)$ has

right limit at $s_o = \delta^{-1}$. This proves the existence of left limit of $x^*(s)$ at $s = \tilde{s}$. □

Proof of Theorem 5.5.8. The continuity of the second-order cone constraints in x, s and its convexity in x are straightforward. We only focus on the uniqueness of $x^*(s)$. In case (i) uniqueness trivially follows from strict convexity of f. In case (ii), we want to show for each cone constraint that for any x_1, x_2 such that $\mu_i' x_1 + s \left\|\Sigma_i x_1\right\|_2 = \mu_i' x_2 + s \left\|\Sigma_i x_2\right\|_2 = b_i$ we must have $\mu_i' x_\theta + s \left\|\Sigma_i x_\theta\right\|_2 < b_i$ for any $x_\theta = \theta x_1 + (1-\theta) x_2$ where $\theta \in (0,1)$. First of all, there exists no $c \geq 0, c \neq 1$ such that $x_1 = c x_2$ or $x_2 = c x_1$ because otherwise $\mu_i' x_1 + s \left\|\Sigma_i x_1\right\|_2 = c(\mu_i' x_2 + s \left\|\Sigma_i x_2\right\|_2) = c b_i \neq b_i$. Second, if there exists some $c < 0$ such that $x_1 = c x_2$ or $x_2 = c x_1$, then $\mu_i' x_\theta + s \left\|\Sigma_i x_\theta\right\|_2$ is piecewise linear in θ and has two pieces with different slopes, therefore $\mu_i' x_\theta + s \left\|\Sigma_i x_\theta\right\|_2 < b_i$ for all $\theta \in (0,1)$. Finally, if x_1 and x_2 are not parallel, then it is easy to verify that $\left\|\Sigma_i x_\theta\right\|_2$ is strictly convex in θ by examining its second order derivative, therefore we have $\mu_i' x_\theta + s \left\|\Sigma_i x_\theta\right\|_2 < b_i$ again. Together with the SCI condition not being satisfied, we can use Lemma D.4.2 to conclude the uniqueness of $x^*(s)$ for all $s \in S$. Lemma D.4.1 then implies the desired conclusion. □

D.5 Finite Sample Performance Guarantees for Univariate Gaussian Validator

This section provides finite-sample errors regarding the performance guarantees presented in Theorem 5.4.4, focusing on two general classes of constraints: differentiable stochastic constraints (Section D.5.1) and linear chance constraints (Section D.5.2).

D.5.1 Differentiable Constraints

In order to derive finite-sample errors, we need stronger versions of Assumptions 5.4.2-5.4.4 and 5.4.9. Assumption 5.4.3 is replaced by boundedness of the fourth order moment:

Assumption D.5.1 $\mathfrak{m}_4 := \left(\mathbb{E}_F\left[\sup_{x \in \mathcal{X}} \left|h(x,\xi) - H(x)\right|^4\right]\right)^{\frac{1}{4}} < \infty$.

The L_2-continuity condition for the constraint function h in Assumption 5.4.4 is now strengthened to a differentiability condition:

Assumption D.5.2 *The random function $h(\cdot, \xi)$ is continuously differentiable on $\mathcal{X}$ for almost every $\xi \in \Xi$, and $\nabla H(x) = \mathbb{E}_F[\nabla h(x, \xi)]$. Assume $\overline{\rho} := \sup_{x \in \mathcal{X}} \rho(\mathrm{Cov}_F(\nabla h(x, \xi))) < \infty$ where $\rho(\cdot)$ denotes the largest eigenvalue of a matrix.*

Note that, in the presence of Assumption 5.4.3, Assumption D.5.2 implies Assumption 5.4.4 through the dominated convergence theorem. When the gradient has a square integrable envelope, i.e., $\mathbb{E}_F[\sup_{x \in \mathcal{X}} \|\nabla h(x, \xi)\|^2] < \infty$, and the decision space $\mathcal{X}$ is compact, Assumption D.5.2 also implies Assumption 5.4.2.

We then assume uniqueness of the optimal parameter, and local differentiability of the solution path and the expectation constraint:

Assumption D.5.3 (Unique optimal parameter and local differentiability) *There exists a unique optimal parameter, i.e., $S^* = \{s^*\}$, and $H(x_S^*) = \gamma$ at the optimal solution $x_S^* = x^*(s^*)$. Moreover, $H(x)$ is continuously differentiable in a neighborhood of x_S^*, and the parameter-to-solution mapping $x^*(s)$ is continuously differentiable in a neighborhood of s^*. There exists a $\delta > 0$ such that for all $s \in [s^* - \delta, s^* + \delta] \subseteq S$ it holds $\frac{1}{2} \leq \nabla H(x^*(s))' \nabla x^*(s) / \nabla H(x_S^*)' \nabla x^*(s^*) \leq 2$ and $\|\nabla x^*(s)\|_2 / \|\nabla x^*(s^*)\|_2 \leq 2$, and that for all $s \leq s^* - \delta$ it holds $H(x^*(s)) \leq H(x^*(s^* - \delta))$.*

We have the following finite-sample performance bounds for Algorithm 11:

Theorem D.5.1 (Finite-sample feasibility guarantee of univariate Gaussian validator) *Suppose Assumptions 5.4.1, 5.4.5-5.4.8 and D.5.1-D.5.3 hold, and $\{s_1, \ldots, s_p\} \cap \{\tilde{s}_1, \ldots, \tilde{s}_{M-1}\} = \emptyset$. Recall the mesh size $\epsilon_S = \sup_{s \in S} \inf_{1 \leq j \leq p} |s - s_j|$. Denote by $c^* := \nabla H(x_S^*)' \nabla x^*(s^*) / \|\nabla x^*(s^*)\|_2$, and by C some universal constant. For any $t > 0$ such that*

$$2\epsilon_S < \mathrm{err}(p, n_2, t) := \frac{4(1 + z_{1-\beta})\mathfrak{m}_4}{c^* \|\nabla x^*(s^*)\|_2} \sqrt{\frac{t \log p}{n_2}} < \frac{\delta}{2}$$

it holds for the parameter $\hat{s}^*$ output by Algorithm 11 that

$$P_{\boldsymbol{\xi}_{1:n_2}}(|\hat{s}^* - s^*| > 2\mathrm{err}(p, n_2, t)) \leq \frac{C}{t}.$$

If

$$2\epsilon_S < \frac{4(1+z_{1-\beta})\mathfrak{m}_4}{c^* \left\|\nabla x^*(s^*)\right\|_2} \cdot \frac{(\log p)^{1/4}}{n_2^{3/8}} < \frac{\delta}{2}$$

we have

$$P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.1)}) \geq 1-\beta-C(1+z_{1-\beta})^2\big(\frac{\mathfrak{m}_4}{\sigma(x_S^*)}\big)^3\big(1+\frac{\sqrt{\bar{\rho}}}{c^*}\big)^{\frac{2}{3}}\big(\frac{(\log p)^2}{n_2}\big)^{\frac{1}{4}}. \quad \text{(D.33)}$$

Proof of Theorem D.5.1. First we present a lemma concerning moment inequalities for the maximum deviation of sample means:

Lemma D.5.2 *Let $\mathcal{G}$ be function class of finite cardinality, and $G(\xi) := \max_{g\in\mathcal{G}}|g(\xi)|$ be the envelope function. Suppose $\xi_1, \ldots, \xi_n$ are i.i.d. observations from a common distribution F, then for any $k \geq 1$ we have*

$$\sqrt{n}\left(\mathbb{E}\big[\max_{g\in\mathcal{G}}\big|\frac{1}{n}\sum_{i=1}^{n} g(\xi_i) - \mathbb{E}_F[g(\xi)]\big|^k\big]\right)^{1/k} \leq C\sqrt{1+\log|\mathcal{G}|}\left(\mathbb{E}_F[(G(\xi))^{\tilde{k}}]\right)^{1/\tilde{k}}$$

where $\tilde{k} = \max(2,k)$, the constant C only depends on k, and $|\mathcal{G}|$ denotes the cardinality of $\mathcal{G}$.

Proof of Lemma D.5.2. This is a direct consequence of Theorem 2.14.1 from Van der Vaart and Wellner (1996). To apply that theorem, note that the covering number of the function class $\mathcal{G}$ is at most $|\mathcal{G}|$, hence its entropy integral is at most $\sqrt{1+\log|\mathcal{G}|}$. □

We use Lemma D.5.2 to derive tail bounds for various maximum deviations. Denote by $H_j = H(x^*(s_j))$, and $\sigma_j^2 = \sigma^2(x^*(s_j))$ for convenience. Applying Lemma D.5.2 to $\{h(x^*(s_j),\cdot) - H_j : j = 1,\ldots,p\}$ with $k=4$ gives

$$\begin{aligned} n_2^2\mathbb{E}_{\boldsymbol{\xi}_{1:n_2}}\big[\big(\max_j\big|\hat{H}_j - H_j\big|\big)^4\big] &\leq C(\log p)^2\mathbb{E}_F[(\max_j|h(x^*(s_j),\xi) - H_j|)^4] \\ &\leq C(\log p)^2\mathfrak{m}_4^4 \end{aligned}$$

where C is a universal constant (because k is fixed at 4) and $\mathbb{E}_{\boldsymbol{\xi}_{1:n_2}}$ denotes the expectation conditioned on Phase one data and with respect to Phase two data. Similarly applying the lemma to

the squared class $\{(h(x^*(s_j),\cdot)-H_j)^2-\sigma_j^2 : j=1,\ldots,p\}$ with $k=2$ gives

$$\begin{aligned} & n_2\mathbb{E}_{\boldsymbol{\xi}_{1:n_2}}\big[\big(\max_j\big|\frac{1}{n_2}\sum_{i=1}^{n_2}(h(x^*(s_j),\xi_i)-H_j)^2-\sigma_j^2\big|\big)^2\big] \\ \le\ & C\log p\mathbb{E}_F[(\max_j|(h(x^*(s_j),\xi)-H_j)^2-\sigma_j^2|)^2] \\ \le\ & C\log p\mathfrak{m}_4^4. \end{aligned}$$

By Markov's inequality, for any $t_1>0$ we have

$$\max_{j=1,\ldots,p}\left|\hat{H}_j-H_j\right|\le\frac{\mathfrak{m}_4 t_1}{\sqrt{n_2}}$$

with probability at least $1-C(\log p)^2/t_1^4$ and

$$\max_{j=1,\ldots,p}\left|\hat{\sigma}_j^2-\sigma_j^2\right|\le\max_j\big|\frac{1}{n_2}\sum_{i=1}^{n_2}(h(x^*(s_j),\xi_i)-H_j)^2-\sigma_j^2\big|+\max_{j=1,\ldots,p}(\hat{H}_j-H_j)^2\le\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}}+\frac{\mathfrak{m}_4^2 t_1^2}{n_2} \quad \text{(D.34)}$$

with probability at least

$$1-\frac{C(\log p)^2}{t_1^4}-\frac{C(\log p)}{t_1^2}.$$

Note that, when the upper bound (D.34) holds, $\max_j\hat{\sigma}_j^2\le\max_j\sigma_j^2+\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}}+\frac{\mathfrak{m}_4^2 t_1^2}{n_2}\le\mathfrak{m}_4^2+\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}}+\frac{\mathfrak{m}_4^2 t_1^2}{n_2}$. Therefore for any $t_1>0$

$$\begin{aligned} \max_{j=1,\ldots,p}\left|\hat{H}_j-z_{1-\beta}\frac{\hat{\sigma}_j}{\sqrt{n_2}}-H_j\right| &\le \frac{\mathfrak{m}_4 t_1}{\sqrt{n_2}}+z_{1-\beta}\sqrt{\frac{\mathfrak{m}_4^2}{n_2}+\frac{\mathfrak{m}_4^2 t_1}{n_2^{3/2}}+\frac{\mathfrak{m}_4^2 t_1^2}{n_2^2}} \\ &\le (1+z_{1-\beta})\mathfrak{m}_4\frac{1+t_1}{\sqrt{n_2}}\le 2(1+z_{1-\beta})\mathfrak{m}_4\frac{t_1}{\sqrt{n_2}} \quad \text{(D.35)} \end{aligned}$$

for all $t_1\ge 1$ with probability at least

$$1-\frac{C(\log p)^2}{t_1^4}-\frac{C(\log p)}{t_1^2}.$$

For every constant $\epsilon<\delta$, the solution path $x^*(s)$ is differentiable for $s\in[s^*-\epsilon,s^*+\epsilon]$. Therefore

for any s_j such that $|s_j - s^*| \leq \epsilon$, by differentiability we have

$$\begin{aligned} |[h(x^*(s_j),\xi) - H_j] - [h(x_S^*,\xi) - H(x_S^*))]| &= \left| \int_{s^*}^{s_j} (\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))' \nabla x^*(s) ds \right| \\ &\leq \int_{s^*-\epsilon}^{s^*+\epsilon} \left| (\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))' \nabla x^*(s) \right| ds. \end{aligned}$$

The right hand side of the above inequality serves as an envelope function of the function class $\{[h(x^*(s_j),\xi) - H_j] - [h(x_S^*,\xi) - H(x_S^*))] : |s_j - s^*| \leq \epsilon\}$. From Assumption D.5.2 we have that $\mathbb{E}_F[|(\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))' \nabla x^*(s)|^2] \leq \bar{\rho} \|\nabla x^*(s)\|_2^2$ for all s, therefore by Jensen's inequality (or Minkowski's integral inequality)

$$\begin{aligned} & \mathbb{E}_F\big[\big(\int_{s^*-\epsilon}^{s^*+\epsilon} \left| (\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))' \nabla x^*(s) \right| ds \big)^2\big] \\ \leq & \big(\int_{s^*-\epsilon}^{s^*+\epsilon} \sqrt{\mathbb{E}_F[|(\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))' \nabla x^*(s)|^2]} ds \big)^2 \\ \leq & \bar{\rho} \big(\int_{s^*-\epsilon}^{s^*+\epsilon} \|\nabla x^*(s)\|_2 \, ds \big)^2 \\ \leq & 16 \bar{\rho} \|\nabla x^*(s^*)\|_2^2 \, \epsilon^2 \end{aligned}$$

an upper bound for the second moment of the envelope. Now applying Lemma D.5.2 with $k = 2$ to $\{[h(x^*(s_j),\xi) - H_j] - [h(x_S^*,\xi) - H(x_S^*))] : |s_j - s^*| \leq \epsilon)\}$, and noting that the cardinality does not exceed p, we have

$$n_2 \mathbb{E}_{\boldsymbol{\xi}_{1:n_2}} \big[\max_{j:|s_j - s^*| \leq \epsilon} \left| \hat{H}_j - H_j - (\hat{H}(x_S^*) - H(x_S^*)) \right|^2 \big] \leq C(\log p) \bar{\rho} \|\nabla x^*(s^*)\|_2^2 \, \epsilon^2$$

which implies through Markov's inequality that for every $t_2 > 0$

$$\max_{j:|s_j - s^*| \leq \epsilon} \left| \hat{H}_j - H_j - (\hat{H}(x_S^*) - H(x_S^*)) \right| \leq \frac{t_2}{\sqrt{n_2}} \tag{D.36}$$

with probability at least

$$1 - \frac{C(\log p) \bar{\rho} \|\nabla x^*(s^*)\|_2^2 \, \epsilon^2}{t_2^2}.$$

Deviation inequalities (D.34), (D.35) and (D.36) are the key elements for establishing finite sample error bounds. Lastly, we also need a bound characterizing the modulus of continuity of the variance $\sigma^2(x^*(s))$. For every s_j such that $|s_j - s^*| \le \epsilon$

$$\begin{aligned}
|\sigma_j^2 - \sigma^2(x_S^*)| &= |\mathbb{E}_F[(h(x^*(s_j),\xi) - H_j)^2] - \mathbb{E}_F[(h(x_S^*,\xi) - H(x_S^*))^2]| \\
&= \Big|\mathbb{E}_F\big[\int_{s^*}^{s_j} 2(h(x^*(s),\xi) - H(x^*(s)))(\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))'\nabla x^*(s) ds\big]\Big| \\
&\le \mathbb{E}_F\big[\int_{s^*}^{s_j} 2|h(x^*(s),\xi) - H(x^*(s))||(\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))'\nabla x^*(s)| ds\big] \\
&\le \mathbb{E}_F\big[\int_{s^*-\epsilon}^{s^*+\epsilon} 2|h(x^*(s),\xi) - H(x^*(s))||(\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))'\nabla x^*(s)| ds\big] \\
&= \int_{s^*-\epsilon}^{s^*+\epsilon} 2\mathbb{E}_F\big[|h(x^*(s),\xi) - H(x^*(s))||(\nabla h(x^*(s),\xi) - \nabla H(x^*(s)))'\nabla x^*(s)|\big] ds \\
&\quad \text{by Fubini's theorem} \\
&\le \int_{s^*-\epsilon}^{s^*+\epsilon} 2\sigma(x^*(s))\sqrt{\bar\rho}\,\|\nabla x^*(s)\|_2\, ds \ \text{ by Cauchy Schwartz inequality} \\
&\le 8\mathfrak{m}_4\sqrt{\bar\rho}\,\|\nabla x^*(s^*)\|_2\,\epsilon.
\end{aligned}$$

That is, for all $\epsilon < \delta$

$$\max_{j:|s_j - s^*|\le\epsilon} |\sigma_j^2 - \sigma^2(x_S^*)| \le 8\mathfrak{m}_4\sqrt{\bar\rho}\,\|\nabla x^*(s^*)\|_2\,\epsilon. \tag{D.37}$$

We first show the deviation inequality for $\hat{s}^*$. If (D.35) happens, and t_1 is such that

$$2\epsilon_S < \epsilon(t_1, n_2) := \frac{4(1+z_{1-\beta})\mathfrak{m}_4}{c^*\,\|\nabla x^*(s^*)\|_2} \cdot \frac{t_1}{\sqrt{n_2}} < \frac{\delta}{2} \tag{D.38}$$

we want to show that $|\hat{s}^* - s^*| \le 2\epsilon(t_1, n_2)$. By Assumption D.5.3, for any $s \in (s^*, s^* + \delta]$ the constraint value $H(x^*(s)) \ge \gamma + \frac{s-s^*}{2}\nabla H(x^*(s^*))'\nabla x^*(s^*) = \gamma + \frac{s-s^*}{2}c^*\,\|\nabla x^*(s^*)\|_2$, and similarly $H(x^*(s)) \le \gamma + \frac{s-s^*}{2}c^*\,\|\nabla x^*(s^*)\|_2$ for all $s \in [s^* - \delta, s^*)$. Therefore $H_j > \gamma + 2(1+z_{1-\beta})\mathfrak{m}_4\frac{t_1}{\sqrt{n_2}}$ for all $s_j \in (s^* + \epsilon(t_1,n_2), s^* + \delta]$ and $H_j < \gamma - 2(1+z_{1-\beta})\mathfrak{m}_4\frac{t_1}{\sqrt{n_2}}$ for all $s_j \in [s^* - \delta, s^* - \epsilon(t_1,n_2))$. Under the condition that $2\epsilon_S < \epsilon(t_1,n_2)$ there must be some $\bar{j}$ for which $s_{\bar{j}} \in (s^* + \epsilon(t_1,n_2), s^* + 2\epsilon(t_1,n_2)) \subset (s^* + \epsilon(t_1,n_2), s^* + \delta]$ and hence $\hat{H}_{\bar{j}} - z_{1-\beta}\frac{\hat{\sigma}_{\bar{j}}}{\sqrt{n_2}} > \gamma$ on one hand. On the other hand the solution path has a derivative $\nabla x^*(s)$ that is non-zero in $[s^* - \delta, s^* + \delta]$ hence the parameter-to-objective

mapping $v(s)$ strictly increases in s in the same interval. Therefore the picked parameter $\hat{s}^* \leq s_{\bar{j}}$. Similarly, there exists some $\underline{j}$ such that $s_{\underline{j}} \in (s^* - 2\epsilon(t_1, n_2), s^* - \epsilon(t_1, n_2)) \subset (s^* - \delta, s - \epsilon(t_1, n_2)]$ and $H_{\underline{j}} < \gamma - 2(1 + z_{1-\beta})\mathfrak{m}_4 \frac{t_1}{\sqrt{n_2}}$. Since $H(x^*(s^* - \delta)) \geq H(x^*(s))$ for all $s \leq s^* - \delta$, we have for all $s_j \leq s_{\underline{j}}$ that $H_j \leq H_{\underline{j}}$ and $\hat{H}_j - z_{1-\beta}\frac{\hat{\sigma}_j}{\sqrt{n_2}} < \gamma$, therefore $\hat{s}^* \leq s_{\underline{j}}$ is impossible. That is, it must be the case that $\hat{s}^* \in (s_{\underline{j}}, s_{\bar{j}}] \subset [s^* - 2\epsilon(t_1, n_2), s^* + 2\epsilon(t_1, n_2)]$. This gives the deviation inequality

$$P_{\boldsymbol{\xi}_{1:n_2}}(|\hat{s}^* - s^*| > 2\epsilon(t_1, n_2)) \leq C\big(\frac{(\log p)^2}{t_1^4} + \frac{\log p}{t_1^2}\big) \leq \frac{C \log p}{t_1^2}$$

provided that (D.38) holds. Since the above bound becomes trivial when $(\log p)/t_1^2 \geq 1$, hence we can assume $(\log p)/t_1^2 < 1$ without loss of generality (and enlarge the universal constant C if necessary) to get

$$P_{\boldsymbol{\xi}_{1:n_2}}(|\hat{s}^* - s^*| > 2\epsilon(t_1, n_2)) \leq \frac{C \log p}{t_1^2}. \tag{D.39}$$

Now we derive the finite sample error for the feasibility confidence level. Using the same notation

$\epsilon(t_1, n_2)$, we write

$$
\begin{aligned}
& P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \\
\geq\ & P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma, |\hat{s}^* - s^*| \leq 2\epsilon(t_1, n_2)) \\
\geq\ & P_{\boldsymbol{\xi}_{1:n_2}}\Big(\hat{H}(x^*(\hat{s}^*)) - H(x^*(\hat{s}^*)) - \frac{z_{1-\beta}\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}} \leq \hat{H}(x^*(\hat{s}^*)) - \frac{z_{1-\beta}\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}} - \gamma, \\
& \qquad \text{and } |\hat{s}^* - s^*| \leq 2\epsilon(t_1, n_2)\Big) \\
\geq\ & P_{\boldsymbol{\xi}_{1:n_2}}(\hat{H}(x^*(\hat{s}^*)) - H(x^*(\hat{s}^*)) - \frac{z_{1-\beta}\hat{\sigma}(x^*(\hat{s}^*))}{\sqrt{n_2}} \leq 0, |\hat{s}^* - s^*| \leq 2\epsilon(t_1, n_2)) \\
=\ & P_{\boldsymbol{\xi}_{1:n_2}}(\hat{H}(x_S^*) - H(x_S^*) - \frac{z_{1-\beta}\sigma(x_S^*)}{\sqrt{n_2}} + \Delta_H + \Delta_\sigma \leq 0, |\hat{s}^* - s^*| \leq 2\epsilon(t_1, n_2)) \\
& \qquad \text{where } \Delta_H = (\hat{H}(x^*(\hat{s}^*)) - H(x^*(\hat{s}^*))) - (\hat{H}(x_S^*) - H(x_S^*)) \text{ and} \\
& \qquad \Delta_\sigma = (z_{1-\beta}/\sqrt{n_2})(\sigma(x_S^*) - \hat{\sigma}(x^*(\hat{s}^*))) \\
\geq\ & P_{\boldsymbol{\xi}_{1:n_2}}\big(\hat{H}(x_S^*) - H(x_S^*) - \frac{z_{1-\beta}\sigma(x_S^*)}{\sqrt{n_2}} + \max_{j:|s_j - s^*| \leq 2\epsilon(t_1,n_2)} \big|\hat{H}_j - H_j - (\hat{H}(x_S^*) - H(x_S^*))\big| + \\
& \qquad \frac{z_{1-\beta}}{\sqrt{n_2}} \max_{j:|s_j - s^*| \leq 2\epsilon(t_1,n_2)} |\sigma(x_S^*) - \hat{\sigma}_j| \leq 0,\ |\hat{s}^* - s^*| \leq 2\epsilon(t_1, n_2)\big) \\
\geq\ & P_{\boldsymbol{\xi}_{1:n_2}}\big(\frac{\sqrt{n_2}(\hat{H}(x_S^*) - H(x_S^*))}{\sigma(x_S^*)} + \frac{\sqrt{n_2}}{\sigma(x_S^*)} \max_{j:|s_j - s^*| \leq 2\epsilon(t_1,n_2)} \big|\hat{H}_j - H_j - (\hat{H}(x_S^*) - H(x_S^*))\big| + \\
& \qquad \frac{z_{1-\beta}}{\sigma(x_S^*)} \max_{j:|s_j - s^*| \leq 2\epsilon(t_1,n_2)} |\sigma(x_S^*) - \hat{\sigma}_j| \leq z_{1-\beta},\ |\hat{s}^* - s^*| \leq 2\epsilon(t_1, n_2)\big).
\end{aligned}
$$

It follows from (D.34) and (D.37) that

$$
\max_{j:|s_j - s^*| \leq 2\epsilon(t_1,n_2)} |\sigma^2(x_S^*) - \hat{\sigma}_j^2| \leq \frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1, n_2)
$$

with probability at least $1 - C(\log p)/t_1^2 - C(\log p)^2/t_1^4$. If $\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1, n_2) \leq \sigma^2(x_S^*)/4$, it follows from mean value theorem that with at least the same probability

$$
\max_{j:|s_j - s^*| \leq 2\epsilon(t_1,n_2)} |\sigma(x_S^*) - \hat{\sigma}_j| \leq \frac{1}{\sigma(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1, n_2)\big). \tag{D.40}
$$

For now we assume $\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1, n_2) \leq \sigma^2(x_S^*)/4$ holds so that the bound

(D.40) is valid. Later on we shall show that this is without loss of generality. We proceed as

$$
\begin{aligned}
&P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \\
\geq &P_{\boldsymbol{\xi}_{1:n_2}}\Big(\frac{\sqrt{n_2}(\hat{H}(x_S^*) - H(x_S^*))}{\sigma(x_S^*)} + \frac{t_2}{\sigma(x_S^*)} + \frac{z_{1-\beta}}{\sigma^2(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2)\big) \\
&\qquad \leq z_{1-\beta}\Big) \\
&- P_{\boldsymbol{\xi}_{1:n_2}}\big(\max_{j:|s_j - s^*|\leq 2\epsilon(t_1,n_2)} \big|\hat{H}_j - H_j - (\hat{H}(x_S^*) - H(x_S^*))\big| > \frac{t_2}{\sqrt{n_2}}\big) - P_{\boldsymbol{\xi}_{1:n_2}}\big(|\hat{s}^* - s^*| > 2\epsilon(t_1,n_2)\big) \\
&- P_{\boldsymbol{\xi}_{1:n_2}}\big(\max_{j:|s_j - s^*|\leq 2\epsilon(t_1,n_2)} |\sigma(x_S^*) - \hat{\sigma}_j| > \frac{1}{\sigma(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2)\big)\big) \\
\geq &P_{\boldsymbol{\xi}_{1:n_2}}\Big(\frac{\sqrt{n_2}(\hat{H}(x_S^*) - H(x_S^*))}{\sigma(x_S^*)} \leq z_{1-\beta} - \frac{t_2}{\sigma(x_S^*)} \\
&\qquad - \frac{z_{1-\beta}}{\sigma^2(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2)\big)\Big) \\
&- \frac{C(\log p)\bar{\rho}\,\|\nabla x^*(s^*)\|_2^2\,(\epsilon(t_1,n_2))^2}{t_2^2} - \frac{C\log p}{t_1^2}
\end{aligned}
$$

by (D.36), (D.39) and (D.40).

To deal with the first probability term, we recall the Berry-Esseen theorem. There exists some universal constant C such that

$$
\sup_{t\in\mathbb{R}} |P_{\boldsymbol{\xi}_{1:n_2}}\big(\frac{\sqrt{n_2}(\hat{H}(x_S^*) - H(x_S^*))}{\sigma(x_S^*)} \leq t\big) - \Phi(t)| \leq \frac{C\mathbb{E}_F[|h(x_S^*,\xi) - H(x_S^*)|^3]}{\sigma^3(x_S^*)\sqrt{n_2}}
$$

where Φ is the cumulative distribution function for the standard normal. Noting that $\mathbb{E}_F[|h(x_S^*,\xi) -$

$H(x_S^*)|^3] \le \mathfrak{m}_4^3$ and that Φ has a bounded derivative, we further bound the confidence level as

$$
\begin{aligned}
& P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \ge \gamma) \\
\ge\ & \Phi\big(z_{1-\beta} - \frac{t_2}{\sigma(x_S^*)} - \frac{z_{1-\beta}}{\sigma^2(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2)\big)\big) - \frac{C\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}} \\
& - \frac{C(\log p)\bar{\rho}\,\|\nabla x^*(s^*)\|_2^2\,(\epsilon(t_1,n_2))^2}{t_2^2} - \frac{C\log p}{t_1^2} \\
\ge\ & 1 - \beta - C\big(\frac{t_2}{\sigma(x_S^*)} + \frac{z_{1-\beta}}{\sigma^2(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2)\big)\big) - \frac{C\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}} \\
& - \frac{C(\log p)\bar{\rho}\,\|\nabla x^*(s^*)\|_2^2\,(\epsilon(t_1,n_2))^2}{t_2^2} - \frac{C\log p}{t_1^2}.
\end{aligned}
$$

Arranging terms gives

$$
\begin{aligned}
& 1 - \beta - P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \ge \gamma) \\
\le\ & C\Big(\frac{t_2}{\sigma(x_S^*)} + \frac{z_{1-\beta}}{\sigma^2(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + \mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2)\big) + \frac{\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}} \\
& + \frac{(\log p)\bar{\rho}\,\|\nabla x^*(s^*)\|_2^2\,(\epsilon(t_1,n_2))^2}{t_2^2} + \frac{\log p}{t_1^2}\Big) \\
\le\ & C\Big(\frac{z_{1-\beta}}{\sigma^2(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + \mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2)\big) + \frac{\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}} \\
& + \frac{\big((\log p)\bar{\rho}\big)^{1/3}\,\|\nabla x^*(s^*)\|_2^{2/3}\,(\epsilon(t_1,n_2))^{2/3}}{\sigma^{2/3}(x_S^*)} + \frac{\log p}{t_1^2}\Big) \qquad \text{(D.41)} \\
& \text{by minimizing the bound over } t_2 \\
\le\ & C\Big(\frac{z_{1-\beta}}{\sigma^2(x_S^*)}\big(\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2)\big) + \frac{\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}} \\
& + \frac{\big((\log p)\bar{\rho}\big)^{1/3}\,\|\nabla x^*(s^*)\|_2^{2/3}\,(\epsilon(t_1,n_2))^{2/3}}{\sigma^{2/3}(x_S^*)} + \frac{\log p}{t_1^2}\Big)
\end{aligned}
$$

where in the last inequality we leave out the terms $\frac{\mathfrak{m}_4^2 t_1^2}{n_2}$ because when $\frac{t_1}{\sqrt{n_2}} \le 1$ it holds that $\frac{t_1^2}{n_2} \le \frac{t_1}{\sqrt{n_2}}$ hence the former can be absorbed into the latter. Previously we assume that $\frac{\mathfrak{m}_4^2 t_1}{\sqrt{n_2}} + \frac{\mathfrak{m}_4^2 t_1^2}{n_2} + 16\mathfrak{m}_4\sqrt{\bar{\rho}}\,\|\nabla x^*(s^*)\|_2\,\epsilon(t_1,n_2) \le \sigma^2(x_S^*)/4$. This is without loss of generality, because otherwise the first error term in (D.41) is of constant order which makes the upper bound trivial. Now expanding

the $\epsilon(t_1, n_2)$ we further bound the error as follows

$$
\begin{aligned}
& 1-\beta-P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*))\geq\gamma) \\
\leq\ & C\Big(\frac{z_{1-\beta}\mathfrak{m}_4^2}{\sigma^2(x_S^*)}\big(\frac{t_1}{\sqrt{n_2}}+\Big[\frac{\bar{\rho}(1+z_{1-\beta})^2}{c^{*2}n_2}\Big]^{1/2}t_1\big)+\frac{\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}} \\
& \quad +\Big[\frac{(\log p)\bar{\rho}(1+z_{1-\beta})^2\mathfrak{m}_4^2}{\sigma^2(x_S^*)c^{*2}n_2}\Big]^{1/3}t_1^{2/3}+\frac{\log p}{t_1^2}\Big) \\
\leq\ & C\Big(\frac{z_{1-\beta}\mathfrak{m}_4^2}{\sigma^2(x_S^*)}\big(1+\frac{\sqrt{\bar{\rho}}(1+z_{1-\beta})}{c^*}\big)\frac{t_1}{\sqrt{n_2}}+(\log p)^{1/3}\Big[\frac{\mathfrak{m}_4}{\sigma(x_S^*)}\big(1+\frac{\sqrt{\bar{\rho}}(1+z_{1-\beta})}{c^*}\big)\frac{t_1}{\sqrt{n_2}}\Big]^{2/3} \\
& \quad +\frac{\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}}+\frac{\log p}{t_1^2}\Big) \\
\leq\ & C\Big(\big(\frac{z_{1-\beta}\mathfrak{m}_4}{\sigma(x_S^*)}+(\log p)^{1/3}\big)\Big[\frac{\mathfrak{m}_4}{\sigma(x_S^*)}\big(1+\frac{\sqrt{\bar{\rho}}(1+z_{1-\beta})}{c^*}\big)\frac{t_1}{\sqrt{n_2}}\Big]^{2/3}+\frac{\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}}+\frac{\log p}{t_1^2}\Big) \\
& \quad \text{since it can be assumed } \frac{\mathfrak{m}_4}{\sigma(x_S^*)}\big(1+\frac{\sqrt{\bar{\rho}}(1+z_{1-\beta})}{c^*}\big)\frac{t_1}{\sqrt{n_2}}\leq 1 \\
\leq\ & C\Big(\frac{(1+z_{1-\beta})^{5/3}(\log p)^{1/3}\mathfrak{m}_4^3}{\sigma^3(x_S^*)n_2^{1/3}}\big(1+\frac{\sqrt{\bar{\rho}}}{c^*}\big)^{2/3}t_1^{2/3}+\frac{\log p}{t_1^2}+\frac{\mathfrak{m}_4^3}{\sigma^3(x_S^*)\sqrt{n_2}}\Big) \quad \text{since } \mathfrak{m}_4\geq\sigma(x_S^*) \\
\leq\ & C\Big(\frac{(1+z_{1-\beta})^{5/3}(\log p)^{1/3}\mathfrak{m}_4^3}{\sigma^3(x_S^*)n_2^{1/3}}\big(1+\frac{\sqrt{\bar{\rho}}}{c^*}\big)^{2/3}t_1^{2/3}+\frac{\log p}{t_1^2}\Big) \qquad \text{(D.42)}
\end{aligned}
$$

where in the last inequality we drop the last term since it's dominated by the first when $t_1 \geq 1$. Note that (D.42) holds only under the condition (D.38). It is straightforward to see that the bound (D.42) is minimized at

$$t_1^* := \frac{(\log p)^{1/4}\sigma^{9/8}(x_S^*)n_2^{1/8}}{(1+z_{1-\beta})^{5/8}\mathfrak{m}_4^{9/8}(1+\sqrt{\bar{\rho}}/c^*)^{1/4}}$$

by equating the two error terms. Consider $\tilde{t}_1 := (\log p)^{1/4}n_2^{1/8}$. It is clear that $\tilde{t}_1 = t_1^*(1+z_{1-\beta})^{5/8}\big(\frac{\mathfrak{m}_4}{\sigma(x_S^*)}\big)^{9/8}(1+\sqrt{\bar{\rho}}/c^*)^{1/4} > t_1^*$, and hence the first term dominates at $t_1 = \tilde{t}_1$. Therefore when (D.38) is satisfied at $t_1 = \tilde{t}_1$, we have

$$1-\beta-P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*))\geq\gamma) \leq C(1+z_{1-\beta})^{\frac{5}{3}}\big(\frac{\mathfrak{m}_4}{\sigma(x_S^*)}\big)^3\big(1+\frac{\sqrt{\bar{\rho}}}{c^*}\big)^{\frac{2}{3}}\big(\frac{(\log p)^2}{n_2}\big)^{\frac{1}{4}}.$$

The desired bound is obtained by replacing $\frac{5}{3}$ with 2 as the exponent of $1+z_{1-\beta}$. □

D.5.2 Linear Chance Constraints

Consider linear chance constraints in the form of $\mathbb{P}_F(a_k'x \le b_k \text{ for } k = 1,\dots,K) \ge 1-\alpha$. We assume the following isotropy condition:

Assumption D.5.4 (Isotropy) *There exist constants D_2, D_3 such that for all unit vector $\nu \in \mathbb{R}^d$ and all $a_k, 1 \le k \le K$, the random variable $a_k'\nu$ has a sub-Gaussian norm at most D_2, i.e., $\mathbb{E}\big[\exp\big(\big(\frac{a_k'x}{D_2}\big)^2\big)\big] \le 2$, and has a density bounded above by D_3. Each b_k is a non-zero constant.*

This assumption stipulates that each a_k has variability of constant order in all directions, and it trivially holds when each a_k is standard Gaussian.

We have the following finite-sample performance bounds for linear chance constraints:

Theorem D.5.3 (Finite-sample CCP feasibility guarantee) *Consider* (5.12) *with a linear chance constraint $h(x,\xi) = \mathbf{1}(a_k'x \le b_k \text{ for } k = 1,\dots,K)$ and $0 < \alpha < \frac{1}{2}$. Suppose Assumptions 5.4.1, 5.4.6-5.4.8, and D.5.3-D.5.4 hold, and $\{s_1,\dots,s_p\} \cap \{\tilde{s}_1,\dots,\tilde{s}_{M-1}\} = \emptyset$. Recall the notations ϵ_S and c^* from Theorem D.5.1. For any $t > 0$ such that*

$$2\epsilon_S < \mathrm{err}(p,n_2,t) := \frac{6(1+z_{1-\beta})}{c^*\,\|\nabla x^*(s^*)\|_2}\Big(\sqrt{\frac{\alpha\log(4pt)}{n_2}} + \frac{\log(4pt)}{n_2}\Big) < \frac{\delta}{2}$$

it holds for the parameter $\hat{s}^*$ output by Algorithm 11 that

$$P_{\boldsymbol{\xi}_{1:n_2}}\big(\,|\hat{s}^* - s^*| > 2\mathrm{err}(p,n_2,t)\big) \le \frac{1}{t}.$$

If

$$2\epsilon_S < \frac{6(1+z_{1-\beta})}{c^*\,\|\nabla x^*(s^*)\|_2}\Big(\sqrt{\frac{\alpha\log(pn_2)}{n_2}} + \frac{\log(pn_2)}{n_2}\Big) < \frac{\delta}{2}$$

we have

$$\begin{aligned} &P_{\boldsymbol{\xi}_{1:n_2}}(x^*(\hat{s}^*) \text{ is feasible for (5.12)}) \\ \ge& 1-\beta - C(1+z_{1-\beta})^2\big(1+\sqrt{\tilde{C}K}\big(\log(\max\{3,\tfrac{n_2}{\tilde{C}}\})\big)^{\frac{1}{4}}\big)\Big(\frac{(\log(pn_2))^3}{\alpha n_2}\Big)^{\frac{1}{4}} \end{aligned} \tag{D.43}$$

where C is a universal constant and

$$\tilde{C} = \frac{D_2^2 D_3 \sqrt{\log(2K/\alpha)}}{c^* \min_{1\le k\le K}|b_k|}.$$

To get a sense of the effect of the dimension d on the finite-sample error (D.43), suppose that $D_2, D_3, K, \{b_k, k = 1, \ldots, K\}$ are all numbers of constant order and we focus on the number c^*. The latter is the derivative of the satisfaction probability $P(x^*(s))$ with respect to the parameter s when the solution path is reparameterized to move at a unit speed. Therefore a proxy for the finite-sample performance of Algorithm 11 is the sensitivity of the satisfaction probability along the direction of the solution path. The more sensitive it is, the better is the finite-sample performance. Note that this sensitivity does not explicitly depend on the dimension.

Here we provide the proof Theorem D.5.3: *Proof of Theorem D.5.3.* The proof follows the same line of argument as that of Theorem D.5.1, but uses a different set of deviation inequalities tailored to bounded random variables. To avoid repetition, we focus on the derivation of these deviation inequalities.

We need the following concentration inequalities for the sample mean and sample variance:

Lemma D.5.4 (Adapted from Maurer and Pontil (2009)) *Suppose $X_i, i = 1, \ldots, n$ are i.i.d. $[0,1]$-valued random variables, $\sigma^2 = \mathrm{Var}(X_1)$, and $\hat{\sigma}^2$ is the sample variance. Then we have for every $\epsilon \in (0,1)$ that*

$$P\Big(\big|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X_1]\big| > \sqrt{\frac{2\sigma^2\log(2/\epsilon)}{n}} + \frac{\log(2/\epsilon)}{3n}\Big) \le \epsilon$$

and

$$P\Big(|\hat{\sigma} - \sigma| > \sqrt{\frac{2\log(2/\epsilon)}{n-1}}\Big) \le \epsilon.$$

Proof of Lemma D.5.4. Theorem 3 in Maurer and Pontil (2009) gives the following Bennett's inequality

$$P\Big(\frac{1}{n}\sum_{i=1}^n X_i < \mathbb{E}[X_1] - \big(\sqrt{\frac{2\sigma^2\log(1/\epsilon)}{n}} + \frac{\log(1/\epsilon)}{3n}\big)\Big) \le \epsilon.$$

Applying the above inequality to $1 - X_i, i = 1, \ldots, n$ and noting that $\text{Var}(1 - X_1) = \text{Var}(X_1)$, we have

$$P\Big(\frac{1}{n}\sum_{i=1}^{n} X_i > \mathbb{E}[X_1] + \big(\sqrt{\frac{2\sigma^2 \log(1/\epsilon)}{n}} + \frac{\log(1/\epsilon)}{3n}\big)\Big) \leq \epsilon.$$

The first inequality in the lemma then comes from a union bound. The second inequality in the lemma is a direct consequence of Theorem 10 from Maurer and Pontil (2009). □

Let $H_j := \mathbb{P}_F(a_k' x^*(s_j) \leq b_k \text{ for all } k = 1, \ldots, K)$ be the satisfaction probability at $x^*(s_j)$, and $\sigma_j^2 := H_j(1 - H_j)$ be the variance. Applying Lemma D.5.4 to each $\mathbf{1}(a_k' x^*(s_j) \leq b_k \text{ for all } k = 1, \ldots, K)$ gives

$$|\hat{H}_j - H_j| \leq \sqrt{\frac{2H_j(1 - H_j)\log(2/t_1)}{n_2}} + \frac{\log(2/t_1)}{3n_2}$$

with probability at least $1 - t_1$, and

$$|\hat{\sigma}_j - \sigma_j| \leq \sqrt{\frac{2\log(2/t_1)}{n_2 - 1}}$$

with probability at least $1 - t_1$. Using a union bound, we have

$$|\hat{H}_j - H_j| \leq \sqrt{\frac{2H_j(1 - H_j)\log(2p/t_1)}{n_2}} + \frac{\log(2p/t_1)}{3n_2} \text{ for all } j = 1, \ldots, p \tag{D.44}$$

with probability at least $1 - t_1$, and that

$$|\hat{\sigma}_j - \sigma_j| \leq \sqrt{\frac{2\log(2p/t_1)}{n_2 - 1}} \text{ for all } j = 1, \ldots, p \tag{D.45}$$

with probability at least $1 - t_1$. When (D.45) happen, we also have

$$\hat{\sigma}_j \leq \sigma_j + |\hat{\sigma}_j - \sigma_j| \leq \sqrt{H_j(1 - H_j)} + \sqrt{\frac{2\log(2p/t_1)}{n_2 - 1}} \text{ for all } j = 1, \ldots, p.$$

Together with (D.44), we can conclude that, with probability at least $1-2t_1$, for all $j=1,\ldots,p$

$$\begin{aligned}
& |\hat{H}_j - \frac{z_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} - H_j| \\
\leq\ & \frac{z_{1-\beta}}{\sqrt{n_2}}\big(\sqrt{H_j(1-H_j)} + \sqrt{\frac{2\log(2p/t_1)}{n_2-1}}\big) + \sqrt{\frac{2H_j(1-H_j)\log(2p/t_1)}{n_2}} + \frac{\log(2p/t_1)}{3n_2} \\
\leq\ & 2\big(z_{1-\beta} + \sqrt{\log(2p/t_1)}\big)\big(\sqrt{\frac{H_j(1-H_j)}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big) \\
\leq\ & 2(1+z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{H_j(1-H_j)}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big) \qquad \text{(D.46)} \\
& \text{if we assume that } p \geq 2 \text{ so that } \log(2p/t_1) > 1.
\end{aligned}$$

Deviation bounds (D.45) and (D.46) are CCP counterparts of (D.34) and (D.35). Now we try to derive the CCP counterpart of (D.36). For any $\epsilon < \delta$ and every parameter value $s_j \in [s^*-\epsilon, s^*+\epsilon]$ we have by differentiability

$$|a_k' x^*(s_j) - a_k' x_S^*| = |\int_{s^*}^{s_j} a_k' \nabla x^*(s) ds| \leq \int_{s^*}^{s_j} |a_k' \nabla x^*(s)| ds \leq \eta_k(\epsilon) := \int_{s^*-\epsilon}^{s^*+\epsilon} |a_k' \nabla x^*(s)| ds.$$

Note that the sub-Gaussian norm $\|\cdot\|_{\psi_2} : \{X \text{ is a random variable} : \|X\|_{\psi_2} < \infty\} \to \mathbb{R}$ is a convex mapping, therefore by Jensen's inequality

$$\|\eta_k(\epsilon)\|_{\psi_2} \leq \int_{s^*-\epsilon}^{s^*+\epsilon} \big\|a_k' \nabla x^*(s)\big\|_{\psi_2} ds \leq \int_{s^*-\epsilon}^{s^*+\epsilon} D_2 \|\nabla x^*(s)\|_2 ds \leq 4D_2 \|\nabla x^*(s^*)\|_2 \epsilon.$$

With the above bound of $\eta_k(\epsilon)$, we want to quantify the closeness of the linear chance constraint

at the solutions $x^*(s_j)$ and x^*_S. We apply a union bound to obtain

$$\begin{aligned}
& |\mathbf{1}(a_k' x^*(s_j) \le b_k \text{ for all } k=1,\dots,K) - \mathbf{1}(a_k' x^*_S \le b_k \text{ for all } k=1,\dots,K)| \\
\le\ & \sum_{k=1}^K |\mathbf{1}(a_k' x^*(s_j) \le b_k) - \mathbf{1}(a_k' x^*_S \le b_k)| \\
\le\ & \sum_{k=1}^K \mathbf{1}(a_k' x^*(s_j) \le b_k < a_k' x^*_S \text{ or } a_k' x^*(s_j) > b_k \ge a_k' x^*_S) \\
\le\ & \sum_{k=1}^K \mathbf{1}(a_k' x^*_S - \eta_k(\epsilon) \le b_k < a_k' x^*_S \text{ or } a_k' x^*_S + \eta_k(\epsilon) > b_k \ge a_k' x^*_S) \\
\le\ & \sum_{k=1}^K \mathbf{1}(|a_k' x^*_S - b_k| \le \eta_k(\epsilon)).
\end{aligned}$$

Noting that difference of two indicator functions takes values in $\{-1,0,1\}$, we have

$$\begin{aligned}
& \mathbb{E}_F\big[\big(\mathbf{1}(a_k' x^*(s_j) \le b_k \text{ for all } k=1,\dots,K) - \mathbf{1}(a_k' x^*_S \le b_k \text{ for all } k=1,\dots,K)\big)^2\big] \\
=\ & \mathbb{E}_F\big[|\mathbf{1}(a_k' x^*(s_j) \le b_k \text{ for all } k=1,\dots,K) - \mathbf{1}(a_k' x^*_S \le b_k \text{ for all } k=1,\dots,K)|\big] \\
\le\ & \sum_{k=1}^K \mathbb{P}_F(|a_k' x^*_S - b_k| \le \eta_k(\epsilon)). \qquad \text{(D.47)}
\end{aligned}$$

In order to derive an upper bound for each of the K probabilities above, we first need a lower bound for $\|x^*_S\|_2$. If there are some $\tilde{k} \in \{1,2,\dots,K\}$ such that $b_{\tilde{k}} < 0$, then

$$1-\alpha = \mathbb{P}_F(a_k' x^*_S \le b_k \text{ for all } k=1,\dots,K) \le \mathbb{P}_F(a_{\tilde{k}}' x^*_S \le b_{\tilde{k}}) \le 2\exp\big(-\frac{\min_k |b_k|^2}{D_2^2 \left\|x^*_S\right\|_2^2}\big)$$

where in the last inequality Assumption D.5.4 is used. This forces $\|x^*_S\|_2 \ge \frac{\min_k |b_k|}{D_2\sqrt{\log(2/1-\alpha)}}$. Otherwise if all $b_k > 0$ then

$$\alpha = \mathbb{P}_F(\max_{k=1,\dots,K} a_k' x^*_S - b_k > 0) \le \sum_{k=1}^K \mathbb{P}_F(a_k' x^*_S > b_k) \le 2K\exp\big(-\frac{\min_k |b_k|^2}{D_2^2 \left\|x^*_S\right\|_2^2}\big)$$

which forces $\|x^*_S\|_2 \ge \frac{\min_k |b_k|}{D_2\sqrt{\log(2K/\alpha)}}$. When $\alpha > 1/2$, the second lower bound dominates hence $\|x^*_S\|_2 \ge \frac{\min_k |b_k|}{D_2\sqrt{\log(2K/\alpha)}}$ always holds. Now we go back to (D.47) and notice that for each k and

every $c > 0$

$$\begin{aligned}
\mathbb{P}_F(|a_k' x_S^* - b_k| \le \eta_k(\epsilon)) &\le \mathbb{P}_F(\eta_k(\epsilon) > \varepsilon) + \mathbb{P}_F(|a_k' x_S^* - b_k| \le \varepsilon) \\
&\le \mathbb{P}_F(\eta_k(\epsilon) > c) + \mathbb{P}_F(|a_k' x_S^* - b_k| \le c) \\
&\le 2\exp\big(-\frac{c^2}{16 D_2^2 \left\|\nabla x^*(s^*)\right\|_2^2 \epsilon^2}\big) + \frac{2 D_3 c}{\left\|x_S^*\right\|_2} \\
&\le 2\exp\big(-\frac{c^2}{16 D_2^2 \left\|\nabla x^*(s^*)\right\|_2^2 \epsilon^2}\big) + \frac{2 D_2 D_3 \sqrt{\log(2K/\alpha)} c}{\min_k |b_k|}.
\end{aligned}$$

With $c = 4 D_2 \left\|\nabla x^*(s^*)\right\|_2 \epsilon \cdot \sqrt{\log\Big(\max\Big\{e, \frac{\min_k |b_k|}{D_2^2 D_3 \|\nabla x^*(s^*)\|_2 \sqrt{\log(2K/\alpha)}\epsilon}\Big\}\Big)}$, the above bound gives

$$\mathbb{P}_F(|a_k' x_S^* - b_k| \le \eta_k(\epsilon)) \le 10\tilde{\epsilon}\sqrt{\log\big(\max\{e, \frac{1}{\tilde{\epsilon}}\}\big)}$$

where $\tilde{\epsilon} := \frac{D_2^2 D_3 \|\nabla x^*(s^*)\|_2 \sqrt{\log(2K/\alpha)}\epsilon}{\min_k |b_k|}$. From the union bound (D.47) it follows that

$$\begin{aligned}
&\mathbb{E}_F\big[\big(\mathbf{1}(a_k' x^*(s_j) \le b_k \text{ for all } k = 1, \dots, K) - \mathbf{1}(a_k' x_S^* \le b_k \text{ for all } k = 1, \dots, K)\big)^2\big] \\
\le\ & \sigma_\epsilon^2 := 10 K \tilde{\epsilon}\sqrt{\log\big(\max\{e, \frac{1}{\tilde{\epsilon}}\}\big)} \qquad \text{(D.48)}
\end{aligned}$$

for all $s_j \in [s^* - \epsilon, s^* + \epsilon]$. In particular, σ_ϵ^2 is a valid upper bound for the variance of each $\tilde{h}(x^*(s_j), \xi) := \mathbf{1}(a_k' x^*(s_j) \le b_k \text{ for all } k = 1, \dots, K) - \mathbf{1}(a_k' x_S^* \le b_k \text{ for all } k = 1, \dots, K)$ since the second moment always upper bounds the variance. Note that $(\tilde{h}(x^*(s_j), \xi) + 1)/2$ is $[0, 1]$-valued, hence applying Lemma D.5.4 to $(\tilde{h}(x^*(s_j), \xi) + 1)/2$ reveals that for all $s_j \in [s^* - \epsilon, s^* + \epsilon]$

$$|\hat{H}_j - H_j - (\hat{H}(x_S^*) - H(x_S^*))| \le \sqrt{\frac{2\sigma_\epsilon^2 \log(2/t_2)}{n_2}} + \frac{2\log(2/t_2)}{3 n_2}$$

with probability at least $1 - t_2$, therefore

$$\max_{j : |s_j - s^*| \le \epsilon} |\hat{H}_j - H_j - (\hat{H}(x_S^*) - H(x_S^*))| \le \sqrt{\frac{2\sigma_\epsilon^2 \log(2p/t_2)}{n_2}} + \frac{2\log(2p/t_2)}{3 n_2} \qquad \text{(D.49)}$$

with probability at least $1 - t_2$, a counterpart of (D.36).

As in the proof of Theorem D.5.1, we first derive the deviation inequality for $\hat{s}^*$. Based on (D.46), we would like to find $\underline{H}$ such that, with high probability, for all $H_j \leq \underline{H}$ we have $\hat{H}_j - \frac{z_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} < 1-\alpha$, as well as $\overline{H}$ such that all $H_j \geq \overline{H}$ satisfies $\hat{H}_j - \frac{z_{1-\beta}\hat{\sigma}_j}{\sqrt{n_2}} > 1-\alpha$. Given the bound (D.46), it suffices to $\underline{H}$ and $\overline{H}$ such that

$$H - 2(1+z_{1-\beta})\sqrt{\log(2p/t_1)}\Big(\sqrt{\frac{H(1-H)}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\Big) \quad > \quad 1-\alpha, \text{ for all } H \geq \overline{H} \tag{D.50}$$

$$H + 2(1+z_{1-\beta})\sqrt{\log(2p/t_1)}\Big(\sqrt{\frac{H(1-H)}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\Big) \quad < \quad 1-\alpha, \text{ for all } H \leq \underline{H} \tag{D.51}$$

For (D.50), since we must have $\overline{H} > 1-\alpha$, it holds that $H(1-H) < \alpha(1-\alpha) < \alpha$ for all $H \geq \overline{H}$. Therefore $\overline{H} := 1-\alpha+2(1+z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big)$ satisfies (D.50). For (D.51), since the left hand side is monotonic in H, we only need to find a $\underline{H}$ for which the inequality in (D.51) holds true. If

$$(1+z_{1-\beta})\sqrt{\log(2p/t_1)}\Big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\Big) < \frac{\sqrt{2}}{4}\alpha \tag{D.52}$$

then one can verify that $\underline{H} := 1-\alpha-2\sqrt{2}(1+z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big)$ satisfies (D.51) by noting that $\underline{H} > 1-2\alpha$ and hence $\underline{H}(1-\underline{H}) < 2\alpha$. In order for (D.52) to hold, we consider p, t_1, n_2 satisfying the following counterpart of (D.38)

$$2\epsilon_s < \epsilon(p,t_1,n_2) := \frac{6(1+z_{1-\beta})\sqrt{\log(2p/t_1)}}{c^* \left\|\nabla x^*(s^*)\right\|_2}\Big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\Big) < \frac{\delta}{2}. \tag{D.53}$$

We explain why (D.53) implies (D.52). Assumption D.5.3 stipulates that $1 \geq H(x^*(s^*+\delta)) \geq H(x^*_S) + \frac{1}{2}\nabla H(x^*_S)'\nabla x^*(s^*)\delta = 1-\alpha+\frac{1}{2}c^*\left\|\nabla x^*(s^*)\right\|_2\delta$, which leads to $c^*\left\|\nabla x^*(s^*)\right\|_2\delta \leq 2\alpha$. The second inequality in (D.53) then gives $(1+z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big) < \alpha/6 < \sqrt{2}\alpha/4$. Similar to the proof of Theorem D.5.1, when (D.53) holds and (D.46) happens, we must have $\hat{s}^* \in [s^* - 2\epsilon(p,t_1,n_2), s^* + 2\epsilon(p,t_1,n_2)]$. Therefore under the condition (D.53)

$$P_{\boldsymbol{\xi}_{1:n_2}}\big(|\hat{s}^* - s^*| > 2\epsilon(p,t_1,n_2)\big) \leq 2t_1. \tag{D.54}$$

Now we proceed to deal with the finite sample confidence error. Following the same steps of bounding the feasibility confidence level, we have

$$\begin{aligned} & P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \\ \geq\ & P_{\boldsymbol{\xi}_{1:n_2}}\big(\frac{\sqrt{n_2}(\hat{H}(x_S^*) - H(x_S^*))}{\sigma(x_S^*)} + \frac{\sqrt{n_2}}{\sigma(x_S^*)} \max_{j:|s_j - s^*| \leq 2\epsilon(p,t_1,n_2)} \big|\hat{H}_j - H_j - (\hat{H}(x_S^*) - H(x_S^*))\big| + \\ & \qquad \frac{z_{1-\beta}}{\sigma(x_S^*)} \max_{j:|s_j - s^*| \leq 2\epsilon(p,t_1,n_2)} |\sigma(x_S^*) - \hat{\sigma}_j| \leq z_{1-\beta},\ |\hat{s}^* - s^*| \leq 2\epsilon(p,t_1,n_2)\big). \end{aligned}$$

We bound the deviation of sample standard deviation as follows

$$\begin{aligned} \max_{j:|s_j - s^*| \leq 2\epsilon(p,t_1,n_2)} |\sigma(x_S^*) - \hat{\sigma}_j| &\leq \max_{j:|s_j - s^*| \leq 2\epsilon(p,t_1,n_2)} |\sigma(x_S^*) - \sigma_j| + \max_{j=1,\ldots,p} |\sigma_j - \hat{\sigma}_j| \\ &\leq \max_{j:|s_j - s^*| \leq 2\epsilon(p,t_1,n_2)} |\sqrt{\alpha(1-\alpha)} - \sqrt{H_j(1-H_j)}| + \max_{j=1,\ldots,p} |\sigma_j - \hat{\sigma}_j|. \end{aligned}$$

The second error is taken care of by (D.45). To bound the first error, we note that by Assumption D.5.3 we have $\max_{j:|s_j - s^*| \leq 2\epsilon(p,t_1,n_2)} |H_j - (1-\alpha)| \leq 2c^* \left\|\nabla x^*(s^*)\right\|_2 \cdot 2\epsilon(p,t_1,n_2) = 24(1 + z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big)$. Therefore if $24(1 + z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big) \leq \alpha/2$, it follows from applying mean value theorem that

$$\max_{j:|s_j - s^*| \leq 2\epsilon(p,t_1,n_2)} |\sqrt{\alpha(1-\alpha)} - \sqrt{H_j(1-H_j)}| \leq \frac{1}{\sqrt{\alpha}} 24(1+z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big).$$

Similar to (D.40), we can argue that $24(1 + z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big) \leq \alpha/2$ can be assumed without loss of generality so that the above bound can be assumed to hold. Together with (D.45), we have

$$\begin{aligned} & \max_{j:|s_j - s^*| \leq 2\epsilon(p,t_1,n_2)} |\sigma(x_S^*) - \hat{\sigma}_j| \\ \leq\ & \frac{24}{\sqrt{\alpha}}(1 + z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big) + \sqrt{\frac{2\log(2p/t_1)}{n_2 - 1}} \text{ for all } j = 1, \ldots, p \end{aligned} \tag{D.55}$$

with probability at least $1 - t_1$. Now we can further bound the confidence level

$$
\begin{aligned}
&P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \\
\geq &P_{\boldsymbol{\xi}_{1:n_2}}\Big(\frac{\sqrt{n_2}(\hat{H}(x_S^*) - H(x_S^*))}{\sigma(x_S^*)} + \frac{1}{\sigma(x_S^*)}\big(\sqrt{2\sigma^2_{2\epsilon(p,t_1,n_2)}\log(2p/t_2)} + \frac{2\log(2p/t_2)}{3\sqrt{n_2}}\big) \\
&\quad + \frac{24 z_{1-\beta}}{\sigma(x_S^*)\sqrt{\alpha}}(1 + z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big) + \frac{z_{1-\beta}}{\sigma(x_S^*)}\sqrt{\frac{2\log(2p/t_1)}{n_2 - 1}} \leq z_{1-\beta}\Big) \\
&\quad - t_2 - 3t_1 \quad \text{by (D.49), (D.54) and (D.55).}
\end{aligned}
$$

Like in the proof of Theorem D.5.1, applying Berry-Esseen theorem to the first probability on the right hand side and absorbing various constants into the universal constant C give

$$
\begin{aligned}
&1 - \beta - P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \\
\leq\ & C\Big(\frac{1}{\sqrt{\alpha n_2}} + \frac{1}{\sqrt{\alpha}}\big(\sqrt{2\sigma^2_{2\epsilon(p,t_1,n_2)}\log(2p/t_2)} + \frac{2\log(2p/t_2)}{3\sqrt{n_2}}\big) \\
&\quad + \frac{z_{1-\beta}}{\alpha}(1 + z_{1-\beta})\sqrt{\log(2p/t_1)}\big(\sqrt{\frac{\alpha}{n_2}} + \frac{\sqrt{\log(2p/t_1)}}{n_2}\big) + \frac{z_{1-\beta}}{\sqrt{\alpha}}\sqrt{\frac{2\log(2p/t_1)}{n_2}} + t_1 + t_2\Big) \\
\leq\ & C\Big(\frac{1}{\sqrt{\alpha}}\big(\sqrt{\sigma^2_{2\epsilon(p,t_1,n_2)}\log(2p/t_2)} + \frac{\log(2p/t_2)}{\sqrt{n_2}}\big) + (1 + z_{1-\beta})^2\sqrt{\frac{\log(2p/t_1)}{\alpha n_2}} + t_1 + t_2\Big) \\
&\quad \text{by keeping dominant terms only} \\
\leq\ & C\Big(\frac{1}{\sqrt{\alpha}}\big(\sqrt{\sigma^2_{2\epsilon(p,1/n_2,n_2)}\log(pn_2)} + \frac{\log(pn_2)}{\sqrt{n_2}}\big) + (1 + z_{1-\beta})^2\sqrt{\frac{\log(pn_2)}{\alpha n_2}} + \frac{1}{n_2}\Big) \\
&\quad \text{by taking } t_1 = t_2 = \frac{2}{n_2} \\
\leq\ & C\Big(\sqrt{\frac{\log(pn_2)}{\alpha}}\sigma_{2\epsilon(p,2/n_2,n_2)} + (1 + z_{1-\beta})^2\frac{\log(pn_2)}{\sqrt{\alpha n_2}}\Big) \\
\leq\ & C\Big(\sqrt{\frac{\log(pn_2)}{\alpha}}\sigma_{2\epsilon(p,2/n_2,n_2)} + (1 + z_{1-\beta})^2\frac{\log(pn_2)}{\sqrt{\alpha n_2}}\Big).
\end{aligned}
$$

It remains to bound the $\sigma_{2\epsilon(p,2/n_2,n_2)}$ term which by the definition (D.48) can be expressed as

$$
\sigma_{2\epsilon(p,2/n_2,n_2)} = \sqrt{20K\tilde{\epsilon}(p,n_2)}\big(\log(\max\{e, \frac{1}{2\tilde{\epsilon}(p,n_2)}\})\big)^{\frac{1}{4}}
$$

where

$$\tilde{\epsilon}(p,n_2)=\frac{6(1+z_{1-\beta})D_2^2D_3\sqrt{\log(2K/\alpha)}}{c^*\min_k|b_k|}\Big(\sqrt{\frac{\alpha\log(pn_2)}{n_2}}+\frac{\log(pn_2)}{n_2}\Big).$$

Note that $\frac{1}{2\tilde{\epsilon}(p,n_2)}\leq\frac{c^*\min_k|b_k|n_2}{D_2^2D_3\sqrt{\log(2K/\alpha)}}=\frac{n_2}{\tilde{C}}$, hence using this upper bound in the logarithm we have

$$\begin{aligned}\sigma_{2\epsilon(p,2/n_2,n_2)} &\leq 11\sqrt{(1+z_{1-\beta})K\tilde{C}}\Big(\log(\max\big\{e,\frac{n_2}{\tilde{C}}\big\})\Big)^{\frac{1}{4}}\Big(\sqrt{\frac{\alpha\log(pn_2)}{n_2}}+\frac{\log(pn_2)}{n_2}\Big)^{\frac{1}{2}}\\ &\leq 11\sqrt{(1+z_{1-\beta})K\tilde{C}}\Big(\log(\max\big\{e,\frac{n_2}{\tilde{C}}\big\})\Big)^{\frac{1}{4}}\Big[\Big(\frac{\alpha\log(pn_2)}{n_2}\Big)^{\frac{1}{4}}+\Big(\frac{\log(pn_2)}{n_2}\Big)^{\frac{1}{2}}\Big]\end{aligned}$$

where the second inequality follows because $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for any $a,b\geq 0$. Substituting $\sigma_{2\epsilon(p,2/n_2,n_2)}$ with its upper bound gives

$$\begin{aligned}&1-\beta-P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*))\geq\gamma)\\ \leq\;& C\Big(\Big[(1+z_{1-\beta})^2+\sqrt{(1+z_{1-\beta})K\tilde{C}}\Big(\log(\max\big\{e,\frac{n_2}{\tilde{C}}\big\})\Big)^{\frac{1}{4}}\Big]\frac{\log(pn_2)}{\sqrt{\alpha n_2}}\\ &\qquad+\sqrt{(1+z_{1-\beta})K\tilde{C}}\Big(\log(\max\big\{e,\frac{n_2}{\tilde{C}}\big\})\Big)^{\frac{1}{4}}\frac{(\log(pn_2))^{3/4}}{(\alpha n_2)^{1/4}}\Big)\\ \leq\;& C(1+z_{1-\beta})^2\Big(1+\sqrt{\tilde{C}K}\Big(\log(\max\big\{e,\frac{n_2}{\tilde{C}}\big\})\Big)^{\frac{1}{4}}\Big)\Big(\frac{(\log(pn_2))^{3/4}}{(\alpha n_2)^{1/4}}+\frac{\log(pn_2)}{\sqrt{\alpha n_2}}\Big)\\ \leq\;& C(1+z_{1-\beta})^2\Big(1+\sqrt{\tilde{C}K}\Big(\log(\max\big\{e,\frac{n_2}{\tilde{C}}\big\})\Big)^{\frac{1}{4}}\Big)\frac{(\log(pn_2))^{3/4}}{(\alpha n_2)^{1/4}}\end{aligned}$$

where the last inequality follows because $\frac{(\log(pn_2))^{3/4}}{(\alpha n_2)^{1/4}}\leq\frac{\log(pn_2)}{\sqrt{\alpha n_2}}$ if $\frac{(\log(pn_2))^{3/4}}{(\alpha n_2)^{1/4}}\leq 1$. Note again that this bound is valid when (D.53) is satisfied at $t_1=\frac{2}{n_2}$. Replacing e, the base of the natural logarithm, with 3 gives the desired bound. □

D.6 Applying Univariate Gaussian Validator to Formulations with Multidimensional Conservativeness Parameters

We consider the case of multidimensional conservativeness parameter, i.e., $S\subset\mathbb{R}^q$ for some $q\geq 2$, and present the asymptotic performance guarantees of the univariate Gaussian validator. We assume the following counterpart of Assumption 5.4.6:

Assumption D.6.1 (Piecewise uniformly continuous solution curve) *The parameter space $S \subset \mathbb{R}^q$ is compact, and there exist M connected and open subsets $S_1, \ldots, S_M$ of S such that (i) $S_i \cap S_{i'} = \emptyset$ for all $i \neq i'$; (ii) $\mathrm{m}(\cup_{i=1}^M S_i) = \mathrm{m}(S)$ where $\mathrm{m}(\cdot)$ denotes the Lebesgue measure on $\mathbb{R}^q$; and (iii) for each $i = 1, \ldots, M$, the optimal solution $x^*(s)$ of $OPT(s)$ exists and is unique for all $s \in S_i$, and $x^*(s)$ as a function of s is uniformly continuous on S_i.*

Similar to the case of scalar parameter, the solution curve $x^*(s)$ on each piece S_i can be continuously extended to the closure $\overline{S_i} := \cap_{S' \text{ is closed, } S_i \subseteq S'} S'$ under this piecewise uniform continuity assumption. Specifically, for every parameter value $s \in S \setminus \cup_{i=1}^M S_i$, we define the extended parameter-to-solution mapping to be

$$x^*(s) := \{\lim_{s' \in S_i, s' \to s} x^*(s') : s \in \overline{S_i}, i = 1, \ldots, M\}.$$

Accordingly, the optimal solution set and optimal parameter set associated with the solution path are defined as

$$\mathcal{X}_S^* := \mathrm{argmin}\{f(x) : H(x) \geq \gamma, x = x^*(s) \text{ for } s \in \cup_{i=1}^M S_i \text{ or } x \in x^*(s) \text{ for some } s \in S \setminus \cup_{i=1}^M S_i\} \tag{D.56}$$

and

$$S^* := \{s \in \cup_{i=1}^M S_i : x^*(s) \in \mathcal{X}_S^*\} \cup \{s \in S \setminus \cup_{i=1}^M S_i : x^*(s) \cap \mathcal{X}_S^* \neq \emptyset\}.$$

We also assume uniqueness of the optimal solution:

Assumption D.6.2 (Unique optimal solution) *The optimal solution set $\mathcal{X}_S^*$ defined in* (D.56) *is a singleton $\{x_S^*\}$.*

Note that in the case of scalar s, uniqueness of the optimal solution is a consequence (Proposition 5.4.3) of several more elementary assumptions among which monotonicity of the robust feasible set with respect to the parameter (Assumption 5.4.8) plays the key role. However, such notion of monotonicity does not completely carry to the mutidimensional case. For example, one may have a formulation $OPT(s)$ such that the robust feasible set satisfies $\mathrm{Sol}(s) \subseteq \mathrm{Sol}(s')$ whenever $s' \leq s$

component-wise, but $\mathrm{Sol}(s)$ and $\mathrm{Sol}(s')$ are in general not comparable.

We also assume the following counterpart of Assumption 5.4.9:

Assumption D.6.3 *For every $\epsilon > 0$ there exists an $s \in \cup_{i=1}^{M} S_i$ such that $H(x^*(s)) > \gamma$ and $\|x^*(s) - x_S^*\|_2 < \epsilon$, where x_S^* is the unique optimal solution from Assumption D.6.2.*

We then have the following asymptotic performance guarantees for Algorithm 11:

Theorem D.6.1 (Asymptotic joint feasibility-optimality guarantee) *Suppose Assumptions 5.4.1-5.4.5 hold for (5.1). Also suppose that Assumptions D.6.1-D.6.3 hold for the formulation $OPT(s)$, and that $\{s_1, \ldots, s_p\} \subseteq \cup_{i=1}^{M} S_i$. Denote by $\epsilon_S := \sup_{s \in S} \inf_{j=1,\ldots,p} \|s - s_j\|_2$ the mesh size, and by x_S^* be the unique optimal solution from Assumption D.6.2. Conditional on Phase one, as Phase two data size $n_2 \to \infty$, we have for the output of Algorithm 11 that (i) $\lim_{n_2 \to \infty, \epsilon_S \to 0} x^*(\hat{s}^*) = x_S^*$ and $\lim_{n_2 \to \infty, \epsilon_S \to 0} d(\hat{s}^*, \mathcal{S}^*) = 0$ almost surely; and (ii) $\liminf_{n_2 \to \infty, \epsilon_S \to 0} P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) \geq 1 - \beta$ if $H(x_S^*) = \gamma$, and $\lim_{n_2 \to \infty, \epsilon_S \to 0} P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) = 1$ if $H(x_S^*) > \gamma$.*

Proof of Theorem D.6.1. The proof is the same as that of Theorem 5.4.4 with straightforward modifications. In particular, $\{\tilde{s}_1, \ldots, \tilde{s}_M\}$ shall be replaced by $S \setminus \cup_{i=1}^{M} S_i$ and the solution set $\mathcal{X}_S$ is now defined as $\mathcal{X}_S := \{x^*(s) : s \in \cup_{i=1}^{M} S_i\} \cup \big(\cup_{s \in S \setminus \cup_{i=1}^{M} S_i} x^*(s) \big)$. □

In order to establish an asymptotically tight feasibility confidence level like in Theorem 5.4.5, we further assume uniqueness of the optimal parameter:

Assumption D.6.4 (Unique optimal parameter) *The optimal parameter set S^* is a singleton $\{s^*\}$, and $s^* \in S_{i^*}$ for some $i^* = 1, \ldots, M$.*

We then have the following guarantee:

Theorem D.6.2 (Asymptotically tight feasibility guarantee) *In addition to the conditions of Theorem D.6.1, suppose Assumption D.6.4 holds. Suppose also that the parameter-to-objective mapping $v(s)$ satisfies $v(s) < v(s')$ whenever $s < s'$ component-wise and that $H(x_S^*) = \gamma$. For each $j = 1, \ldots, p$, let*

$$\tilde{j} := \mathrm{argmin}_{j'} \{ \|s_j - s_{j'}\|_2 : s_{j'} < s_j \text{ component-wise}, s_{j'} \text{ lies on the same piece as } s_j \}$$

and if there is no such feasible j' simply let $\tilde{j} := j$. If the mesh is such that

$$\max_{j=1,\dots,p} |H(x^*(s_j)) - H(x^*(s_{\tilde{j}}))| = o\big(\frac{1}{\sqrt{n_2}}\big) \tag{D.57}$$

then we have for the output of Algorithm 11 that $\lim_{n_2\to\infty,\epsilon_S\to 0 \text{ s.t. (D.57) } holds} P_{\boldsymbol{\xi}_{1:n_2}}(H(x^*(\hat{s}^*)) \geq \gamma) = 1 - \beta$.

Proof of Theorem D.6.2. The proof follows exactly that of Theorem 5.4.5 with straightforward modifications. For example, when bounding $\hat{H}(x^*(\hat{s}^*)) - z_{1-\beta}\frac{\hat{\sigma}(\hat{s}^*)}{\sqrt{n_2}}$ in the proof of Theorem 5.4.5 we replace the parameter value $s_{j^*}^{i_{j^*}^*}$ output by the algorithm with $s_{j^*}^{i_{j^*}^*-1}$ and use the condition (5.21), whereas now we shall replace the output parameter value $\hat{s}^* = s_{j^*}$ with $s_{\tilde{j}^*}$ and then use (D.57) to obtain the same bound. □