Robust Actuarial Risk Analysis

Jose Blanchet* Henry Lam† Qihe Tang‡ Zhongyi Yuan§

Abstract

This paper investigates techniques for the assessment of model error in the context of insurance risk analysis. The methodology is based on finding robust estimates for actuarial quantities of interest, which are obtained by solving optimization problems over the unknown probabilistic models, with constraints capturing potential nonparametric misspecification of the true model. We demonstrate the solution techniques and the interpretations of these optimization problems, and illustrate several examples including calculating loss probabilities and conditional value-at-risk.

1 Introduction

This paper studies a methodology to quantify the impact of model assumptions that are uncertain or possibly incorrect in actuarial risk analysis. The motivation of our study is that, in many situations, coming up with a highly accurate model is challenging. This could be due to a lack of data, e.g., in describing a low-probability event, or modeling issues, e.g., in addressing a hidden and potentially sophisticated dependence structure among risk factors. Often times, actuaries and statisticians resort to models and calibration procedures that capture stylized features based on experience or expert knowledge, but bearing the risk of deviating too much from reality and thus leading to suboptimal decision-making. The methodology studied in this paper aims to quantify the incurred errors from these approaches to an extent that we shall describe.

On a high level, the methodology we study in this paper has the following characteristics:

*Department of Management Science and Engineering, Stanford University, Stanford, CA.
†Department of Industrial Engineering and Operations Research, Columbia University, New York, NY.
‡Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA.
§Department of Risk Management, Pennsylvania State University, University Park, PA.

1) Its starting point is a *baseline* model that, for any good reasons (e.g., a balance of fidelity and tractability), is currently employed by the actuary.

2) The method employs optimization theory to find a bound for the underlying risk metric of interest, where the optimization is non-parametrically imposed over all probabilistic models that are in the neighborhood of the baseline model.

3) Typically the worst-case model (i.e., the optimal solution obtained from the optimization procedure) can be written in terms of the baseline model. So the proposed procedure can be understood as a correction to allow for the possibility of model misspecification.

The key idea of the methodology lies in a suitable formulation of the optimization problem in 2) that leads to 3). This is an optimization cast over the space of models. This formulation has a feasible region described by a neighborhood of the baseline model, the latter believed to be close to a "true" or correct one - thanks to the expertise of the actuary. However, despite in the vicinity of the baseline, the location of the true model is unknown. The neighborhood that defines the feasible region, roughly speaking, serves to provide a more accessible indication of where the true model is. When this neighborhood happens to contain the true model, the optimization problem, namely a maximization or a minimization of the risk metric, will output an upper or a lower bound for the corresponding true value.

Thus, instead of outputting a risk metric value obtained from a single baseline model, the bound from our optimization can be viewed as a worst-case estimate from a collection of models that likely include the truth. In this sense, the bound is robust to model misspecification (to the extent of the neighborhood we impose). Our approach can also be viewed as a robustification of traditional scenario analysis or stress testing. Rather than testing the impact of risk metric due to specific shocks (by running the model at different scenarios or parameter values, say), our bound is capable of capturing the impact to the risk metric due to any perturbation of the model in the much bigger non-parametric space. We will illustrate these interpretations in several examples such as calculating loss probabilities and conditional value-at-risk.

The approach and the formulations that we discuss in this paper have roots in areas such as economics, operations research and statistics. The notion of optimization over models appears in the context of decision-making under ambiguity, i.e., when parts of the underlying model is

uncertain, the decision-maker resorts to optimizing decisions over the worst-case scenario, resulting in a minimax problem where the maximization is over the set of uncertain models. In the case of probabilistic models, this machinery involves an optimization over distributions, which is the framework we study in this paper. In stochastic control, such an approach has been used in deriving best control policies under ambiguous transition distributions [54, 52, 37]. In economics, the work of two Nobel laureates L. P. Hansen and T. J. Sargent [35] studies optimal decision making and its macroeconomic implications under ambiguity. Similar ideas have been used in quantitative finance including portfolio optimization [29, 30], as well as in physics and biology applications [3, 4]. In operations research, optimization over probabilistic models dates back to [60] in the context of inventory management, and has been used in [8, 9, 62] when some moment information is assumed known. The literature of distributionally robust optimization, which has been growingly active in recent years, e.g., [21, 32, 66, 38, 50], studies reformulations and efficient algorithms to handle uncertainty in probabilistic assumptions in various stochastic optimization settings. The specific type of constraints we study, namely a neighborhood defined via statistical distance, has been used in [6, 48, 42, 43, 36] under the umbrella of so-called ϕ-divergence, which in particular includes the Kullback-Leibler divergence or the relative entropy that we employ heavily. Other distances of recent interest include, e.g, the Wasserstein distance [28, 15]. In the context of extreme risks relevant to actuarial applications, [14, 5, 10, 23] study the use of distances, e.g., Renyi divergence, to capture model uncertainty in the tail, and [44, 49] study the incorporation of tail shape information, along with the shape-constrained distributional optimization literature [56, 34, 47, 63]. In the multivariate setting, [26, 58, 64, 24, 57, 27] study analytical solution and computation methods for bounding worst-case copulas. [20] investigates the use of moments and entropy maximization in portfolio selection to hedge downside risks of a mortality portfolio. Lastly, the calibration of the neighborhood size in distributionally robust optimization has been investigated via the viewpoints of, e.g., hypothesis testing (e.g., [7]), empirical likelihood and profile inference (e.g., [65, 46, 11, 25, 41, 33]) and Bayesian (e.g, [7]). For consistency, throughout this paper, we will adopt the terminology from the operations research literature and call our methodology *distributionally robust optimization*.

Our contribution in this paper is to bring the combination of the ideas in the above areas

to the attention of the actuarial community. In addition, given that risk assessment is of special importance for actuaries, we also discuss the implications of the distributionally robust methodology in the setting of tail events beyond these past literatures. We will illustrate our modeling approach and interpretations, how it can be used in actuarial problems, and solution methods employing a mix of analytical tools, simulation and convex optimization. We choose to present examples that are relatively simple for pedagogical purposes, but our goal is to convince the reader that the proposed methodology is substantially general.

2 Basic Distributionally Robust Problem Formulation

Suppose that in a generic application, one is interested in computing a performance measure in the form of an expected value of a function of underlying risk factors, i.e., $E_{true}\left(h\left(X\right)\right)$, where X is a random variable (or random vector) taking values in R^d, and $h : R^d \to R$ is a performance function. The notation $E_{true}\left(\cdot\right)$ denotes the expectation operator associated with an underlying "true" or correct probabilistic model, which is unknown to the actuary.

To quantify model error, we propose, in the most basic form, to use the following pair of maximization and minimization problems, which we call the basic distributionally robust (BDR) problem formulation:

$$\begin{aligned} &\min / \max \;\; E\left(h\left(X\right)\right) \\ &\text{s.t. } D\left(P||P_0\right) \leq \delta, \end{aligned} \tag{1}$$

The first involves solvingwhere

$$D(P\|P_0) = E_P\left(\log\left(\frac{dP}{dP_0}\left(X\right)\right)\right) = \int \log\left(\frac{dP}{dP_0}\left(x\right)\right) dP\left(x\right), \tag{2}$$

is the so-called Kullback-Leibler (KL) divergence [40] between P and P_0, with the integral in (2) taken over the region in which X takes its values and dP/dP_0 is the likelihood ratio between the probability models P and P_0. Optimization (1) has a decision variable P, and the expectation in its objective function is with respect to P. Finally, δ should be suitably calibrated (chosen as small

as possible) to guarantee that

$$D(P_{true}\|P_0) \leq \delta. \tag{3}$$

The KL divergence defined in (2) plays an important role in information theory and statistics (with connections to concepts such as entropy [18] and Fisher information [19]; it is also known as the relative entropy whose properties have been substantially studied. The KL divergence measures the discrepancy between two probability distributions, in the sense that $D(P\|P_0) = 0$ if and only if P is identical to P_0 (also see discussion below). Thus, the constraint in (1) can be viewed as a neighborhood around P_0 in the space of models, where the size of the neighborhood is measured by the KL divergence.

The reason for selecting δ depicted above is the following. First, by choosing δ so that inequality (3) holds, we guarantee that P_{true} is in the feasible region of (1). This translates, in turn, to that the interval obtained from the minimum and maximum values of (1) contains the true performance measure $E_{true}(h(X))$. Second, δ is chosen as small as possible because then this obtained interval is the shortest, signaling a higher accuracy of our estimate.

The probability space of X described above can be substantially more general, including infinite dimensional objects such as when X denotes a stochastic process like Brownian motion. However, to avoid technicalities, we will confine ourselves in the framework $X \in R^d$. In fact, to facilitate the discussion further, first we discuss the case where P is discrete finite. In this case, (1) can be written as

$$\begin{aligned}
&\min / \max \ \sum_{k=1}^{M} p\,(k)\, h\,(k) \\
&\text{s.t.} \\
&D(p\|p_0) = \sum_{k=1}^{M} p\,(k) \log\left(\frac{p\,(k)}{p_0\,(k)}\right) \leq \delta\ , \\
&\sum_{k=1}^{M} p\,(k) = 1, \quad p\,(k) \geq 0 \text{ for } k \geq 1\ ,
\end{aligned} \tag{4}$$

where the decision variable is the probability mass function $\{p(k) : k = 1, \ldots, M\}$ for some support size M, and $\{p_0(k) : k = 1, \ldots, M\}$ is the baseline probability mass function.

Let us briefly discuss two key properties of the KL divergence that make formulation (4) ap-

pealing. First, by Jensen's inequality, we have that

$$D(p\|p_0) = -\sum_{k=1}^{M} p\left(k\right) \log\left(\frac{p_0\left(k\right)}{p\left(k\right)}\right) \geq -\log\left(\sum_{k=1}^{M} p\left(k\right) \frac{p_0\left(k\right)}{p\left(k\right)}\right) = 0,$$

and $D(p\|p_0) = 0$ if and only if $p\left(k\right) = p_0\left(k\right)$ for all k. In other words, the $D\left(\cdot\right)$ allows us to compare the discrepancies between any two models, and the models agree if and only if there is no discrepancy. That is, in principle, the space of decision variables can include any distributions in the form $\{p\left(k\right) : 1 \leq k \leq M\}$, without confining to any parametric form. Thus the framework is capable to include model misspecification in a non-parametric manner.

It should be noted that $D\left(\cdot\right)$ is not a distance in the mathematical sense, because it does not obey the triangle inequality. However, and as the second key property of the KL divergence, $D(p\|p_0)$ is a convex function of $p\left(\cdot\right)$. To see this, first note that the function $l\left(x\right) = x\log\left(x\right)$ is convex on $(0, \infty)$. Second, observe that if $\{p\left(k\right) : 0 \leq k \leq M\}$ and $\{q\left(k\right) : 0 \leq k \leq M\}$ are two probability distributions and $\alpha \in (0, 1)$, then, because of Jensen's inequality, for any fixed k,

$$l\left(\frac{\alpha p\left(k\right) + \left(1-\alpha\right) q\left(k\right)}{p_0\left(k\right)}\right) \leq \alpha l\left(\frac{p\left(k\right)}{p_0\left(k\right)}\right) + \left(1-\alpha\right) l\left(\frac{q\left(k\right)}{p_0\left(k\right)}\right).$$

Thus, we conclude

$$\begin{aligned} D(\alpha p + (1-\alpha) q\|p_0) &= \sum_{k=1}^{M} p_0\left(k\right) l\left(\frac{\alpha p\left(k\right) + \left(1-\alpha\right) q\left(k\right)}{p_0\left(k\right)}\right) \\ &\leq \sum_{k=1}^{M} \left[\alpha p_0(k) l\left(\frac{p\left(k\right)}{p_0\left(k\right)}\right) + \left(1-\alpha\right) p_0(k) l\left(\frac{q\left(k\right)}{p_0\left(k\right)}\right)\right] \\ &= \alpha D(p\|p_0) + \left(1-\alpha\right) D(q\|p_0). \end{aligned}$$

Consequently, BDR problem (4) is a convex optimization problem with a linear objective function. These types of problems have been well-studied and are computationally tractable using operations research tools.

Later in the paper we will discuss the implications of KL divergence as a notion of discrepancy and propose other notions and constraints. For now, we argue that there are direct variations of the BDR problem formulation that can be easily accommodated within the same convex optimization

framework. For this discussion, it helps to think of $p(k)$ as the mortality distribution at each age (so $k = 1, \ldots, 100$ for instance). Suppose that an actuary is less confident in the estimate for $p_{true}(k)$ for small values of k; that is, assume that $p_0(k) \approx p_{true}(k)$ for large values of k, but the actuary is uncertain about how similar $p_0(k)$ is to $p_{true}(k)$ for small values of k (this arises if the mortality estimate is more credible for some age groups for instance). Then we can replace the first inequality constraint in (4) by introducing a weighting function $\{w(k) : 1 \leq k \leq M\}$ with $w(k) > 0$ which is *increasing,* thus obtaining

$$\sum_{k=1}^{M} w(k)\, p(k) \log\left(\frac{p(k)}{p_0(k)}\right) \leq \delta\,. \tag{5}$$

To understand why $w(\cdot)$ should be increasing in order to account for model errors from misspecifying $p_0(k)$ for small values of k, consider the following exemplary case. Suppose for some $k_0 > 0$, we have 1) $w(k) = \varepsilon > 0$ for $k \leq k_0$ and ε small, and 2) $w(k) = 1$ for $k > k_0$. Then observe that the constraint (5) is relatively insensitive to the value of $p(k)$ for $k \leq k_0$. Therefore, the convex optimization program will have more freedom for the $p(k)$ on $k \leq k_0$ to jitter and improve the objective function without having a significant impact on feasibility.

Another variation of the BDR formulation includes moment constraints. For instance, suppose additional information is known for the expected time-until-death for individuals who have a particular underlying medical condition and are at least 30 years old (from a series of medical studies, say). Using such information one might impose a constraint of the form $E(X|X \geq 30) \in [a_-, a_+]$ for a specified range $[a_-, a_+]$, or equivalently $E(XI(X \geq 30)) \in [P(X \geq 30)a_-, P(X \geq 30)a_+]$ (throughout the paper, we use $I(A)$ to denote the indicator of A, i.e., $I(A) = 1$ if A occurs and $= 0$ if not). Using this information, one can add the constraints

$$a_- \sum_{k=30}^{M} p(k) \leq \sum_{k=30}^{M} p(k)\, k \leq a_+ \sum_{k=30}^{M} p(k)\,, \tag{6}$$

which are linear inequalities, and therefore the resulting optimization problem is still convex and tractable. In Section 9 we will continue discussing other constraints that can inform the BDR formulation with alternate forms of expert knowledge.

At this point, several questions might be in order: How do we solve the BDR optimization problem and its variations? How can we understand such solution intuitively? What is the role

of the constraint in (1)? How do we choose δ? How do we extend the methodology to deal with possibly multidimensional distributions? Our goal is to address these questions throughout the rest of this paper.

3 Solving the BDR Formulation

This section describes how to solve (4) and then transitions toward the more general problem (1) in an intuitive way. We concentrate on the problem of maximization; the minimization counterpart is analogous, and we will summarize the differences at the end of our discussion.

3.1 The Maximization Form

We introduce Lagrange multipliers to solve the convex optimization problem (4). The Lagrangian takes the form

$$g\left(p\left(1\right),...,p\left(M\right),\lambda_1,\lambda_2\right)=\sum_{k=1}^{M}p\left(k\right)h\left(k\right)-\lambda_1\left(\sum_{k=1}^{M}p\left(k\right)\log\left(\frac{p\left(k\right)}{p_0\left(k\right)}\right)-\delta\right)-\lambda_2\left(\sum_{k=1}^{M}p\left(k\right)-1\right).$$

The Karush-Kuhn-Tucker (KKT) [16] conditions in our (convex optimization) setting characterize an optimal solution. Denoting an optimal solution for (4) as $\{p_+\left(k\right):1\leq k\leq M\}$, and the corresponding Lagrange multipliers as λ_1^+ and λ_2^+, the KKT conditions are as follows (we use "+" to denote an optimal solution for the maximization formulation, as opposed to "−" for the minimiza-

tion counterpart, which we shall discuss momentarily as well):

$$\frac{\partial g}{\partial p\left(k\right)}\left(p_{+}\left(1\right),...,p_{+}\left(M\right),\lambda_{1}^{+},\lambda_{2}^{+}\right)=h\left(k\right)-\lambda_{1}^{+}\left(1+\log\left(\frac{p_{+}\left(k\right)}{p_{0}\left(k\right)}\right)\right)-\lambda_{2}^{+}\leq 0,$$
$$k=1,\ldots,M, \tag{7}$$
$$p_{+}(k)\frac{\partial g}{\partial p\left(k\right)}\left(p_{+}\left(1\right),...,p_{+}\left(M\right),\lambda_{1}^{+},\lambda_{2}^{+}\right)=p_{+}(k)\left[h\left(k\right)-\lambda_{1}^{+}\left(1+\log\left(\frac{p_{+}\left(k\right)}{p_{0}\left(k\right)}\right)\right)-\lambda_{2}^{+}\right]=0,$$
$$k=1,\ldots,M, \tag{8}$$
$$\sum_{k=1}^{M}p_{+}\left(k\right)\log\left(\frac{p_{+}\left(k\right)}{p_{0}\left(k\right)}\right)\leq\delta,\quad\sum_{k=1}^{M}p_{+}\left(k\right)=1,\ \ p_{+}\left(k\right)\geq 0\text{ for }1\leq k\leq M, \tag{9}$$
$$\lambda_{1}^{+}\geq 0,\lambda_{2}^{+}\text{ is free,} \tag{10}$$
$$\lambda_{1}^{+}\left(\sum_{k=1}^{M}p_{+}\left(k\right)\log\left(\frac{p_{+}\left(k\right)}{p_{0}\left(k\right)}\right)-\delta\right)=0. \tag{11}$$

The relations (7) and (8) correspond to the so-called stationarity conditions, (9) the primal feasibility, (10) the dual feasibility, and (11) the complementary slackness condition.

Define

$$\mathcal{M}_{+}=\text{argmax}\{h\left(k\right):1\leq k\leq M\},$$

i.e., $\mathcal{M}_{+}$ is the index set on which $h\left(\cdot\right)$ achieves its maximum value. Also, denote $h^{*}=\max\{h(k):1\leq k\leq M\}$ as the maximum value of $h(\cdot)$. We analyze the solution by dividing it into two cases, depending on whether $\log\left(\frac{1}{P_{0}(X\in\mathcal{M}_{+})}\right)\leq\delta$:

Case 1: $\log\left(\frac{1}{P_{0}(X\in\mathcal{M}_{+})}\right)\leq\delta$**.**

Consider

$$p_{+}(k)=\frac{p_{0}\left(k\right)I\left(k\in\mathcal{M}_{+}\right)}{\sum_{j\in\mathcal{M}_{+}}p_{0}\left(j\right)}=P_{0}\left(X=k\mid X\in\mathcal{M}_{+}\right). \tag{12}$$

for $k\in\mathcal{M}_{+}$, and $p_{+}(k)=0$ otherwise, i.e., $p_{+}(\cdot)$ is the conditional distribution of X given $X\in\mathcal{M}_{+}$. Note that this choice of $p_{+}(\cdot)$ gives

$$\sum_{k=1}^{M}p_{+}\left(k\right)\log\left(\frac{p_{+}\left(k\right)}{p_{0}\left(k\right)}\right)=\log\left(\frac{1}{P_{0}\left(X\in\mathcal{M}_{+}\right)}\right)\leq\delta$$

so that (9) holds. Moreover, choosing $\lambda_{1}^{+}=0$ and $\lambda_{2}^{+}=h^{*}$, we have (7), (8), (10) and (11) all hold. Thus, (12) is an optimal solution in this case.

Case 2: $\log\left(\frac{1}{P_0(X\in\mathcal{M}_+)}\right) > \delta$**.**

Consider setting

$$h\left(k\right) - \lambda_1^+\left(1 + \log\left(\frac{p_+\left(k\right)}{p_0\left(k\right)}\right)\right) - \lambda_2^+ = 0,\ k = 1, \ldots, M,$$

giving $p_+(k) = p_0(k)\exp((h(k) - \lambda_1^+ - \lambda_2^+)/\lambda_1^+)$. Denoting $\theta_+ = 1/\lambda_1^+$ and $\psi_+ = \lambda_2^+/\lambda_1^+ + 1$, we can write

$$p_+\left(k\right) = p_0\left(k\right)\exp\left(\theta_+ h\left(k\right) - \psi_+\right). \tag{13}$$

To satisfy $\sum_{k=1}^M p_+(k) = 1$ in (9), we must have $\psi_+ = \log\sum_{k=1}^M p_0\left(k\right)\exp\left(\theta_+ h\left(k\right)\right)$, i.e., ψ_+ is the logarithmic moment generating function of $h(X)$ under $p_0(\cdot)$ at θ_+. Now, to satisfy (11), we enforce θ_+ to be the positive root of the equation

$$\sum_{k=1}^M p_+\left(k\right)\log\left(\frac{p_+\left(k\right)}{p_0\left(k\right)}\right) = \theta_+\frac{\sum_{k=1}^M p_0\left(k\right)\exp\left(\theta_+ h\left(k\right)\right)h\left(k\right)}{\sum_{k=1}^M p_0\left(k\right)\exp\left(\theta_+ h\left(k\right)\right)} - \log\left(\sum_{k=1}^M p_0\left(k\right)\exp\left(\theta_+ h\left(k\right)\right)\right) = \delta, \tag{14}$$

where the first equality is obtained by plugging in the expression of $p_+(\cdot)$ in (13). We argue that such a positive root must exist. Note that the left hand side of (14) is continuous and increasing (continuity is immediate, and monotonicity can be verified by somewhat tedious but elementary differentiation). When $\theta_+ \to 0$ in (14), the left hand side goes to 0. When $\theta_+ \to \infty$, it becomes

$$\begin{aligned} & \theta_+ h^* + O(\exp(-c\theta_+)) - \left(\theta_+ h^* + \log\sum_{k\in\mathcal{M}_+} p_0(k) + O(\exp(-c\theta_+)\right) \\ = \ & -\log\sum_{k\in\mathcal{M}_+} p_0(k) + O(\exp(-c\theta_+)) \end{aligned}$$

for some constant $c > 0$, by singling out the dominant exponential factor $\exp(\theta_+ h^*)$ and noticing a relative exponential decay in the remaining terms in the expression in (14). But note that $-\log\sum_{k\in\mathcal{M}_+} p_0(k) + O(\exp(-c\theta_+)) > \delta$ as $\theta_+ \to \infty$ in our considered case. Thus we must have a positive root for (14). Consequently, one can verify straightforwardly that the choice of $p_+(\cdot)$ in (13) with $\theta_+ > 0$ solving (14) satisfies all of (7), (8), (9), (10) and (11).

We comment that Case 1 can be regarded as a degenerate case and rarely arises in practice,

while Case 2 is the more important case to consider. From (13) in Case 2, $p_+(\cdot)$ can be interpreted as a member of a "natural exponential family", also known as an "exponential tilting" distribution that arises often in statistics, large deviations analysis and importance sampling in Monte Carlo simulation. On the other hand, the form of $p_+(\cdot)$ in (12) in Case 1 can be interpreted as a limit of (13) as $\theta_+ \to \infty$, namely

$$p_+(k) = \lim_{\theta\to\infty} p_0(k) \frac{\exp(\theta h(k))}{\sum_{j=1}^M p_0(j)\exp(\theta h(j))} = \frac{p_0(k) I(k \in \mathcal{M}_+)}{\sum_{j\in\mathcal{M}_+} p_0(j)} = P_0(X = k \mid X \in \mathcal{M}_+).$$

3.2 The Minimization Form

The minimization form of the BDR problem is analogous to the maximization one. In this case, the Lagrangian takes the form

$$g(p(1),...,p(M),\lambda_1,\lambda_2) = \sum_{k=1}^M p(k) h(k) + \lambda_1 \left(\sum_{k=1}^M p(k)\log\left(\frac{p(k)}{p_0(k)}\right) - \delta\right) + \lambda_2\left(\sum_{k=1}^M p(k) - 1\right).$$

With the corresponding KKT conditions and similar discussion as before, we arrive at the two analogous cases. Here, we define

$$\mathcal{M}_- = \text{argmin}\{h(k) : 1 \le k \le M\}.$$

Case 1: $\log\left(\frac{1}{P_0(X\in\mathcal{M}_-)}\right) \le \delta$**.**

An optimal solution is given by

$$p_-(k) = P_0(X = k | X \in \mathcal{M}_-).$$

Case 2: $\log\left(\frac{1}{P_0(X\in\mathcal{M}_-)}\right) > \delta$**.**

An optimal solution is given by

$$\begin{aligned} p_-(k) &= p_0(k)\exp(-\theta_- h(k) - \psi(\theta_-)), \\ \psi(\theta_-) &= \log\sum_{k=1}^M p_0(k)\exp(-\theta_- h(k)), \end{aligned}$$

with $\theta_- > 0$ satisfying

$$-\theta_- \frac{\sum_{k=1}^{M} p_0(k) \exp(-\theta_- h(k)) h(k)}{\sum_{k=1}^{M} p_0(k) \exp(-\theta_- h(k))} - \log\left(\sum_{k=1}^{M} p_0(k) \exp(-\theta_- h(k))\right) = \delta.$$

3.3 Solutions for the General Formulation

The insights of the solutions obtained in Sections 3.1 and 3.2 can be extended to the general BDR formulation (1). First, denoting $E_0(\cdot)$ as the expectation under the baseline model $P_0(\cdot)$, (1) can be written equivalently as optimizing over all random variables $Z(\omega)$, where ω is an element of the underlying outcome space Ω, in the form

$$\begin{aligned} &\max \ E_0(h(X) Z) \\ &\text{s.t.} \\ &E_0(Z \log(Z)) \leq \delta \\ &Z(\omega) \geq 0, \text{ for all } \omega, \text{ and } \ E_{P_0}(Z) = 1 \ . \end{aligned} \tag{15}$$

The object $Z(\omega)$ is the likelihood ratio between P and P_0, which is well-defined since the KL divergence is defined only for P absolutely continuous with respect to P_0. The solution to (15), denoted $(Z_+(\omega) : \omega \in \Omega)$, can be used to obtain the solution $P_+(\cdot)$ to (1), by defining $P_+(\cdot)$ via

$$P_+(X \in A) = E_0(Z_+ I(X \in A)).$$

Conceptually, the general problem formulation is not much different from the finite outcome case (4). It can be shown (e.g., [54, 17]) that exactly the same form of the optimal solution applies. For instance, if $f_0(\cdot)$ is the density of X in R^d, then

$$\begin{aligned} Z_+ &= \frac{f_+(X)}{f_0(X)} = \exp(\theta_+ h(X) - \psi(\theta_+)), \\ \psi(\theta_+) &= \log E_0(\exp(\theta_+ h(X))), \end{aligned}$$

with

$$\theta_+ \frac{E_0\left(\exp\left(\theta_+ h\left(X\right)\right) h\left(X\right)\right)}{E_0\left(\exp\left(\theta_+ h\left(X\right)\right)\right)} - \log\left(E_0\left(\exp\left(\theta_+ h\left(X\right)\right)\right)\right) = \delta$$

in case $\log\left(1/P_0\left(X \in \mathcal{M}_+\right)\right) > \delta$, where

$$\mathcal{M}_+ = \text{argmax}\{h\left(x\right) : x \in R^d\}.$$

However, if $\log\left(1/P_0\left(X \in \mathcal{M}_+\right)\right) \leq \delta$, then the optimal solution is a probability distribution given by

$$P_+\left(X \in A\right) = P_0\left(X \in A | X \in \mathcal{M}_+\right).$$

3.4 Numerical Example of BDR Formulation

Let us assume that X is the time-until-death of an individual entering a whole life insurance contract that pays \$1 benefit to the beneficiaries at the time of death. Assume that the force of interest (continuously compounded) is given by $r > 0$. We are interested in the expected net present value of the benefits paid, denoted as $h\left(X\right) = \exp\left(-rX\right)$. This expected value is thus given by $E_{true}\left(h\left(X\right)\right) = E_{true}\left(\exp\left(-rX\right)\right)$. A more realistic example might consider stochastic interest rates or a portfolio of different types of contracts, but here we stay with this simple example for clear illustration.

Consider the more concrete setting where X takes values only on the set $\{1, ..., 100\}$, the benefit is paid in integer years, and the individual in consideration is 20 years old so that the individual could die at ages 21, 22,..., 120. Now, imagine a situation in which we know that the individual has a medical condition for which not much is known, so there might be substantial variability in our estimation of the distribution of X. Or perhaps there is a significant likelihood that medical advances might be expected to occur in the next 10 years or so, and therefore, if $p_0\left(\cdot\right)$ were calibrated based on historical data, $p_0\left(\cdot\right)$ might simply be an incorrect model.

To illustrate numerically the BDR formulation (4), we set the baseline distribution $p_0(k)$ from the recent static mortality table under the Internal Revenue Code ([2], Appendix, "unisex" column), and consider the estimation of the worst-case expected payoff when the true mortality distribution deviates from $p_0(k)$ as described by (4).

Figure 1 shows the maximum and minimum values of (4) under different δ, when the force of interest is set at $r = .015$. The maximum value increases with δ, while the minimum value decreases with δ, both at a decreasing rate. The effects on the maximum values seem to be stronger than the minimum values when the distribution changes from the baseline, as shown by a larger magnitude of the upward movement as δ increases. Note that the bounds, especially the upper one, grow quite wide as δ increases. This indicates that under no other information, the worst-case impact of model uncertainty can be quite significant. This impact can be reduced if one adds auxiliary information, either from additional data source or from expert knowledge (e.g., moment equalities or inequalities in (6) and in the subsequent Section 9).

Figures 2, 3 and 4 show the shapes of the distributions giving rise to the maximum and minimum expected payoffs, compared with the baseline distribution, at $\delta = 1$, 2 and 5 respectively. We can see a sharp concentration of mass close to 0 for the maximal distribution, as higher chance of sooner death leads to a larger expected present value of payoff. In contrast, more mass is located toward older ages for the minimal distribution, leading to a reduction in the expected present value of payoff. As δ increases, the accumulation of the mass at 0 for the maximal distribution tends to be heavier, as well as the shift of mass to older ages for the minimal distribution. The mass shift for the maximal distribution seems to be more dramatic than for the minimal one, which is consistent with Figure 1 that shows a wider bound for the maximization problem.

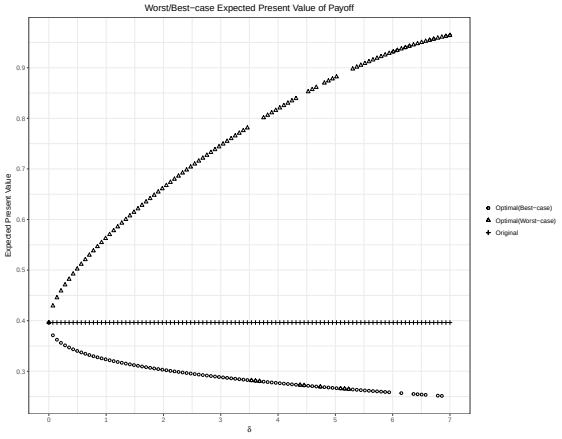

Figure 1: Robust estimates against δ

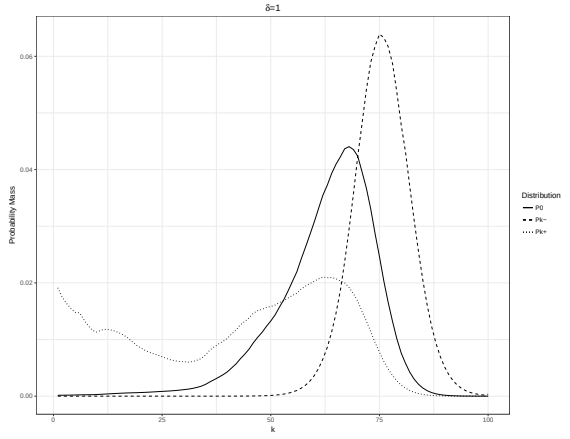

Figure 2: Optimal probability mass function at $\delta = 1$

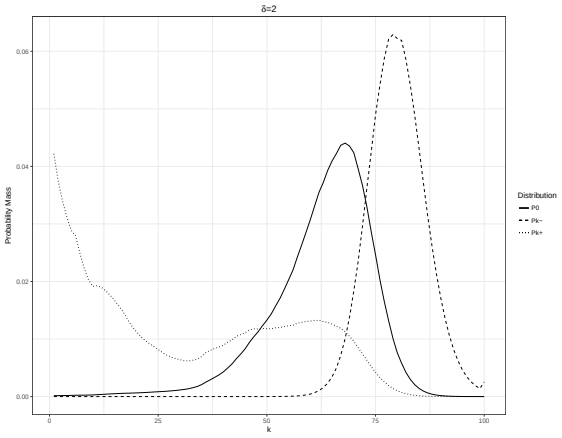

Figure 3: Optimal probability mass function at $\delta = 2$

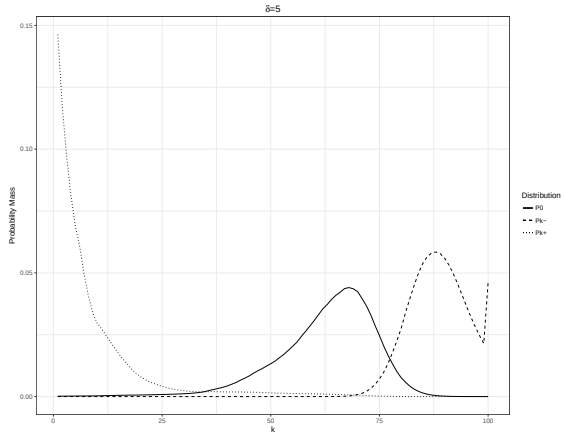

Figure 4: Optimal probability mass function at $\delta = 5$

4 Distributionally Robust Analysis of Dependence

We consider a variation of the BDR formulation that is suitable for quantifying the impact of dependence. We revisit the example of whole life insurance in Section 3.4 as an illustration, but now we assume that a couple is interested in a contract that pays \$1 benefit at the time of the first death between the couple. In this case, the payoff equals $h(X,Y) = \exp(-r\min(X,Y))$, where X and Y are the time-until-death of the individual and his or her spouse, respectively. We assume that both spouses are 20 years old at the time of signing the risk premium valuation, so that the pair (X,Y) is supported on the set $\{1,...,100\} \times \{1,...,100\}$.

We are interested in the potential impact on the estimated actuarial net present value of the benefits due to unaccounted dependence structures. Suppose that, this time, enough information is available to estimate the distributions of X and Y marginally with relatively high accuracy, but the joint distribution is difficult to estimate. That is, the marginals are known to be

$$\begin{aligned} P_{true}(Y=j) &= \sum_{i=1}^{100} P_0(X=i, Y=j), \\ P_{true}(X=i) &= \sum_{j=1}^{100} P_0(X=i, Y=j). \end{aligned}$$

A joint baseline model $p_0(i,j) = P(X=i, Y=j)$ can be conjectured (e.g., using an independent marginal assumption).

To capture the above type of uncertainty, the distributionally robust (maximization) formulation

is given by

$$\begin{aligned}
&\max \sum_{i=1}^{M}\sum_{j=1}^{M} p\left(i,j\right) h\left(i,j\right) \\
&\text{s.t.} \\
&\sum_{i=1}^{M}\sum_{j=1}^{M} p\left(i,j\right) \log\left(\frac{p\left(i,j\right)}{p_0\left(i,j\right)}\right) \leq \delta, \\
&\sum_{j=1}^{M} p\left(i,j\right) = P_0\left(X=i\right) \text{ for all } i, \ \sum_{i=1}^{M} p\left(i,j\right) = P_0\left(Y=j\right) \text{ for all } j, \\
&p\left(i,j\right) \geq 0 \text{ for all } i,j,
\end{aligned} \tag{16}$$

where we write (16) in a more general fashion that outputs the worst-case expected value $E[h(X,Y)]$ over $P(X,Y)$ supported on $\{(i,j) : 1 \leq i \leq M, 1 \leq j \leq M\}$. The first constraint is a KL divergence from the baseline $p_0(i,j)$ like in (1). The second constraint stipulates that the marginal distributions are known to be $P_0(X=i)$ and $P_0(Y=j)$. In the joint whole life insurance example, M would be set as 100.

We discuss how to solve (16), concentrating on the case analogous to Case 2 analyzed in the BDR formulation in Section 3.1 (i.e., assuming the neighborhood size δ is not too big, which is the more interesting case).

It is worth mentioning that $p = p_0$ is a feasible solution, and therefore we can verify the Slater condition (see [16]). Such a condition guarantees that strong duality holds and that the KKT conditions are necessary and sufficient for optimality. Consequently, commercial packages can be used to solve a convex optimization problem like (16) very quickly. Here, we will describe an iterative procedure based on exponential tilting similar to that for problem (4). In particular, an analogous argument in using the stationarity KKT condition in Case 2 there leads to an optimal solution of the form

$$p_+\left(i,j\right) = P_0\left(X=i, Y=j\right) \exp\left(\theta_+ h\left(i,j\right) - \alpha_{\theta_+}\left(i\right) - \beta_{\theta_+}\left(j\right)\right),$$

where

$$\sum_j \exp\left(\theta_+ h\left(i,j\right) - \beta_{\theta_+}\left(j\right)\right) P_0\left(X=i, Y=j\right) \;=\; \exp\left(\alpha_{\theta_+}\left(i\right)\right) P_0\left(X=i\right), \tag{17}$$

$$\sum_i \exp\left(\theta_+ h\left(i,j\right) - \alpha_{\theta_+}\left(i\right)\right) P_0\left(X=i, Y=j\right) \;=\; \exp\left(\beta_{\theta_+}\left(j\right)\right) P_0\left(Y=j\right). \tag{18}$$

where (17) and (18) guarantee the marginal distributions match the known values. The form above motivates the following iterative scheme (see [22, 31]): Given a θ, define

$$p_\theta(i,j) = \frac{\exp\left(\theta h\left(i,j\right)\right) p_0\left(i,j\right)}{\sum_k \sum_l \exp\left(\theta h\left(k,l\right)\right) p_0\left(k,l\right)}.$$

Then iteratively update $p_\theta(i,j)$ so that at each step, for each $i,j = 1,\ldots,M$ replace $p_\theta(i,j)$ with

$$\frac{p_\theta(i,j) P_0(X=i)}{\sum_l p_\theta(i,l)}$$

and then for each $i,j = 1,\ldots,M$, replace $p_\theta(i,j)$ with

$$\frac{p_\theta(i,j) P_0(Y=j)}{\sum_k p_\theta(k,j)}$$

until convergence. The iteration ensures the marginal distributions of the resulting probability mass function are equal to $P_0(X=i), i=1,\ldots,M$, and $P_0(Y=j), j=1,\ldots,M$, respectively, leading to a final $p_\theta(\cdot,\cdot)$. Then, we solve

$$\min_\theta \sum_{i=1}^M \sum_{j=1}^M p_\theta\left(i,j\right) h\left(i,j\right) - \frac{1}{\theta}\left(\sum_{i=1}^M \sum_{j=1}^M p_\theta\left(i,j\right) \log\left(\frac{p_\theta\left(i,j\right)}{p_0\left(i,j\right)}\right) - \delta\right)$$

to get θ_+. An optimal mass function is then $p_+(i,j) = p_{\theta_+}(i,j)$.

4.1 Numerical Example on Joint Mortality

We illustrate numerically the solution to optimization (16). Consider, as discussed above, the estimation of the expected payoff $h(X,Y) = \exp(-r\min(X,Y))$ in the described setting. We set the baseline joint distribution $p_0(i,j)$ from the same mortality table used in Section 3.4, but now

assuming independence of the times-until-death of the couple, i.e. $p_0(i,j) = P_0(X=i)P_0(Y=j)$. This independent baseline model describes the simplest assumption made by insurers when pricing life products, and we are interested in a robust estimate of the expected payoff when this assumption is violated.

Figure 5 shows the worst scenario values of (16) under different values of δ, setting $r = .015$. As depicted, there is a concavely increasing trend for the worst-case expected payoff as the level of dependency deviates from independence. While the estimate is 0.0729 according to the independent model, it can potentially rise to 0.0734 when the true model is misspecified from independence to a level of $\delta = 0.002$. Note that the magnitude of the changes in worst-case estimates is significantly less dramatic than the example in Section 3.4. This is because here we have injected complete marginal information, and the model is allowed to deviate only in terms of dependence structure. From the small change magnitude, we can see that the marginal information is an important piece that substantially reduces the freedom of deviation.

Figures 6 and 7 further show the approximate surfaces of the optimal joint probability mass function at $\delta = 0.013$ and the original independent joint probability mass function respectively. Their differences are more clearly visualized in Figure 8. First, the maximum difference of the mass functions is close to .01, which is quite small. This reconciles with the observation in Figure 5 that the deviations of the models are significantly confined due to known marginal information. Next, in terms of the shape of the differences, the worst-case distribution falls below the baseline in a region centered around $(70, 70)$. This seems to be the region where the baseline probability mass function peaks. The worst-case distribution spreads out this peak and allocates more masses to other regions to achieve a higher objective function. The positive difference seems clustered around two regions and is interpreted as a consequence of the spreading period Note that the integration of the differences must be zero, so both positive and negative differences must be present.

5 Distributionally Robust Rare-Event Analysis

Consider $h(X) = I(X \in B)$ for some set B, where we recall $I(\cdot)$ as the indicator function. We are interested in the probability $P_{true}(X \in B) = E_{true}[I(X \in B)]$, which is known to be small. We

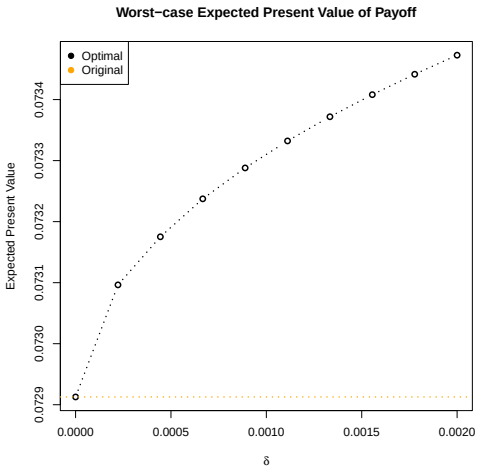

Figure 5: Robust estimate against δ

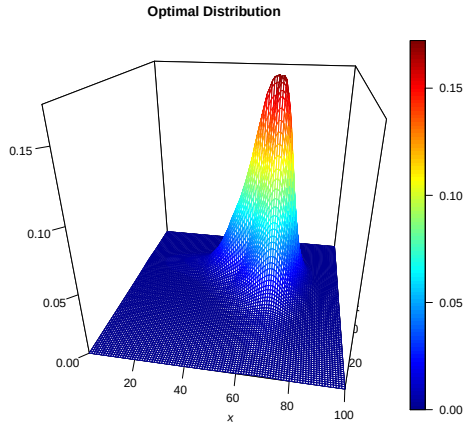

Figure 6: Optimal probability mass function at $\delta = 0.013$

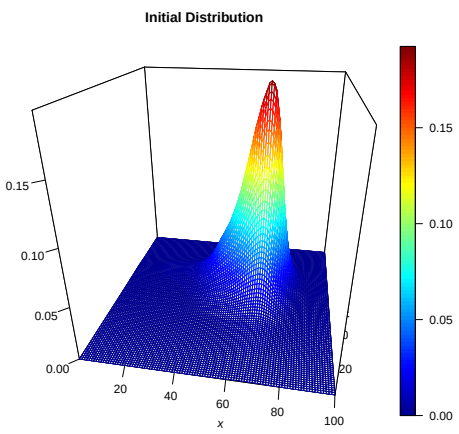

Figure 7: Baseline probability mass function

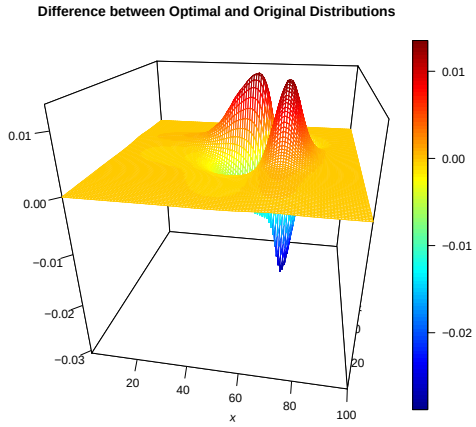

Figure 8: Difference between optimal and baseline probability mass functions

wish to solve the BDR maximization problem

$$\begin{aligned} &\max \ P\left(X \in B\right) \\ &s.t. \ \ D\left(P||P_0\right) \leq \delta, \end{aligned} \tag{19}$$

where a baseline distribution P_0 of X has been suitably calibrated. Since we are in a rare-event setting, it is reasonable to assume that $\log\left(1/P_0\left(X \in B\right)\right) > \delta$. Therefore, applying the general

solution form of the BDR maximization problem in Section 3.3, a worst-case distribution is given by $P_+(\cdot)$ defined such that for each set A,

$$\begin{aligned} P_+(X \in A) &= \frac{E_0(\exp(\theta_+ I(X \in B))\, I(X \in A))}{E_0(\exp(\theta_+ I(X \in B)))} \\ &= \frac{\exp(\theta_+) P_0(X \in A, X \in B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} + \frac{P_0(X \in A, X \notin B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)}, \end{aligned} \tag{20}$$

for some $\theta_+ \in (0, \infty)$. To obtain a clearer interpretation of this worst-case distribution, let us write

$$\begin{aligned} P_+(X \in A) &= \frac{(\exp(\theta_+) - 1) P_0(X \in A, X \in B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} + \frac{P_0(X \in A)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} \\ &= \alpha(\theta_+) P_0(X \in A | X \in B) + (1 - \alpha(\theta_+)) P_0(X \in A), \end{aligned}$$

where

$$\alpha(\theta_+) = \frac{(\exp(\theta_+) - 1) P_0(X \in B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} > 0$$

and $\theta_+ > 0$ satisfies

$$\theta_+ \frac{\exp(\theta_+) P_0(X \in B)}{(\exp(\theta_+) - 1) P_0(X \in B) + 1} - \log((\exp(\theta_+) - 1) P_0(X \in B) + 1) = \delta. \tag{21}$$

Consequently, in simple words, in the context of distributionally robust performance analysis of rare-event probabilities, the worst-case measure is a (specific) mixture between the conditional distribution of X given $X \in B$ and the unconditional distribution, under the baseline measure (regardless of how it was calibrated). This allows the actuary to obtain bounds on rare-event probabilities with the BDR formulation easily using only the baseline measure.

Moreover, note that the worst scenario value is given by

$$P_+(X \in B) = \frac{\exp(\theta_+) P_0(X \in B)}{\exp(\theta_+) P_0(X \in B) + P_0(X \notin B)} = \frac{\exp(\theta_+) P_0(X \in B)}{1 + (\exp(\theta_+) - 1) P_0(X \in B)}. \tag{22}$$

Now, we investigate the asymptotic behavior of the worst scenario value (22) when we fix δ and let $P_0(X \in B) \to 0$. This asymptotics at a fixed model uncertainty level δ while the probability of interest is close to 0 is relevant from an applied standpoint in rare-event analysis. It models the

situation in which we have limited data, or postulate a specific imperfect parametric model P_0, yet we want to estimate probabilities that are very small.

To perform the asymptotic analysis on $P_+(X \in B)$, we first need to understand the behavior of θ_+ that satisfies (21). We carry out a heuristic development that follows and can be rigorized by the argument in Theorem 1 in [10]. Suppose $\delta > 0$ is fixed and $P_0(X \in B) \to 0$. Let us postulate that $\theta_+ \exp(\theta_+) P_0(X \in B) = \eta$ for some $\eta > 0$ that remains bounded away from zero and infinity as $P_0(X \in B) \to 0$. This implies that $\theta_+ \to \infty$ as $P_0(X \in B) \to 0$, and therefore we have $\exp(\theta_+) P_0(X \in B) \to 0$ as $P_0(X \in B) \to 0$. Consequently, $(\exp(\theta_+) - 1)P_0(X \in B) \to 0$ and the left hand side of (21) is asymptotically equivalent to $\theta_+ \exp(\theta_+) P_0(X \in B)$, so that $\theta_+ \exp(\theta_+) P_0(X \in B) \approx \delta$ as $P_0(X \in B) \to 0$. This implies $\theta_+ \approx \log(1/P_0(X \in B))$. In turn, from (22), we get that

$$P_+(X \in B) \approx \frac{\delta}{\theta_+} \approx \frac{\delta}{\log(1/P_0(X \in B))}.$$

To understand decision-making implications, suppose that X models the potential losses of an insurance company, and the baseline loss probability $P_0(X > t) = \exp(-\Lambda_0(t))$ for some function $\Lambda_0(\cdot)$, i.e., under P_0, X has a cumulative hazard function $\Lambda_0(\cdot)$. Now, say we wish to compute b so that $P_{true}(X > b) \leq .005$. Suppose that $\delta = .1$, but the model P_{true} is not known. Then, based on the worst-case distribution of our robust analysis above, we must choose b so that

$$P_+(X > b) \approx \frac{\delta}{\Lambda_0(b)} = \frac{.1}{\Lambda_0(b)} \leq .005, \tag{23}$$

which implies that $b \approx \Lambda_0^{-1}(20)$. For example, if X is exponentially distributed with unit mean under P_0, so that $\Lambda_0(b) = b$, then this gives $b \approx 20$. In contrast, if one trusts the exponential model fully, then one would choose $\exp(-b) = .005$, which yields $b \approx 5.3$.

Paraphrasing, if b is the value of the statutory solvency capital needed to withstand losses with probability at least .995, and if the actuary uses an exponential model with unit mean, then a deviation of .1 units measured in KL divergence between the assumed (exponential) model and the unknown reality might underestimate the statutory capital by a factor of about $20/5.3 \approx 3.7$.

Another important message is that the BDR formulation might provide very conservative estimates for rare-event probabilities when δ (i.e., the size of the uncertainty) is large relative to the

rare-event probability of interest. Suppose again that X is exponentially distributed with unit mean under P_0, then (23) indicates that $P_+(X > b) \approx .1/b$, giving a Pareto-type tail behavior under P_+. In Section 9, we will discuss other types of distances that can give less conservative estimates for rare-event probabilities.

5.1 Numerical Example for Rare-Event Probability

Consider calculating the loss probability $P_{true}(L > b)$ for a given reserve level b, where the loss model is of the simple form $L = X_1 + \cdots + X_n$. Under the baseline model P_0, X_i's follow a multivariate Gaussian distribution, i.e., $(X_1, \ldots, X_d) \sim N(\boldsymbol{\mu}, \Sigma)$ for some mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, with specific distributional parameters given momentarily. Consequently,

$$P_0(L > b) = \overline{\Phi}\left(\frac{b - \mathbf{1}'\boldsymbol{\mu}}{\sqrt{\mathbf{1}'\Sigma\mathbf{1}}}\right),$$

where $\mathbf{1}$ is a vector of entries equal to one, $'$ denotes transpose, and $\overline{\Phi}(\cdot)$ is the tail distribution function of a standard Gaussian random variable.

Now suppose that the actuary is uncertain about whether $(X_1, ..., X_d)$ follows a Gaussian model. So we formulate the corresponding BDR problem (19) with the objective function being $P(L > b)$. We first check whether $\log(1/P_0(L > b)) \leq \delta$, analogous to Case 1 in Section 3.1. If this is the case, then the worst-case density of L is

$$f_+(x) = \frac{\phi\left((x - \mathbf{1}'\boldsymbol{\mu})/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right)}{P_0(L > b)\sqrt{\mathbf{1}'\Sigma\mathbf{1}}} I(x > b), \tag{24}$$

where $\phi(\cdot)$ is the density of a standard Gaussian random variable, and $P_+(L > b) = 1$. To see this, note that $h(l) = I(l > b)$, and (24) is precisely the conditional distribution of L given that $L > b$ under the model P_0, which is the solution depicted in Case 1 in Section 3.1.

Otherwise, if $\log(1/P_0(L > b)) > \delta$, then the worst-case density is given by

$$f_+(x) = \begin{cases} \frac{\exp(\theta_+)\phi\left((x-\mathbf{1}'\boldsymbol{\mu})/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}}{\exp(\theta_+)P_0(L>b)+P_0(L\leq b)} & \text{for} \quad x > b \\ \frac{\phi\left((x-\mathbf{1}'\boldsymbol{\mu})/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}}{\exp(\theta_+)P_0(L>b)+P_0(L\leq b)} & \text{for} \quad x \leq b \end{cases}, \tag{25}$$

where $\theta_+ > 0$ is the root of the equation

$$\theta_+ \frac{\exp(\theta_+) P_0(L > b)}{\exp(\theta_+) P_0(L > b) + P_0(L \leq b)} - \log(\exp(\theta_+) P_0(L > b) + P_0(L \leq b)) = \delta.$$

These are obtained from (20) and (21), or analogously from Case 2 in Section 3.1. Alternately, using the interpretation of the associated exponential tilting discussed in Section 3.1, we have

$$\begin{aligned} f_+(x) &\propto \phi\left(\left(x - \mathbf{1}'\boldsymbol{\mu}\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right) \exp(\theta_+ h(x)) \\ &\propto \phi\left(\left(x - \mathbf{1}'\boldsymbol{\mu}\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right) \exp(\theta_+) I(x > b) \\ &\quad + \phi\left(\left(x - \mathbf{1}'\boldsymbol{\mu}\right)/\sqrt{\mathbf{1}'\Sigma\mathbf{1}}\right) I(x \leq b), \end{aligned}$$

which coincides with (25). Observe that the worst-case density is a weighted version of the original one, thus largely preserving the shape in the corresponding regions. It assigns more (uniform) weight to the region where $x > b$ in a way that is as large as possible but preserves the validity of the KL and the integrating-to-one constraints.

We illustrate the above numerically for $n = 5$, with means of X_i's being $\mu_1 = 1.92, \mu_2 = 1.42, \mu_3 = 1.13, \mu_4 = 1.80, \mu_5 = 1.54$ (five numbers generated uniformly between 1 and 2) and covariance matrix

$$\Sigma = \begin{bmatrix} 1.11 & -0.21 & -0.42 & -0.72 & 1.30 \\ -0.21 & 4.82 & 0.89 & -0.31 & -1.50 \\ -0.42 & 0.89 & 1.05 & 0.61 & -0.52 \\ -0.72 & -0.31 & 0.61 & 2.58 & -0.45 \\ 1.30 & -1.50 & -0.52 & -0.45 & 1.91 \end{bmatrix},$$

which is randomly generated from a Wishart distribution with 5 degrees of freedom and an identity scale matrix. Setting $b = 10$, the probability $P_0(L > b)$ is now given by 0.23. Figure 9 shows the worst-case density under $\delta = 0.1$ and the baseline density. Note that the worst-case density puts more mass on the right side of b, and less on its left side, in order to boost the likelihood of a big loss. In doing so, there is spike occurring exactly at b. This spike occurs because the adjustment leading to the worst-case density attempts to maintain the shape of distribution as much as possible, so

that in each of the regions below and above b the density is multiplied by a constant weight. Finally, Figure 10 shows, much like Figure 1, a concave growth pattern for the worst-case probability as δ increases.

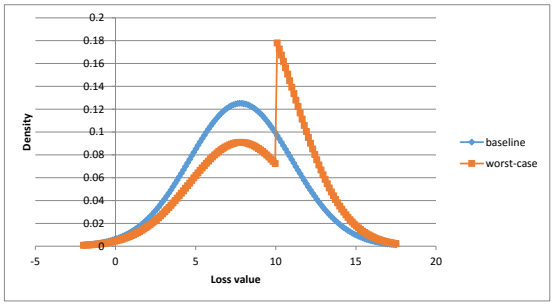

Figure 9: Approximate worst-case density under $\delta = 0.1$

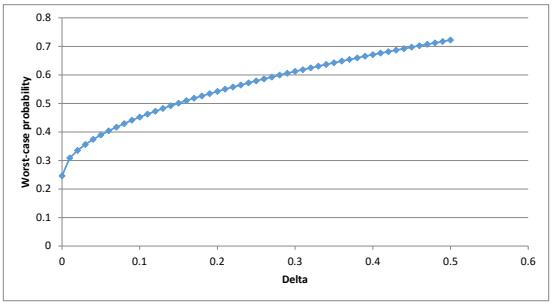

Figure 10: Worst-case probability against δ

6 The Feasible Region in the BDR Formulation

This section discusses the selection of the parameter $\delta > 0$. We present two main approaches: One is estimation using historical data. Another is to understand the choice of δ in terms of systematic stress testing. In the second approach, we will also establish the connection with conventional stress testing approaches and parametric sensitivity analysis.

6.1 Data-Driven Estimation

For simplicity, we assume that the true model has a density $f_{true}(\cdot)$, and the baseline model has a density $f_0(\cdot)$. The most direct approach is to estimate

$$D(P_{true}||P_0) = \int \log\left(\frac{f_{true}(x)}{f_0(x)}\right) f_{true}(x)\, dx. \tag{26}$$

Some proposed procedures that guarantee convergence can be found in, e.g., [51, 55]. The problem with this approach, however, is that the rate of convergence of the estimator depends on

the dimension of the underlying density. Intuitively, this is because this approach indirectly needs to estimate density non-parametrically, which has a similar dependence on the dimension. An alternate approach that is less conservative is based on empirical likelihood that we discuss next.

To explain, first suppose that we observe $X_1, ..., X_n$ as an independent and identically distributed (i.i.d.) sample from X and form the empirical probability mass function

$$\mu_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x),$$

which is uniform on the set $\{X_1, ..., X_n\}$. Now, for any given set of weights $w = (w_1, ..., w_n)$ such that $w_i \geq 0$ and $\sum_{i=1}^{n} w_i = 1$, define

$$v_n(x, w) = \sum_{i=1}^{n} w_i I(X_i = x).$$

In simple words, $v_n(\cdot, w)$ is a probability mass function that assigns probability w_i to the value X_i. Then consider the so-called profile log-likelihood function [53]

$$\mathcal{R}_n(\gamma) = \min\left\{ D(v_n(\cdot, w) \,||\mu_n(\cdot)) : E_{v_n(\cdot,w)}(h(X)) = \sum_{x\in\{X_1,...,X_n\}} v_n(x, w)\, h(x) = \sum_{i=1}^{n} w_i h(X_i) = \gamma \right\}.$$

It turns out, under the null hypothesis that $E_{true}(h(X)) = \gamma$ and other mild conditions (including $Var_{true}(h(X)) < \infty$), that we have

$$2n\mathcal{R}_n(\gamma) \Rightarrow \chi^2_1.$$

That is, under the null hypothesis, $2n\mathcal{R}_n(\gamma)$ follows approximately a chi-square distribution with one degree of freedom. This result is classical in the theory of empirical likelihood [53]. To make our discussion as self-contained as possible, we provide formal derivation of the result in the appendix.

The connection with the BDR problem formulation can be established as follows. Assume P_0 is built from the empirical distribution function of the observed data; that is, we use $\mu_n(\cdot)$ as the distribution of X under the model P_0. If we knew the specific value $\gamma = E_{true}(h(X))$, it would make sense to choose $\delta \geq \mathcal{R}_n(\gamma)$ to solve the BDR problem, because the worst-case distribution,

P_+^n, of the corresponding empirical BDR problem, namely

$$\begin{aligned} \max \quad & \sum_{x \in \{X_1,\ldots,X_n\}} v_n(x)\, h(x) \\ \text{s.t.} \quad & D\left(v_n(\cdot, w) \,||\mu_n(\cdot)\right) \leq \delta \,, \end{aligned} \tag{27}$$

will yield a worst scenario value, $E_+\left(h\left(X\right)\right) = \sum_{k=1}^n w_k^+ h\left(X_k\right) \geq \gamma$. So, as $n \to \infty$, the BDR formulation will result in a correct upper bound for $E_{true}\left(h\left(X\right)\right)$.

Consequently, if one chooses P_0 based on the empirical distribution, it is reasonable to select δ so that

$$P\left(\delta > \mathcal{R}_n\left(\gamma\right)\right) \approx P\left(2n\delta > \chi_1^2\right) = .95, \tag{28}$$

thus providing a 95% confidence upper bound for γ. In other words, we can set $\delta = \chi^2_{1,.95}/(2n)$, where $\chi^2_{1,.95}$ is the 95-percentile of χ_1^2. This approach hinges on the availability of a sufficiently large sample size (in the same order as invoking the central limit theorem).

6.2 A Systematic Approach to Stress Testing

In the absence of enough data to perform a non-parametric calibration of P_0 and δ as suggested in the previous subsection, we suggest using the BDR formulation as an approach to perform systematic stress testing. The idea is to understand how the BDR optimization performs as we vary δ.

To motivate, consider the example of evaluating the expected shortfall or the conditional value-at-risk (C-VaR) of a given portfolio of risk exposures. C-VaR is the conditional expected size of the deficit faced by the company, given the unlikely event of insufficiency of statutory solvency capital. According to Solvency II, the capital should be sufficient to withstand losses with probability not lower than .995 during a one-year time horizon [1]. To stress test C-VaR, it is customary to apply arbitrary shocks into the system (i.e., stressing the system by assuming that extreme events occur). However, such shocks are typically selected in an arbitrary way. Moreover, in the presence of multiple shocks affecting different risk factors, it might be difficult to argue that a particular combination of shocks is more reasonable than an alternative combination of shocks.

In contrast, the approach that we study here can be used to stress test more systematically. Suppose that an insurance company has a given baseline model, P_0, which has been calibrated using a combination of past observations and expert knowledge. The model P_0 is used to compute a given risk measure (or performance measure), say, C-VaR. Periodically, either by internal procedures or due to regulatory constraints, the company is requested to perform stress testing. The result of these stress tests is the determination of capital requirement, which might typically be higher than the C-VaR obtained under P_0.

Let us assume that such a stress-testing procedure has occurred multiple times in the company—say, n times—using the customary approach described before. So the company has built certain experience of the increase in the capital requirement relative to the C-VaR determined by the stress-testing procedure. Based on this experience, one could calibrate values $\delta_1, ..., \delta_n$ so that the solution to the BDR formulation matches the increased capital requirement determined during the n sessions of stress testing. Ultimately, by choosing a single δ appropriately to embed all these experiences, the BDR formulation automatically takes care of inducing the shocks that can potentially cause the highest damage.

To demonstrate in a simple setting how the above could work, suppose that the shocks used in stress-testing before are introduced by changing the parameters in P_0, which is assumed to have a density lying in a parametric family, say f_{θ_0}, where $\theta_0 = (\theta_0^1, \ldots, \theta_0^m)$ denotes a set of m baseline parameters. In the past stress-testing, the company changes the value of some θ_0^j to θ_1^j. Now, note that the KL divergence between f_θ and f_{θ_0}, for any θ within a neighborhood of θ_0 in the parametric space, is

$$\int \log \frac{f_\theta(x)}{f_{\theta_0}(x)} f_\theta(x) dx = E_\theta(\log f_\theta(X) - \log f_{\theta_0}(X))$$

where $E_\theta(\cdot)$ denotes the expectation under f_θ. By a Taylor series expansion, the above is approximately (see, e.g., [3, 4])

$$\begin{aligned} & E_\theta\left(\log f_\theta(X) - \left(\log f_\theta(X) + \nabla \log f_\theta(X)(\theta_0 - \theta) + \frac{1}{2}(\theta_0 - \theta)' \nabla^2 \log f_\theta(X)(\theta_0 - \theta)\right)\right) \\ = \; & \frac{1}{2}(\theta_0 - \theta)' \mathcal{I}(\theta_0 - \theta), \end{aligned}$$

where $\nabla \log f_\theta$ and $\nabla^2 \log f_\theta$ denote the gradient and Hessian of $\log f_\theta$, and $\mathcal{I} = -E_\theta(\nabla^2 \log f_\theta(X))$ is the Fisher information matrix. The last equality follows from the fact that the expected value of the so-called score function $\nabla \log f_\theta$, i.e, $E_\theta(\nabla \log f_\theta(X))$, is 0. With this expression, suppose that, in a simple setting, the company in a past period changes θ_0^j to θ_1^j with $\theta_1^j - \theta_0^j = d^j$. This means that by selecting $\delta = (1/2)\mathcal{I}_j(d^j)^2$, where $\mathcal{I}_j$ is the j-th diagonal entry of $\mathcal{I}$, the worst scenario value of our BDR formulation gives a valid bound on the changed performance measure of interest from such a parametric stress test. Suppose that this has been done many times, so that $d^j, j = 1, \ldots, m$, now denote the maximum perturbation difference used for each of the m parameters. Then choosing $\delta = \max_{j=1,\ldots,m}(1/2)\mathcal{I}_j(d^j)^2$ will give a robust stress test result that incorporates all past experience.

7 Simulation-Based Solution Procedure

The optimization problems that we consider in the previous sections may sometimes be difficult to solve, in the sense that an optimal model P_+ is not expressible in closed form, especially in the context of general continuous distributions. For example, X may comprise a complicated function or the convolution of many elementary variables. In this situation, one can resort to Monte Carlo simulation to approximate the optimization problem.

This approach is as follows. Assume that P_0 is not analytically tractable or that we do not have access to a closed-form expression of a density of X under P_0, but we can simulate i.i.d. sample $X_1, ..., X_n$ of X from P_0. Say we are interested in using BDR formulation (1) to bound $E_{true}\left(h\left(X\right)\right)$. Using the re-expression (15), an empirical version of (15) takes the form given in (27), which we write explicitly here as

$$\begin{aligned} &\max \sum_{i=1}^{n} h\left(X_i\right) w_i \\ &\text{s.t.} \ \sum_{i=1}^{n} w_i \log\left(n w_i\right) \leq \delta \ , \quad \sum_{i=1}^{n} w_i = 1, \ w_i \geq 0 \text{ for all } 1 \leq i \leq n. \end{aligned} \tag{29}$$

We then proceed to solve (29) using the tools for the discrete mass case in Section 3.1. Formulation (29) can be viewed as a so-called sample average approximation (SAA) (see [61, 39]), whose general

idea is to replace expected value with empirical average in an optimization problem. Formulation (29) bears some differences with conventional SAA since its decision variable also changes with the Monte Carlo sample (depending on where the support of the sample is). [30, 31] use this approach to quantify model risk in some financial contexts.

7.1 Simulation-Based BDR Formulation: *t*-Copula Baseline Model

We illustrate the above simulation-based approach. Connecting with the example in Section 5.1, suppose now that the loss is given by $L = X_1 + \cdots + X_d$, where $(X_1, \ldots, X_d)$ has Gaussian marginal distributions $X_i \sim N(\mu_i, \sigma_i^2)$ and the dependency is modeled by a t-copula. A t-copula is a multivariate distribution denoted by

$$C^t_{\nu,\varrho}(u_1, \ldots, u_d) = \mathbf{t}_{\nu,\varrho}\left(t_\nu^{-1}(u_1), \ldots, t_\nu^{-1}(u_d)\right), \qquad (u_1, \ldots, u_d) \in (0,1)^d,$$

where $\mathbf{t}_{\nu,\varrho}$ is the joint distribution function of a d-dimensional t-distribution with degree of freedom ν, mean $\mathbf{0}$, and dispersion (or scale) matrix $\varrho = (\rho_{ij})$ which is positive definite, and t_ν^{-1} is the quantile function of a standard univariate t distribution with degree of freedom ν.

Then the distribution function of $(X_1, \ldots, X_d)$ is

$$f(x_1, \ldots, x_d) = C^t_{\nu,\varrho}(\Phi_{\mu_1,\sigma_1^2}(x_1), \ldots, \Phi_{\mu_d,\sigma_d^2}(x_d)),$$

where $\Phi_{\mu,\sigma^2}(\cdot)$ denotes the distribution function of $N(\mu, \sigma^2)$.

It is difficult to evaluate $P(L > b)$ in closed form, which motivates the use of Monte Carlo simulation. An unbiased estimate of $P_0(L > b)$ can be obtained by outputting $L = \Phi^{-1}_{\mu_1,\sigma_1^2}(t_\nu(Z_1)) + \cdots + \Phi^{-1}_{\mu_d,\sigma_d^2}(t_\nu(Z_d))$, where $(Z_1, \ldots, Z_d)$ is drawn from the multivariate t distribution $\mathbf{t}_{\nu,\varrho}$. Repeat this n times; say we get $L_1, \ldots, L_n$.

To solve the empirical BDR formulation, we use the sampled $L_1, \ldots, L_n$ and combine with our results in Section 5 as follows. We first check whether $\log(n/|\{j : L_j > b\}|) \le \delta$, where $|\{j : L_j > b\}|$

is the cardinality of the set $\{j : L_j > b\}$. If this is the case, we let

$$w_i^+ = \begin{cases} \frac{1}{|\{j:L_j>b\}|} & \text{for } i \text{ such that } L_i > b \\ 0 & \text{for } i \text{ such that } L_i \leq b \end{cases}.$$

This is the approximate worst-case distribution for L, which gives a corresponding approximate worst scenario value 1.

Otherwise, if $\log(n/|\{j : L_j > b\}|) > \delta$, then output

$$w_i = \begin{cases} \frac{\exp(\theta_+)}{\exp(\theta_+)|\{i:L_i>b\}|+|\{i:L_i\leq b\}|} & \text{for } i \text{ such that } L_i > b \\ \frac{1}{\exp(\theta_+)|\{i:L_i>b\}|+|\{i:L_i\leq b\}|} & \text{for } i \text{ such that } L_i \leq b, \end{cases}$$

where $\theta_+ > 0$ satisfies

$$\frac{\theta_+ \exp(\theta_+)|\{i : L_i > b\}|}{\exp(\theta_+)|\{i : L_i > b\}| + |\{i : L_i \leq b\}|} - \log\left(\frac{1}{n}(\exp(\theta_+)|\{i : L_i > b\}| + |\{i : L_i \leq b\}|)\right) = \delta.$$

The probability weights $(w_i)_{i=1,\ldots,n}$ on $(L_i)_{i=1,\ldots,n}$ form an approximation for a worst-case distribution for L, and

$$\frac{\exp(\theta_+)|\{i : L_i > b\}|}{\exp(\theta_+)|\{i : L_i > b\}| + |\{i : L_i \leq b\}|}$$

is the approximate worst scenario value $P_+(L > b)$.

To illustrate numerically, we consider $d = 5$ and $(X_1, \ldots, X_5)$ each following a Gaussian marginal distribution with $\mu_1 = 2.20, \mu_2 = 2.73, \mu_3 = 2.73, \mu_4 = 2.42, \mu_5 = 2.27$ (five numbers generated uniformly between 2 and 3) and $\sigma_1 = 0.92, \sigma_2 = 0.39, \sigma_3 = 0.11, \sigma_4 = 0.56, \sigma_5 = 0.33$ (five numbers generated uniformly between 0 and 1), respectively. We use a t-copula with a degree of freedom 10 and the following dispersion matrix:

$$\Sigma = \begin{bmatrix} 8.19 & -0.92 & -3.54 & 3.35 & -3.96 \\ -0.92 & 6.09 & 1.07 & 1.37 & -0.08 \\ -3.54 & 1.07 & 7.10 & 1.35 & -1.70 \\ 3.35 & 1.37 & 1.35 & 3.14 & -3.73 \\ -3.96 & -0.08 & -1.70 & -3.73 & 8.73 \end{bmatrix},$$

which is generated from a Wishart distribution with 5 degrees of freedom and an identity scale matrix. We use $n = 1,000$ to generate the baseline sample from the t-copula model. Setting $b = 10$, the estimated loss probability is 0.232 with 95% confidence interval $[0.206, 0.258]$. The histograms of the optimally weighted sample at $\delta = 0.1$ and the baseline sample are plotted in Figures 11 and 12. As can be seen, more weights are put on the right side of b, in a uniform manner, to boost the large loss probability. We also plot the worst-case probability against δ in Figure 13, which, similar to the example in Section 5.1, shows a concavely increasing pattern.

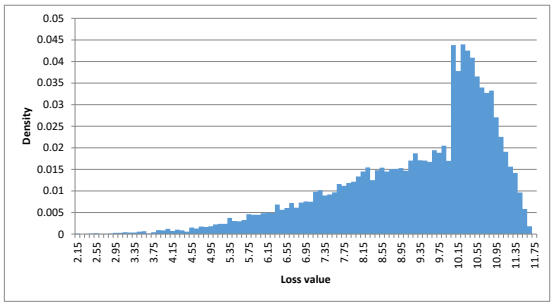

Figure 11: Approximate worst-case density under $\delta = 0.1$

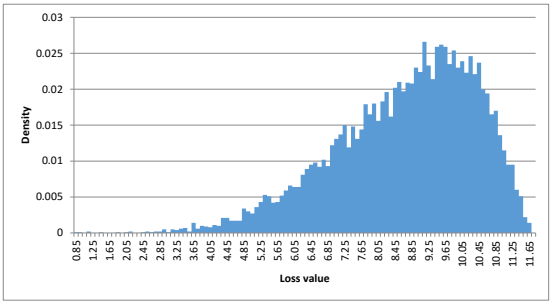

Figure 12: Approximate original density

8 Robust Conditional Value at Risk

This section demonstrates how our distributionally robust analysis can be further developed to handle risk-analytic problems involving optimization. For this discussion, we focus on the problem of computing, say, the 95% C-VaR of a random variable L that represents potential losses in a given year, i.e.,

$$\text{C-VaR}_{true}(\alpha) = E_{true}(L - b | L > b),$$

where $P_{true}(L > b) = 1 - \alpha$ with $\alpha = .95$.

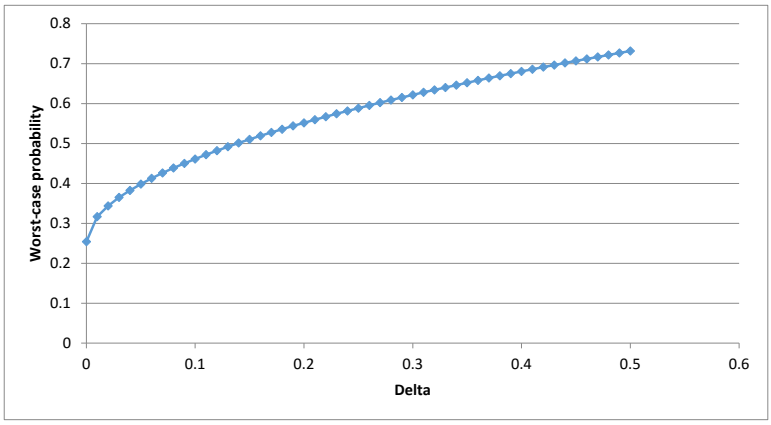

Figure 13: Worst-case probability against δ

It is well known [59] that if L has a continuous distribution under P_{true}, then

$$\text{C-VaR}_{true}(\alpha) = \min_{\theta} \left(\theta + \frac{E_{true}(\max(L-\theta, 0))}{1-\alpha} \right).$$

To robustify this calculation, we extend the BDR formulation discussed earlier in a natural way, to solve an upper and a lower bound for $\text{C-VaR}_{true}(\alpha)$ in the form

$$\min / \max_{D(P||P_0) \leq \delta} \min_{\theta} \left(\theta + \frac{E(\max(L-\theta, 0))}{1-\alpha} \right), \tag{30}$$

where the outer optimization problems are taken over probability models P.

Problem (30) is challenging to solve in closed form, so it is natural to consider SAA along the discussion in Section 7. To approximate

$$\text{C-VaR}_0(\alpha) = \min_{\theta} \left(\theta + \frac{E_0(\max(L-\theta, 0))}{1-\alpha} \right)$$

for a given baseline model P_0, we first simulate i.i.d. copies $L_1, ..., L_n$ under P_0 and then solve

$$\widehat{\text{C-VaR}}_0(\alpha, n) = \min_{\theta} \left(\theta + \frac{1}{n} \sum_{i=1}^{n} \frac{\max(L_i - \theta, 0)}{1-\alpha} \right). \tag{31}$$

One property of this estimate is, from the theory of SAA, that under mild assumptions (for instance, if L has a continuous density under P_0),

$$\widehat{\text{C-VaR}}_0(\alpha, n) \approx \text{C-VaR}_0(\alpha) + \frac{\widehat{\sigma}(\theta_0^*(n))}{n^{1/2}} Z, \tag{32}$$

where Z is a standard Gaussian random variable,

$$\widehat{\sigma}^2(\theta_0^*(n)) = \frac{1}{n-1} \sum_{j=1}^{n} \left(\theta_0^*(n) + \frac{\max(L_i - \theta_0^*(n), 0)}{1-\alpha} - \widehat{\text{C-VaR}}_0(\alpha, n) \right)^2,$$

and $\theta_0^*(n)$ is the solution obtained by solving (31). In simple words, $\widehat{\sigma}^2(\theta_0^*(n))$ is the empirical estimator of the variance of the objective function evaluated at the estimated optimal solution in (31).

Approximating (30) involves adding an outer optimization problem. That is, we still keep $L_1, ..., L_n$ simulated under P_0, but we consider

$$\begin{aligned} &\min / \max_{w_1, ... w_n} \min_{\theta} \left(\theta + \sum_{i=1}^{n} w_i \frac{\max(L_i - \theta, 0)}{1-\alpha} \right) \\ &s.t. \\ &\sum_{i=1}^{n} w_i \log(n w_i) \leq \delta \\ &\sum_{i=1}^{n} w_i = 1, \text{ and } w_i \geq 0 \text{ for all } 1 \leq i \leq n. \end{aligned} \tag{33}$$

The above discussion is similar to the BDR settings discussed before. However, there are also some differences. One of these is that calibrating a suitable δ from the non-parametric, data-driven view in Section 6.1 needs some modification in the C-VaR setting. It turns out that, with i.i.d. data of size n, δ should be selected so that $P\left(2n\delta > \chi_2^2\right) = 1 - \beta$, if we want the optimization output to bound the true C-VaR with $(1-\beta)$ confidence. Note that the degree of freedom of the chi-square

distribution has increased from one to two compared with the discussion in Section 6. The reason for this change is the appearance of the inner minimization in problem (33), which introduces another equality constraint in the empirical likelihood derivation outlined in the Appendix, in order to capture the requirement that the derivative of the objective function with respect to θ should vanish. This derivation is similar to that given in Section 6, and further details can be found in [45, 46]. Under additional assumptions, the calibration can be further tightened; see, e.g., [25].

Another point to note is that while the outer maximization formulation in (33) is a convex program, the minimization one is not. Heuristic procedures, such as multi-start or alternating minimization should be employed in the later case.

8.1 Numerical Example for C-VaR Estimation

We consider the problem of estimating C-VaR in which we assume that the loss L follows a standard normal distribution under P_0. We set $\alpha = 0.9$ and generate $n = 1,000$ observations. Figure 14 shows the upper and lower robust bounds from (33) against different values of δ, showing how the width of the robust interval widens at a decreasing rate as δ increases.

Moreover, we also test the performance of the 95% (i.e., $\beta = 0.05$) confidence bounds using (33), by selecting δ such that $P(2n\delta > \chi^2_2) = 1 - \beta$. We carry out the cases $n = 50$ and $n = 100$. Table 1 reports the point estimate of the coverage probability, mean lower and upper bounds, and the mean and standard deviation of the interval width for empirical likelihood, while Table 2 shows the results using the classical SAA theory via (32). These quantities are obtained from repeating the experiments multiple times, each time generating a new data set. The coverage probability, for instance, outputs the proportion of repetitions in which the constructed interval covers the truth. We see that empirical likelihood gives a higher and more accurate coverage, though the SAA counterpart gives tighter and less varied interval width, at least for this example we conducted. This demonstrates some benefits of using our BDR formulation with empirical likelihood calibration over classical SAA, in generating intervals that cover the truth more frequently and hence are more robust. On the downside, however, these more robust intervals pay the price of being generally wider. The bottom line is that for a user who is risk-averse, the BDR approach with empirical likelihood calibration should be preferable over SAA.

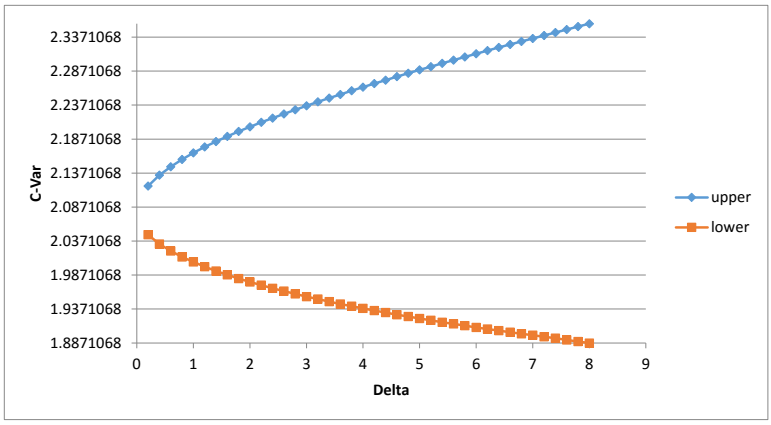

Figure 14: Worst-case C-VaR against δ

n	Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
50	0.90	1.22	2.33	1.11	0.43
100	0.94	1.32	2.26	0.94	0.26

Table 1: Statistical performances of empirical likelihood for different sample sizes

n	Coverage probability	Mean lower bound	Mean upper bound	Mean interval width	Standard deviation of interval width
50	0.86	1.21	2.26	1.05	0.47
100	0.84	1.34	2.05	0.71	0.21

Table 2: Statistical performances of standard confidence interval of SAA for different sample sizes

9 Additional Considerations

We discuss two alternatives that can be used for robust performance analysis. The first one involves the use of moment constraints, and the second one involving different notions of discrepancy.

9.1 Robust Performance Analysis via Moment Constraints

In some situations, it may not be possible to construct an explicit baseline distribution P_0. Alternatively, if information on moments is available, we might consider worst-case optimization under

such information, in the form

$$\begin{aligned} \max \quad & E\left(h(X)\right) \\ \text{s.t.} \quad & E\left(v_i(X)\right) \leq \alpha_i,\ i = 1, \ldots, s \\ & E(v_i(X)) = \alpha_i,\ i = s+1, \ldots, m, \end{aligned} \tag{34}$$

where the maximization is over all probability models P (we focus on maximization here to avoid redundancy). This is a general formulation that has m moment constraints, and $v_i(\cdot)$ can represent any function. For instance, for moment constraints involving means and variances, we can select $v_1(x) = x$ and $v_2(x) = -x$, $v_3(x) = x^2$, $v_4(x) = -x^2$, and $\alpha_1 = \overline{\mu}$, $\alpha_2 = -\underline{\mu}$, $\alpha_3 = \overline{\sigma}$, $\alpha_4 = -\underline{\sigma}$, and all constraints could be inequalities. There is a general procedure for solving these problems which builds on linear programming. The most tricky part involves finding the support of the distribution. Observe that, if the support of the worst-case distribution is known, then problem (34) is just a problem with a linear objective function and linear constraints, and is solvable using standard routines.

Finding the support involves a sequential search. More precisely, the procedure for solving (34) is shown in Algorithm 1 (which borrows from, e.g., [9]).

We discuss Algorithm 1 in the following several aspects:

1. *Interpretation:* The output of the procedure is an exact optimal value of (34). The worst-case probability distribution is a finite-support discrete distribution on $\{x_1, \ldots, x_\tau\}$ with weights $\{p_1^k, \ldots, p_\tau^k\}$ obtained in the last iteration.

2. *Comparison with the BDR formulation:* Unlike the BDR formulation, (34) does not have a baseline input distribution to begin with.

3. *Computational efficiency:* Step 1 in each iteration of Algorithm 1 can be carried out by standard linear programming solver, which can output both the optimal $\{p_j\}$ and the dual multipliers $\{\theta^k, \pi_1^k, \ldots, \pi_m^k\}$. Step 2 is a one-dimensional line search if X is one-dimensional.

4. *Minimization counterpart:* For a minimization problem, simply replace h with $-h$ in the whole procedure of Algorithm 1, except in the last step we output $\sum_{j=1}^{\tau} h(x_j) p_j^k$.

Algorithm 1 Generalized linear programming procedure for solving (34)

Initialization: An arbitrary probability distribution on the support $\{x_1, \ldots, x_L\}$, where $l \leq m+1$, that lies in the feasible region in (34). Set $\tau = l$.

Procedure: For each iteration $k = 1, 2, \ldots$, given $\{x_1, \ldots, x_\tau\}$:

1. Master problem solution: Solve

$$\begin{array}{ll} \max & \sum_{j=1}^{\tau} h(x_j) p_j \\ \text{s.t.} & \sum_{j=1}^{\tau} v_i(x_j) p_j \leq \alpha_i, \ i = 1, \ldots, s \\ & \sum_{j=1}^{\tau} v_i(x_j) p_j = \alpha_i, \ i = s+1, \ldots, m \\ & \sum_{j=1}^{\tau} p_j = 1 \\ & p_j \geq 0, \ j = 1, \ldots, \tau \end{array}$$

Let $\{p_1^k, \ldots, p_\tau^k\}$ be the optimal solution. Find the dual multipliers $\{\theta^k, \pi_1^k, \ldots, \pi_m^k\}$ that satisfy

$$\begin{array}{l} \theta^k + \sum_{i=1}^{m} \pi_i^k v_i(x_j) = h(x_j), \text{ if } p_j > 0, j = 1, \ldots, \tau \\ \theta^k + \sum_{i=1}^{m} \pi_i^k v_i(x_j) \geq h(x_j), \text{ if } p_j = 0, j = 1, \ldots, \tau \\ \pi_i^k \geq 0, i = 1, \ldots, s \end{array}$$

2. Subproblem solution: Find $x_{\tau+1}$ that maximizes

$$\rho(x; \theta^k, \pi_1^k, \ldots, \pi_m^k) = h(x) - \theta^k - \sum_{i=1}^{m} \pi_i^k v_i(x)$$

If $\rho(x_{\tau+1}; \theta^k, \pi_1^k, \ldots, \pi_m^k) > 0$, then let $\tau = \tau + 1$; otherwise, stop the procedure, and $\{x_1, \ldots, x_\tau\}$ are the optimal support points, with $\{p_1^k, \ldots, p_\tau^k\}$ the associated weights.

After the last iteration, output

$$\sum_{j=1}^{\tau} h(x_j) p_j^k$$

9.2 Renyi Divergence as Discrepancy Notion

While we have presented our approach based on the KL divergence or relative entropy, other distance notions can be used. For example, we could consider the BDR formulation using the so-called Renyi divergence of degree $\alpha > 1$, defined via

$$D_\alpha\left(P||P_0\right) = \frac{1}{\alpha - 1}\log E_0\left(\left(\frac{dP}{dP_0}\right)^\alpha\right).$$

As $\alpha \to 1$, we recover the KL divergence or relative entropy, i.e.,

$$D_\alpha\left(P||P_0\right) \to D\left(P||P_0\right) = E\left(\log\left(\frac{dP}{dP_0}\right)\right).$$

The corresponding distributionally robust formulation takes the form

$$\min / \max \;\; E\left(h\left(X\right)\right) \quad \text{s.t.} \;\; D_\alpha\left(P||P_0\right) \leq \delta,$$

and the optimal solution is obtained roughly as follows. First, given $\theta_1, \theta_2 > 0$, define

$$Z_+\left(\theta_1, \theta_2\right) = \max(\theta_1 + \theta_2 h\left(X\right), 0)^{1/(1-\alpha)},$$

with θ_1, θ_2 chosen so that

$$E_0\left(Z_+\left(\theta_1, \theta_2\right)\right) = 1, \quad E_0\left(Z_+\left(\theta_1, \theta_2\right)^\alpha\right) = \exp\left(\delta\left(\alpha - 1\right)\right),$$

which comes from KKT conditions analogous to the one in Section 3.1. Then, a worst-case distribution is defined such that, for any set A,

$$P_+\left(X \in A\right) = E_0\left(Z_+\left(\theta_1, \theta_2\right) I\left(X \in A\right)\right).$$

See, e.g., [14, 5] for related derivations.

The use of Renyi divergence typically leads to less conservative estimates (as we discuss momentarily in the next subsection), but the parameter δ might be more difficult to estimate (e.g., the

empirical likelihood machinery discussed in Section 6.1 may not apply readily). Next, we consider rare-event estimation to develop more intuition and contrast the effect of using Renyi divergence from KL.

9.2.1 Robust Rare-Event Analysis via Renyi Divergence

We consider the case in which $h(X) = I(X \in B)$, so that we have

$$P_+(X \in B) = (\theta_1 + \theta_2)^{1/(1-\alpha)} P_0(X \in B),$$

and

$$\begin{aligned} E_0(Z_+(\theta_1,\theta_2)^\alpha) &= \theta_1^{\alpha/(1-\alpha)} P_0(X \notin B) + (\theta_1+\theta_2)^{\alpha/(1-\alpha)} P_0(X \in B) = \exp(\delta(\alpha-1)), \\ E_0(Z_+(\theta_1,\theta_2)) &= \theta_1^{1/(1-\alpha)} P_0(X \notin B) + (\theta_1+\theta_2)^{1/(1-\alpha)} P_0(X \in B) = 1. \end{aligned}$$

Letting $(\theta_1+\theta_2)^{\alpha/(1-\alpha)} = \eta / P_0(X \in B)$, and substituting in the previous display, we have

$$\begin{aligned} \theta_1^{\alpha/(1-\alpha)}(1 - P_0(X \in B)) + \eta &= \exp(\delta(\alpha-1)), \\ \theta_1^{1/(1-\alpha)}(1 - P_0(X \in B)) + \eta^{1/\alpha} P_0(X \in B)^{1-1/\alpha} &= 1. \end{aligned}$$

As $P_0(X \in B) \to 0$, since $\alpha > 1$, we can select $\theta_1 \approx 1$ and $(\theta_1+\theta_2)^{1/(1-\alpha)} \approx \eta^{1/\alpha} / P_0(X \in B)^{1/\alpha}$, with $\eta \approx \exp(\delta(\alpha-1)) - 1$, concluding that

$$P_+(X \in B) \approx (\exp(\delta(\alpha-1)) - 1)^{1/\alpha} P_0(X \in B)^{1-1/\alpha}.$$

Let us revisit the example discussed at the end of Section 5. Assume that X is exponentially distributed with mean one under P_0. Select $B = [b, \infty)$ and $\alpha = 2$. Then we have

$$P_+(X > b) \approx (\exp(\delta) - 1)^{1/2} \exp(-b/2).$$

In contrast, when using the KL divergence for the case of exponentially distributed X, we obtain a much lower rate of decay (of the form δ/b for b large; see (23)). In this sense, KL divergence

induces a much more cautious robust methodology than Renyi divergence. To illustrate this last point, we compare the BDR formulations using KL and Renyi divergences with $\alpha = 2$. We use $h(X) = I(X > b)$ and an exponential baseline with rate 1, where we set b to be the 99-percentile of this baseline. We solve the BDR formulations using the technique in Section 5 and the discussion above. Figure 15 shows that the bounds obtained from the Renyi divergence are tighter than KL in this small probability estimation context. This is noticeable especially for the upper bounds where the KL divergence gives values more than double in magnitude than Renyi as δ increases.

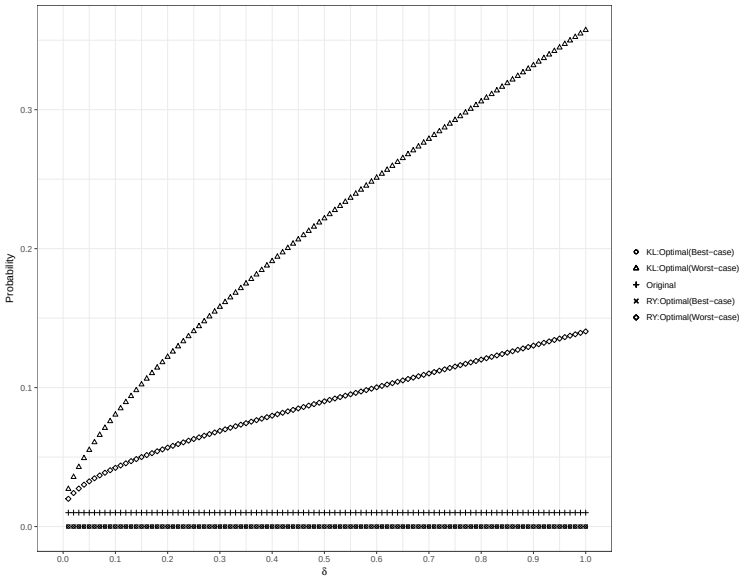

Figure 15: Robust estimates against δ using Renyi and KL divergences

10 Conclusions

We have discussed a systematic approach to quantify potential model errors based on the BDR formulation, which is a convex optimization problem in the space of probability distributions. This approach roots in areas including economics, statistics and operations research, and aims to compute worst-case bounds on the performance or risk measure when the model deviates from a baseline model in the KL sense. We have demonstrated several properties and advantages of this approach, including its non-parametric nature, generality in applying to various situations, and computational tractability.

Building upon the BDR formulation, we have demonstrated how to extend our approach to

applications that are relevant to actuarial practice, including studying the impact of dependence in multivariate models, the impact of model uncertainty in rare-event estimation and in computing risk-analytic measures that involve optimization such as C-VaR. We have also discussed methods to calibrate the feasible region size in BDR formulations, from the viewpoint of empirical likelihood and stress testing. Lastly, we have investigated additional ways to define the model uncertainty region based on Renyi divergence or imposition of moment constraints.

11 Appendix: Approximating Distribution for the Size of the Feasible Region

Similar to the use of the KKT conditions for the solution of the BDR problem formulation, we obtain that the optimal solution satisfies

$$w_i(\theta) = \frac{\exp(\theta h(X_i))}{\sum_{j=1}^n \exp(\theta h(X_j))},$$

where θ satisfies $\sum_{i=1}^n w_i(\theta)(h(X_i) - \gamma) = 0$, namely,

$$\frac{\sum_{i=1}^n \exp(\theta h(X_i))(h(X_i) - \gamma)}{\sum_{j=1}^n \exp(\theta h(X_j))} = 0. \tag{35}$$

Now suppose that indeed $E_{true}(h(X) - \gamma) = 0$, and let us consider $\bar{h}(x) = h(x) - \gamma$. We can rewrite (35)—after multiplying by $\sum_{j=1}^n \exp(\theta h(X_j) - \gamma\theta)/n^{1/2}$) —as

$$\frac{1}{n^{1/2}} \sum_{i=1}^n \exp\left(\theta \bar{h}(X_i)\right) \bar{h}(X_i) = 0.$$

Then let $\theta = \eta/n^{1/2}$ and perform a Taylor expansion to conclude that

$$\frac{1}{n^{1/2}} \sum_{i=1}^n \bar{h}(X_i) \left(1 + \eta \frac{\bar{h}(X_i)}{n^{1/2}} + \cdots\right) = 0. \tag{36}$$

Recall that the central limit theorem states the approximation in distribution

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\bar{h}\left(X_i\right)\approx\sigma Z,$$

where $\sigma^2 = Var_{true}\left(h\left(X\right)\right)$, and Z is standard Gaussian. Therefore, solving for η in (36) and ignoring lower-order error terms, we obtain that

$$\eta\approx-\frac{1}{n^{1/2}}\frac{\sum_{i=1}^{n}\bar{h}\left(X_i\right)}{\frac{1}{n}\sum_{i=1}^{n}\bar{h}\left(X_i\right)^2}\approx\frac{Z}{\sigma}.$$

Now consider $\mathcal{R}_n\left(\gamma\right)$, which can be written as

$$\begin{aligned}
& \mathcal{R}_n\left(\gamma\right) \\
= & \sum_{i=1}^{n} w_i\left(\theta\right)\log\left(\frac{\exp\left(\theta\bar{h}\left(X_i\right)\right)}{n^{-1}\sum_{j=1}^{n}\exp\left(\theta\bar{h}\left(X_j\right)\right)}\right) \\
= & \theta\sum_{i=1}^{n} w_i(\theta)\bar{h}(X_i) - \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\theta\bar{h}(X_i))\right) \\
= & \frac{\theta\frac{1}{n}\sum_{i=1}^{n}\exp(\theta\bar{h}(X_i))\bar{h}(X_i)}{\frac{1}{n}\sum_{i=1}^{n}\exp(\theta\bar{h}(X_i))} - \log\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\theta\bar{h}(X_i))\right) \\
= & \frac{\theta\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)+\frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2+\cdots\right)}{1+\frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i)+\frac{\theta^2}{2n}\sum_{i=1}^{n}\bar{h}(X_i)^2+\cdots} \\
& -\left(\frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i)+\frac{\theta^2}{2n}\sum_{i=1}^{n}\bar{h}(X_i)^2-\frac{\theta^2}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)\right)^2+\cdots\right) \\
\approx & \frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i)+\frac{\theta^2}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2 \\
& -\theta^2\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)\right)^2-\frac{\theta}{n}\sum_{i=1}^{n}\bar{h}(X_i)-\frac{\theta^2}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2-\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)\right)^2\right) \\
= & \frac{\theta^2}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)^2-\left(\frac{1}{n}\sum_{i=1}^{n}\bar{h}(X_i)\right)^2\right) \\
\approx & \frac{\eta^2\sigma^2}{2n} \\
\approx & \frac{Z^2}{2n},
\end{aligned}$$

where the third last step follows by collecting terms up to θ^2. Therefore, we conclude that $2n\mathcal{R}_n(\gamma) \Rightarrow \chi^2_1$, as indicated in our discussion leading to (28).

12 Acknowledgments

(1) This project was supported by the Society of Actuaries (SOA) research grant entitled "Modeling and Analyzing Extreme Risks in Insurance." The authors would like to thank the SOA for its permission to publish some materials excerpted from our project reports [12, 13].
(2) We would like to thank the two referees for their careful reading and helpful feedback.

References

[1] Directive 2009/138/ec of the European Parliament and of the Council of 25 November 2009 on the taking-up and pursuit of the business of Insurance and Reinsurance (Solvency II). *Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:335:0001:0155:en:PDF*.

[2] Updated static mortality tables for defined benefit pension plans for 2016. *Available at https://www.irs.gov/pub/irs-drop/n-15-53.pdf*.

[3] Arampatzis, G., M. A. Katsoulakis, and Y. Pantazis (2015a). Accelerated sensitivity analysis in high-dimensional stochastic reaction networks. *PloS one 10*(7), e0130825.

[4] Arampatzis, G., M. A. Katsoulakis, and Y. Pantazis (2015b). Pathwise sensitivity analysis in transient regimes. In *Stochastic Equations for Complex Systems*, pp. 105–124. Springer.

[5] Atar, R., K. Chowdhary, and P. Dupuis (2015). Robust bounds on risk-sensitive functionals via rényi divergence. *SIAM/ASA Journal on Uncertainty Quantification 3*(1), 18–33.

[6] Ben-Tal, A., D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science 59*(2), 341–357.

[7] Bertsimas, D., V. Gupta, and N. Kallus (2014). Robust SAA. *arXiv preprint arXiv:1408.4445*.

[8] Bertsimas, D. and I. Popescu (2005). Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization 15*(3), 780–804.

[9] Birge, J. R. and J. H. Dulá (1991). Bounding separable recourse functions with limited distribution information. *Annals of Operations Research 30*(1), 277–298.

[10] Blanchet, J., C. Dolan, and H. Lam (2014). Robust rare-event performance analysis with natural non-convex constraints. In *Proceedings of the 2014 Winter Simulation Conference*, pp. 595–603. IEEE Press.

[11] Blanchet, J. and Y. Kang (2016). Sample out-of-sample inference based on wasserstein distance. *arXiv preprint arXiv:1605.01340*.

[12] Blanchet, J., H. Lam, Q. Tang, and Z. Yuan. Applied robust performance analysis for actuarial applications. *Society of Actuaries. Available at https://www.soa.org/research-reports/2016/research-applied-robust-performance/*.

[13] Blanchet, J., H. Lam, Q. Tang, and Z. Yuan. Applied robust performance analysis for actuarial applications: A guide. *Society of Actuaries. Available at https://www.soa.org/research-reports/2016/research-applied-robust-performance/*.

[14] Blanchet, J. and K. R. Murthy (2016a). On distributionally robust extreme value analysis. *arXiv preprint arXiv:1601.06858*.

[15] Blanchet, J. and K. R. Murthy (2016b). Quantifying distributional model risk via optimal transport. *arXiv preprint arXiv:1604.01446*.

[16] Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge University Press.

[17] Breuer, T. and I. Csiszár (2013). Measuring distribution model risk. *Mathematical Finance*.

[18] Cover, T. M. and J. A. Thomas (2012). *Elements of information theory*. John Wiley & Sons.

[19] Cox, D. R. and D. V. Hinkley (1979). *Theoretical statistics*. CRC Press.

[20] Cox, S. H., Y. Lin, R. Tian, and L. F. Zuluaga. Mortality portfolio risk management. *Journal of Risk and Insurance 80*(4), 853–890.

[21] Delage, E. and Y. Ye (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research 58*(3), 595–612.

[22] Deming, W. E. and F. F. Stephan (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics 11*(4), 427–444.

[23] Dey, S. and S. Juneja (2012). Incorporating fat tails in financial models using entropic divergence measures. *arXiv preprint arXiv:1203.0643*.

[24] Dhara, A., B. Das, and K. Natarajan (2017). Worst-case expected shortfall with univariate and bivariate marginals. *arXiv preprint arXiv:1701.04167*.

[25] Duchi, J., P. Glynn, and H. Namkoong (2016). Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*.

[26] Embrechts, P. and G. Puccetti (2006). Bounds for functions of multivariate risks. *Journal of Multivariate Analysis 97*(2), 526–547.

[27] Embrechts, P., G. Puccetti, and L. Rüschendorf (2013). Model uncertainty and var aggregation. *Journal of Banking & Finance 37*(8), 2750–2764.

[28] Esfahani, P. M. and D. Kuhn (2015). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*.

[29] Glasserman, P. and X. Xu (2013). Robust portfolio control with stochastic factor dynamics. *Operations Research 61*(4), 874–893.

[30] Glasserman, P. and X. Xu (2014). Robust risk measurement and model risk. *Quantitative Finance 14*(1), 29–58.

[31] Glasserman, P. and L. Yang (2016). Bounding wrong-way risk in CVA calculation. *Mathematical Finance*, doi:10.1111/mafi.12141.

[32] Goh, J. and M. Sim (2010). Distributionally robust optimization and its tractable approximations. *Operations Research 58*(4-part-1), 902–917.

[33] Gotoh, J.-y., M. J. Kim, and A. Lim (2015). Robust empirical optimization is almost the same as mean-variance optimization.

[34] Hanasusanto, G. A., V. Roitch, D. Kuhn, and W. Wiesemann (2015). Ambiguous joint chance constraints under mean and dispersion information. Technical report.

[35] Hansen, L. P. and T. J. Sargent (2008). *Robustness*. Princeton University Press.

[36] Hu, Z. and L. J. Hong (2013). Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*.

[37] Iyengar, G. N. (2005). Robust dynamic programming. *Mathematics of Operations Research 30*(2), 257–280.

[38] Jiang, R. and Y. Guan (2012). Data-driven chance constrained stochastic program. *Mathematical Programming*, 1–37.

[39] Kleywegt, A. J., A. Shapiro, and T. Homem-de Mello (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization 12*(2), 479–502.

[40] Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

[41] Lam, H. (2016a). Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Forthcoming in Operations Research. Available at arXiv preprint arXiv:1605.09349*.

[42] Lam, H. (2016b). Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research 41*(4), 1248–1275.

[43] Lam, H. (2018). Sensitivity to serial dependency of input processes: A robust approach. *Management Science 64*(3), 1311–1327.

[44] Lam, H. and C. Mottet (2017). Tail analysis without parametric models: A worst-case perspective. *Operations Research 65*(6), 1696–1711.

[45] Lam, H. and E. Zhou (2015). Quantifying uncertainty in sample average approximation. In *Proceedings of the 2015 Winter Simulation Conference*, pp. 3846–3857. IEEE Press.

[46] Lam, H. and E. Zhou (2017). The empirical likelihood approach to quantifying uncertainty in sample average approximation. *Operations Research Letters 45*(4), 301–307.

[47] Li, B., R. Jiang, and J. L. Mathieu (2016). Ambiguous risk constraints with moment and unimodality information. *Available at Optimization Online*.

[48] Love, D. and G. Bayraksan (2015). Phi-divergence constrained ambiguous stochastic programs for data-driven optimization. Technical report, The Ohio State University, Columbus.

[49] Mottet, C. and H. Lam (2017). On optimization over tail distributions. *arXiv preprint arXiv:1711.00573*.

[50] Natarajan, K., D. Pachamanova, and M. Sim (2008). Incorporating asymmetric distributional information in robust value-at-risk optimization. *Management Science 54*(3), 573–585.

[51] Nguyen, X., M. J. Wainwright, and M. I. Jordan (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *Information Theory, IEEE Transactions on 56*(11), 5847–5861.

[52] Nilim, A. and L. El Ghaoui (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research 53*(5), 780–798.

[53] Owen, A. B. (2001). *Empirical likelihood*. CRC press.

[54] Petersen, I. R., M. R. James, and P. Dupuis (2000). Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *Automatic Control, IEEE Transactions on 45*(3), 398–412.

[55] Póczos, B. and J. Schneider (2011). On the estimation of alpha-divergences.

[56] Popescu, I. (2005). A semidefinite programming approach to optimal-moment bounds for convex classes of distributions. *Mathematics of Operations Research 30*(3), 632–657.

[57] Puccetti, G. and L. Rüschendorf (2012). Computation of sharp bounds on the distribution of a function of dependent risks. *Journal of Computational and Applied Mathematics 236*(7), 1833–1840.

[58] Puccetti, G. and L. Rüschendorf (2013). Sharp bounds for sums of dependent risks. *Journal of Applied Probability 50*(01), 42–53.

[59] Rockafellar, R. T. and S. Uryasev (2000). Optimization of conditional value-at-risk. *Journal of Risk 2*, 21–42.

[60] Scarf, H., K. Arrow, and S. Karlin (1958). A min-max solution of an inventory problem. *Studies in the mathematical theory of inventory and production 10*, 201–209.

[61] Shapiro, A., D. Dentcheva, and A. Ruszczynski (2014). *Lectures on stochastic programming: Modeling and Theory*, Volume 16. SIAM.

[62] Smith, J. E. (1995). Generalized Chebychev inequalities: theory and applications in decision analysis. *Operations Research 43*(5), 807–825.

[63] Van Parys, B. P., P. J. Goulart, and M. Morari (2015). Distributionally robust expectation inequalities for structured distributions. *Optimization Online*.

[64] Wang, B. and R. Wang (2011). The complete mixability and convex minimization problems with monotone marginal densities. *Journal of Multivariate Analysis 102*(10), 1344–1360.

[65] Wang, Z., P. W. Glynn, and Y. Ye (2015). Likelihood robust optimization for data-driven problems. *Computational Management Science 13*(2), 241–261.

[66] Wiesemann, W., D. Kuhn, and M. Sim (2014). Distributionally robust convex optimization. *Operations Research 62*(6), 1358–1376.