Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

LEARNING STOCHASTIC MODEL DISCREPANCY

Matthew Plumlee
Henry Lam

Department of Industrial and Operations Engineering
University of Michigan
1205 Beal Ave
Ann Arbor, MI 48109, USA

ABSTRACT

The vast majority of stochastic simulation models are imperfect in that they fail to fully emulate the entirety of real dynamics. Despite this, these imperfect models are still useful in practice, so long as one knows how the model is inexact. This inexactness is measured by a discrepancy between the proposed stochastic model and a true stochastic distribution across multiple values of some decision variables. In this paper, we propose a method to learn the discrepancy of a stochastic simulation using data collected from the system of interest. Our approach is a novel Bayesian framework that addresses the requirements for estimation of probability measures.

1 INTRODUCTION

Stochastic, or simulation, models are only approximations to the reality. A conjectured model may not align with the true system because of unobserved complexity. Moreover, some highly accurate models, even if formulable, may not be implementable due to computational barriers and time constraints, in which case a simpler surrogate model is adopted. In all these cases, there is a discrepancy between the model and the reality.

This important issue is traditionally handled through *model validation* and *calibration* in the simulation literature. Model validation refers to the task of checking whether a developed stochastic model sufficiently reflects the reality, often by comparing simulation outputs with real output data. A conventional method compares the relevant outputs from these two sources statistically (e.g. Schruben 1980, Balci and Sargent 1982). If the stochastic model is found to be inaccurate, ad-hoc calibration is performed to improve the model recursively until the model is sufficiently consistent with the reality (Sargent 2005). While intuitive, this conventional method may require building increasingly sophisticated models and it lacks a mechanism to capture persistent discrepancies outside the structure conjectured by the model builders.

More broadly, the term calibration refers to the class of methods that can adjust the model to bring it closer to reality. Model calibration using data has been well studied in the deterministic model case (e.g., Kennedy and O'Hagan 2001). General methodologies to handle the stochastic model case, however, have not gone much beyond the proposal to re-calibrate given an insufficient coherence between simulation and real data revealed by testing. A unique property of the stochastic case is that we must account for the discrepancy that considers the properties of probability distributions, which constitute the typical representation of the observed outputs. Most existing statistical calibration methods in the deterministic domain, e.g. Higdon et al. (2004), Higdon et al. (2008), Storlie et al. (2015), are thus not directly applicable. In the stochastic domain, Goeva et al. (2014) studied calibration for the distribution of the simulation input model, but assuming the simulation logic is correctly specified. A stream of work in the queueing inference literature studied the calibration of input processes and system performances (e.g., Whitt 1981, Basawa et al. 1996, Pickands III and Stine 1997, Bingham and Pitts 1999, Hall and Park 2004, Larson 1990), given congestion or transaction data. These work are based on specific queueing structures that could be approximated either analytically or via diffusion limits, and, moreover, that are not subject to misspecification errors.

This paper studies a general framework to learn, and correct, for discrepancies between stochastic outputs from simulation models and the reality. In contrast to the past literature, we do not restrict the simulation logic or presume the logic matches reality. Instead, we focus on situations where responses from the real system are recorded for multiple scenarios. Taking into account the discrepancies between the real data and the model outputs at these different scenarios, we seek to give an estimate for the performance measure of interest at a described scenario. Our key methodological contribution is to build a Bayesian learning framework that operates on the probabilistic representation inherent in simulation models by using likelihood ratios as our main inference objects.

Pedagogical example To fix the conceptual ideas in this paper, we borrow from the call center data originally analyzed in Brown et al. (2005). This dataset is associated with a call center where a customer is placed into a queue once a call is received until one of n servers is available. From these data, the sample mean of the waiting time (from entry to service for a customer) from 9:00 to 10:00 am is calculated. In this narrow time period, the arrival rate and service rate, which is time inhomogenous according to Brown et al. (2005), should be approximately homogenous. The average arrival rate in this data was 1.76 customers per minute and the average service time was 3.72 minutes per customer. In this paper, we also account for the number of servers operating in the system at any given time, which appears to differ between days. To our reading, this subset of the data was by-and-large ignored in Brown et al. (2005)'s original analysis.

Our model for this call center will be an n-server first-come-first-serve queue. Following practice, both the interarrival and service times will be modelled as exponentially distributed. After a warm up period, the sample average of the waiting time is measured over the course of a one hour window. This, in principal, agrees with Brown et al. (2005). Moreover, this modeling framework likely emulates many common practices. In the spirit of ad-hoc calibration, two additional features were added: (i) the arrival rate is randomly generated each day from a log normal distribution with associated mean parameter 1.8 and variance 0.4 and (ii) with probability 2/3, a customer will abandon the queue if the wait time is longer than an exponential random variable with mean 1.5, otherwise they will remain in the queue until served. Adding both of these features resulted in a simulation model that was closer to the observed data.

When this example is revisited in this article, it will become clear that the simulation model output does not exactly agree with the collected responses. Even though the parameters of the model are well-estimated and the model has been deliberately calibrated, the model is not perfect thus it does not agree with the collected responses from the real system. This creates the problem in practice: how does one go about calibrating this model in a precise way using these observations?

2 GENERAL SETTING

In this paper, the system of interest outputs a discrete (or discretized) response over a space $\mathscr{Y}$ with cardinality q. This response depends on a vector of design variables, denoted x, and each potential combination are indexed from $i = 1,\ldots,m$, where m is the total feasible combinations of these variables. The variable x can be broadly defined to include input variables that can not necessarily be "controlled". An example of the response is the waiting time in call centers (which is the running pedagogical example) or hospitals (Helm and Van Oyen 2014). In the first example, design variables could be the number of servers, the system capacity, and the arrival rate. In the second example, the design variables could be the rate of elective admissions. The probability measure π_i describes the distribution of the response of the real system. The term design point will be understood to mean the point in the space corresponding to a combination of design variables.

The objective is to make decisions about the distribution of the response's distribution π_i for all i between 1 and m. In reality, a user rarely has direct access to the true distribution functions. Perhaps the simplest tool to avoid this problem is by directly observing many independent responses from the system, possibly from a designed experiment (Li et al. 2015). Using this data,

$$\widehat{\pi}_i(\gamma) = \frac{1}{N_i} \sum_{k=1}^{N_i} I\left(y_{ik} = \gamma\right) \tag{1}$$

is an estimate of the probability mass function for the true response. Here, I is the indicator function and y_{ik} is the kth recorded response on the ith setting. For each of the m design points, there are N_i recorded responses corresponding to the ith decision variable. As N_i grows, $\widehat{\pi}_i(\gamma)$ will converge to the true mass function, $\pi_i(\gamma)$, under reasonable conditions.

In practice, this purely empirical estimate is not sufficient to make decisions. A user needs many samples in order to directly estimate this quantity to sufficient precision with $\widehat{\pi}_i(\gamma)$ alone. This problem is compounded because doing this for all possible x_i is almost never a reasonable option. A user may not have access to the some of the design points. For example, this can happen when a potential design point has been proposed but has not yet been implemented. So for many i, N_i is 0, which makes directly estimating π_i purely from observed responses impossible.

Instead of attempting to directly conduct inference on the unknown $\pi_i(\cdot)$, operation researchers often choose a surrogate model to use instead. These models are described using state-of-the-art understanding of a given system, perhaps simplified for computational reasons. The example mentioned in the introduction is the use of first-come first-serve multi-server queues with a Poisson arrival process and an exponentially distributed service times. Despite being a reasonable model, it shows obvious discrepancies with the real data.

Let $\tilde{\pi}_i$ be the surrogate model's distribution associated with the design point x_i. For simplicity, assume that the surrogate model is exactly provided in the sense that simulation is cheap relative to the cost of physical experimentation, and that all parameters in this model are fixed. If the surrogate model is assumed exactly equal to the true model, then $\pi_i(\gamma)$ can be found precisely with the relation $\pi_i(\gamma) = \tilde{\pi}_i(\gamma)$, and no data is needed. But, as discussed before, there is almost always a discrepancy between π_i and $\tilde{\pi}_i$ in practice. Thus, in this paper, we will operate from the assumption that the surrogate model is *inexact*, and that data is collected from the system. In general, this means that $\pi_i(\gamma) \neq \tilde{\pi}_i(\gamma)$ for at least some $\gamma \in \mathscr{Y}$ and some $i \in \{1,\ldots,m\}$, where π_i needs to be statistically inferred. As of yet, there is no systematic way to address this uncertainty and learn the inexactness to the authors' knowledge. This motivated our approach to automatically correct the deficiencies using data, by viewing the model as a black box. The next section will explain the definition of discrepancy in simulation models so that the proposed correction and uncertainty quantification scheme is well defined. Before that, we revisit our running example:

Pedagogical example (continued) The response is discretized into the four categories < 1, $1-2$, $2-3$, > 3 and thus $q = 4$. In this example, $5-9$ servers were considered and thus $m = 5$. We also have $\tilde{\pi}$ to a reasonable approximation by simulating from the described model 1000 times for each design point. Figure 1 shows the frequency diagram of the data along with the surrogate model. The data we use here does not include 5 servers and 9 servers. Using this simulation model, we can conjecture what would happen at these staffing levels. Going by the simulation model alone, one would expect that when 5 servers are present, the average waiting time should often be larger than 3, and when 9 servers are present, the average waiting time will almost never be above 1.

But when looking at the collected responses next to the simulation model, the picture is more muddled. According to the simulation model, the chance of the average waiting time between 1 and 2 minutes is around 8% when 8 servers are used. But in our collected responses, about 25% of the 23 recorded observations are in this range. This indicates that the model is inexact. This knowledge should affect our prediction of the probability of the response being between 1 and 2 when 9 servers are present. Our methodology presented in the next sections will explicitly bring such knowledge into consideration.

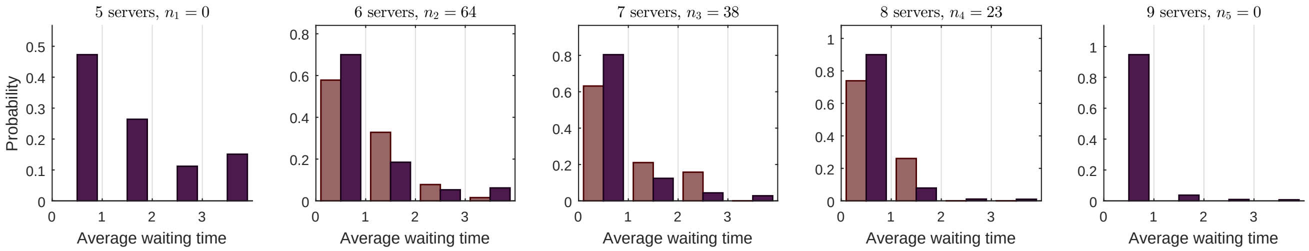

Figure 1: Frequency diagram of the real data (the left bars) and the probabilities computed from the surrogate model (the right bars) for the pedagogical example.

3 DISCREPANCIES AND VALID DISCREPANCIES

This article suggests directly of modeling the *discrepancy* between π_i and $\tilde{\pi}_i$. The discrepancy δ_i is defined as

$$\delta_i(\gamma) = \frac{\pi_i(\gamma)}{\tilde{\pi}_i(\gamma)}, \text{ for all } \gamma \in \mathscr{Y}. \tag{2}$$

i.e., the ratio between π_i and $\tilde{\pi}_i$. A value of $\delta_i(\gamma)$ that is exactly 1 for all $\gamma \in \mathscr{Y}$ would represent a surrogate model with no discrepancy. The δ_i reflects the the relative probability of the response between the true response and simulated response. Model inexactness can then be cast as $\delta_i(\gamma) \neq 1$ for some $\gamma \in \mathscr{Y}$ and i between 1 and m.

There are direct analogies between δ_i and the likelihood ratio (or adjustment factor) in the context of importance sampling (e.g., Owen (2013), Chapter 9; Asmussen and Glynn (2007), Chapter IV; Glasserman (2003), Chapter 4). However, whereas importance sampling is used to speed up Monte Carlo schemes under a known stochastic model, and the associated π_i and δ_i in that situation are therefore known and are designed by the user, ours are not fully observable. In the uncertainty quantification context, δ_i also appears in worst-case optimizations formulated to quantify model errors. These optimizations compute conservative estimates of performance measures subject to beliefs or uncertainty on the true model relative to the surrogate (often known as the baseline model). For instance, Hu et al. (2012) considered Gaussian models with mean and covariance uncertainty represented by linear matrix inequalities. Glasserman and Xu (2014) considered nonparametric uncertainty measured by Kullback-Leibler divergence. A notable difference between these lines of work and ours is the role of δ_i. While the former is a decision variable in an optimization program that computes a worst-case bound, ours is an object to be *inferred* from data to improve the prediction accuracy.

For simplicity, we assume that our surrogate probabilities $\tilde{\pi}_i(\gamma)$ are all strictly positive, thus this value of $\delta_i(\gamma)$ will naturally be ≥ 0 and finite. Any $\tilde{\pi}_i(\gamma)$ for which this is not naturally true can be slightly adjusted by adding small probability for all $\gamma \in \mathscr{Y}$ and renormalizing the distribution such that the masses sum to one.

In addition to the fact that $0 \leq \delta_i(y) < \infty$, there is another important constraint on δ_i that relates to the fact that π_i is a probability distribution. Because

$$\sum_{\gamma \in \mathscr{Y}} \pi_i(\gamma) = 1, \text{ then } \sum_{\gamma \in \mathscr{Y}} \delta_i(\gamma)\tilde{\pi}_i(\gamma) = 1.$$

Thus the L^2 inner product between δ_i and $\tilde{\pi}_i$ over $\mathscr{Y}$ must be exactly equal to 1 for all i from 1 to m.

From the proceeding discussion,

Definition Say $\tilde{\pi}_i$ is a probability distribution over the discrete sample space $\mathscr{Y}$ with $\tilde{\pi}_i(\gamma) > 0$ for all $\gamma \in \mathscr{Y}$. Any mapping $h : \mathscr{Y} \to \mathbb{R}$ is a *valid discrepancy* with respect to $\tilde{\pi}_i$ if

(i) $0 \leq h(\gamma) < \infty$ for all $\gamma \in \mathscr{Y}$ and

(ii) $\sum_{\gamma \in \mathscr{Y}} h(\gamma)\tilde{\pi}_i(\gamma) = 1$.

All δ_i as defined in (2) will meet the criteria outlined above, regardless of the true distribution (provided the true distribution is indeed a proper probability distribution over $\mathscr{Y}$).

These two conditions are consistent with the traditional assumptions on likelihood ratios. We highlight them here because they characterize the type of objects that we conduct inference on. Moreover, they show a distinction between inferring stochastic and deterministic model discrepancies, in that probabilistic structure plays a role in the former but not the latter (e.g. compare to Kennedy and O'Hagan (2001)).

We also comment at this point that the finite support assumption of the output distribution, hence the discrepancy, avoids the technical difficulties in both the statistical inference and the computation involving infinite-dimensional parameters. For continuous distributions, one approach is to project them onto a finite number of statistics, such as moments. Such investigation is out of the scope of this work.

4 LEARNING DISCREPANCY

Calibration problems are often located in data-poor environments (in terms of observed responses) which are also rich in prior knowledge. Beyond the notion of a valid discrepancy provided in Section 3, there is often other information available. Even though observed responses do not exist for every design point, similar discrepancies might be anticipated for similar design points. This similarity is measured by the distance in terms of the design variables associated with each point. Recall the example of waiting times in call centers; the discrepancy when the number of servers is 5 is expected to be more similar to the discrepancy when the number of servers is 6 relative to when the number of servers is 8.

To incorporate this type of prior information in a data-poor environment, this paper will use a Bayesian learning approach. Under this framework, the observed responses have a certain likelihood of occurring given the knowledge of the discrepancy. Combining the likelihood of these observed responses with the prior information creates a posterior for the discrepancy.

More concretely, we let p generally represent a probability density and $p(a|b)$ implies the conditional probability density of a given b. Let a vector d represent a vectorized version of the discrepancy:

$$d = [\delta_1(\gamma_1), \ldots, \delta_1(\gamma_q), \cdots, \delta_m(\gamma_1), \ldots, \delta_m(\gamma_q)]^\mathsf{T}. \tag{3}$$

The term *data* substitutes for the event of all observed responses from the true system:

$$\text{data} = \left\{Y_{ij} = y_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, N_i\right\}.$$

Our objective is thus to conjecture on

$$p\left(d \,|\text{data}\right),$$

which would define the *posterior* of d. Using Bayes rule,

$$p\left(d \,|\text{data}\right) \propto p(\text{data}|d)p(d), \tag{4}$$

where $\propto$ stands for equality up to a constant multiplier. The computational methods that we will propose later require only this proportional value.

The first element in the right hand side of (4) is referred to as the likelihood and the second element in the right hand side of (4) is referred to as the prior distribution. The likelihood is straightfowardly defined based on π. Our prior distribution will be such that

$$p\left(d = g + G\theta\right) \propto \exp\left(\varepsilon 1^\mathsf{T} \log(g + G\theta) - \lambda\, \theta^\mathsf{T} \theta\right),$$

where ε is a small positive constant, λ is a tuning parameter left to the user, $\theta \in \mathbb{R}^{pq}$ relates to d via an affine map $d = g + G\theta$ with g and G designed such that we meet the criteria of a valid discrepancy. Based on this prior distribution, our posterior on θ is

$$p\left(\theta\,|\text{data}\right) \propto \exp\left((N+\varepsilon)^{\mathsf{T}}\log(g+G\theta)-\lambda\,\theta^{\mathsf{T}}\theta\right), \tag{5}$$

and we can recover d with $d = g+G\theta$. Note that N is the length mq vector of counts

$$N = \left[\sum_{j=1}^{n_1} I(y_{1j}=\gamma_1),\ldots,\sum_{j=1}^{n_1} I(y_{1j}=\gamma_q),\cdots,\sum_{j=1}^{n_m} I(y_{mj}=\gamma_1),\ldots,\sum_{j=1}^{n_m} I(y_{mj}=\gamma_q)\right]^{\mathsf{T}}.$$

We explain how the chosen prior distribution agrees with the concept of a valid discrepancy. Basically, it can be interpreted as a multivariate Normal prior on the discrepancy subject to the probabilistic constraints. The vector g and matrix G are derived from a positive definite matrix R and a mean vector μ. These R and μ can be considered like the mean and covariance for the prior on the bias. For example, it is reasonable to use a vector of 1s as the prior mean for μ. But a prior distribution on d of Normal with a mean vector of 1s and a covariance matrix that is proportional to R will not satisfy the probabilistic definition of a valid discrepancy with probability one in general. To resolve this, we let g and G correspond to

$$g = \mu + R\tilde{\Pi}^{\mathsf{T}}\left(\tilde{\Pi}R\tilde{\Pi}^{\mathsf{T}}\right)^{-1}(1-\tilde{\Pi}\mu)$$

and

$$G = \sqrt{R - R\tilde{\Pi}^{\mathsf{T}}\left(\tilde{\Pi}R\tilde{\Pi}^{\mathsf{T}}\right)^{-1}\tilde{\Pi}R}.$$

These can be thought of as the conditional mean and square root of the covariance after conditioning on (ii) in the definition of a valid discrepancy. To the authors' knowledge, this method of conditioning normal distributions to resolve the probabilistic definition of a valid discrepancy has not been used elsewhere in the Bayesian statistical literature.

We also need to consider condition (i) when assigning a prior distribution, which is taken care of by the inclusion of the ε in our prior. This can be considered as replacing a sharp constraint with what are known in optimization as *barrier functions* (Ben-Tal and Nemirovski 2001, pp 274). That is, $I(\delta_i(\gamma_j)) \geq 0)$ is replaced with the barrier

$$\exp(\varepsilon\log(\delta_i(\gamma_j))),$$

where $\varepsilon > 0$ is some small constant.

Thus we conclude the following result based on the described prior distribution:

Theorem 1 Say that $\tilde{\pi}_i(\gamma) > 0$ for all $i \in \{1,\ldots,m\}$ and $\gamma \in \mathscr{Y}$. There exists a constant C such that

$$C\exp\left(\varepsilon 1^{\mathsf{T}}\log(g+G\theta)-\lambda\,\theta^{\mathsf{T}}\theta\right),$$

is a valid probability density on $\mathbb{R}^{mq}$. Moreover, if θ is drawn from this distribution on $\mathbb{R}^{mq}$, then with probability one

$$\text{(i) } g+G\theta > 0 \text{ and (ii) } \tilde{\Pi}\left(g+G\theta\right) = 1.$$

Pedagogical example (continued) Consider our pedagogical example described in Sections 1 and 2. Figure 2 shows the results of the Bayesian analysis of the data. The top panels demonstrate what would happen if the probabilistic constraints that control the notion of a valid discrepancy were ignored in our analysis. Clearly, the non-sensical results with probabilities larger than 1 would not meet any criteria for decision making. While not present here, sometimes these bounds can dip below zero, inducing severe inferential problems. Using the proposed approach that takes into account probabilistic constraints, these issues appear resolved as illustrated in the bottom panels.

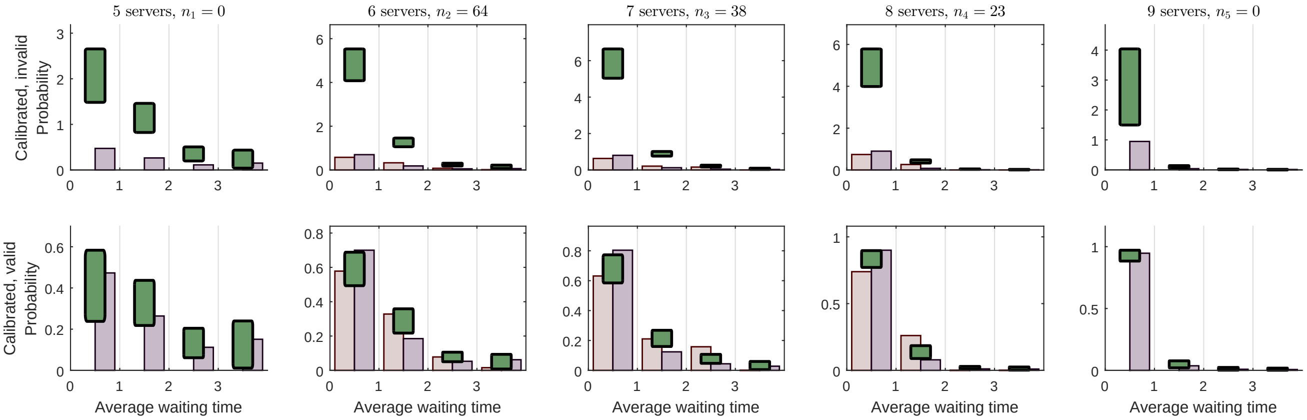

Figure 2: Frequency diagram of the observed responses and the probabilities computed from the surrogate model for the pedagogical example with predictive intervals marked by the rectangles for π based on the observed responses, the surrogate model, and our prior information. The top panels use prior information that ignores validity and the bottom panels use prior information that includes the notion of a valid discrepancy.

A few issues were brought up in the discussion of this example in Section 2. Specifically, when there were 9 servers, how likely was it that the average waiting time would be between 1 and 2 minutes? There is no data, but the simulation model combined with the observed responses and our prior information gives us an estimate of somewhere between 2.5% and 8%. This accounts for the discrepancy that we observed based on the recorded responses at 8 servers. Overall, this range appears to agree with both the data and the surrogate outputs. We are not confident, for example, that staffing 5 servers will produce the same results as the simulation model, which has average waiting times over 3 minutes about 15% of the time. Based on the recorded responses, this could be 22%, but it could also be as low as about 2%. More data would produce tighter bounds; see the result with 64 data points when the number of servers is 6. The confidence intervals are much thinner due to more available data.

5 POSTERIOR COMPUTATIONS

This section discusses Bayesian computation methods for a quantity of interest, described by a function $\zeta(\cdot,\cdot) : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, from the posterior. That is, we find upper and lower bounds for

$$\sum_{i=1}^{m} \sum_{\gamma \in \mathscr{Y}} \zeta(x_i,\gamma)\pi_i(\gamma) = \sum_{i=1}^{m} \sum_{\gamma \in \mathscr{Y}} \zeta(x_i,\gamma)\tilde{\pi}_i(\gamma)\delta_i(\gamma).$$

Let

$$\tilde{z} = \left[\zeta(x_1,\gamma_1)\tilde{\pi}_1(\gamma_1),\ldots,\zeta(x_1,\gamma_q)\tilde{\pi}_1(\gamma_q),\cdots,\zeta(x_m,\gamma_1)\tilde{\pi}_m(\gamma_1),\ldots,\zeta(x_m,\gamma_q)\tilde{\pi}_m(\gamma_q)\right].$$

5.1 Sampling-based approach

Given w samples from the length mq vectors of the discrepancy parameter $\theta^{(1)},\ldots,\theta^{(w)}$, samples from the posterior of the quantity of interest are

$$\{\tilde{z}^{\mathsf{T}} g + \tilde{z}^{\mathsf{T}} G\theta^{(k)}, \quad k = 1,\ldots,w\}.$$

The 95% credible interval is then the gap between the 2.5% and 97.5% empirical quantiles.

Because of the nonregularity of the posterior (and the fact that there is an unknown proportionality constant), direct Monte Carlo sampling is not possible. Sophisticated Markov chain Monte Carlo samplers were designed explicitly for this purpose (Gelman et al. 2014, Chapters 11 and 12). Perhaps the simplest mechanism for sampling from these nonregular type is a simple Metropolis Hastings algorithm with a symmetric proposal (Gelman et al. 2014, pp 278-280). Here, some initial $\theta^{(0)}$ is chosen, typically as the maximizer of the posterior in (5). Then an arbitrary symmetric distribution is selected such that the next point is $\theta^{(0)}+\rho^{(0)}$, where $\rho^{(0)}$ is drawn from this symmetric distribution. Starting with $i=0$, the next sample from step i is

$$\theta^{(i+1)} = \begin{cases} \theta^{(i)}+\rho^{(i)}, \text{ with probability } \min\left(1, \frac{p(\theta^{(i)}+\rho^{(i)}|data)}{p(\theta^{(i)}|data)}\right) \\ \theta^{(i)}, \text{ otherwise.} \end{cases}$$

In the above $\rho^{(i)}$ are independent draws from this symmetric distribution.

Another approach to consider, which has grown in popularity, is the idea of using Hamiltonian dynamics to modify the Metropolis Hastings algorithm with a non-symmetric proposal distribution. The key benefit of this approach is that it considers the derivative of the posterior when constructing proposal samples. The leap-frog algorithm, taken from Neal (2011), pp 121) and modified here for our specific density function, is that given $\theta^{(i)}$ the next sample will be:

Algorithm 1 Hamiltonian Monte Carlo Sampler

```
  $\theta^{\text{prop}} = \theta^{(i)}$
  Draw $v_0$ as a $mq$ vector of independent standard Gaussian random variables,
  Let $v = v_0 + \frac{1}{2}\left((N+\varepsilon 1)^{\mathsf{T}}(g+G\theta^{\text{prop}})^{-1}\right)G - \lambda\theta^{\text{prop}}$
  for $L$ times do
      $\theta^{\text{prop}} = \theta^{\text{prop}} + ev$
      $v = v + \left((N+\varepsilon 1)^{\mathsf{T}}(g+G\theta^{\text{prop}})^{-1}\right)G - 2\lambda\theta^{\text{prop}}.$
  $\theta^{(i+1)} = \begin{cases} \theta^{(prop)}, \text{ with probability } \min\left(1, \frac{p(\theta^{(prop)}|data)}{p(\theta^{(i)}|data)} \frac{\exp\left(\|v+\frac{1}{2}\left((N+\varepsilon)^{\mathsf{T}}(g+G\theta^{\text{prop}})^{-1}\right)G-\lambda\theta^{\text{prop}}\|^2/2\right)}{\exp(\|v_0\|^2/2)}\right) \\ \theta^{(i)}, \text{ otherwise.} \end{cases}$
```

The numerical constant $e>0$ and the integer $L>0$ are parameters of this algorithm that can be adjusted to reach a distribution more inline with the results in the problem.

Lastly, consider slice sampling (Neal 2003). This sampler, similar to principal to the Hit and Run algorithm (Bélisle et al. 1993) and the approach of Chen and Schmeiser (1998). Because of the unique properties of our posterior (namely, convexity of the level sets), the reflective slice sampling has particular appeal. The basic idea is to sample uniformly from level-sets of the density by choosing a direction and reflecting off the barrier of the level set. Start with some positive constant α and let θ^* be the maximizer of the posterior. The algorithm is that given $\theta^{(i)}, \alpha>0$ the next sample will be:

Algorithm 2 Slice Sampler

```
  Set $\theta^{(i+1)} = \theta^{(i)}$.
  Draw random direction $h$
  for $L$ times do
      Find largest step size $\Delta$ along $h$ such that $p(\theta^{(i+1)}+\Delta h|data) \geq p(\theta^*|data)\exp(-\alpha)$.
      Set $\theta^{(i+1)} = \theta^{(i+1)} + \Delta h$.
      Let $v = \left((N+\varepsilon 1)^{\mathsf{T}}(g+G\theta^{(i+1)})^{-1}\right)G - 2\lambda\theta^{(i+1)}$ and update $h = h - 2v\frac{v^{\mathsf{T}}h}{v^{\mathsf{T}}v}$.
  Find largest step size $\Delta$ along $h$ such that $p(\theta^{(i+1)}+\Delta h|data) \geq p(\theta^*|data)\exp(-\alpha)$.
  Let $U$ be a uniform random variable between 0 and 1 and $\theta^{(i+1)} = \theta^{(i+1)} + U\Delta h$.
  Let $\alpha = \log p(\theta^*|data) - \log p(\theta^{(i+1)}|data) - \log(U)$.
```

The numerical constant $L > 0$ is a parameter of this algorithm that can be adjusted. There is no need for an acceptance/rejection step. The major computational the burden of this sampler lies in our ability to find the maximum step size quickly and effectively.

Pedagogical example (continued) We attempted the samplers from Section 5.1 on the posterior described in the running example. The results are located in Table 1. Attempts were made to optimize all three methods and balance the computational time by using different total amounts of samples collected. Half the samples for each sampler were discarded as a warm-up period. Though these results are limited, they indicate that the choice of sampler will play a major role in being able to efficiently sample from the posterior of the quantity of interest. The method with the best effective sample size on a subsample of 200 was the slice sampling approach. This is despite that the number of iterations is low compared to the other two methods.

Table 1: The results for the sampling approaches using the queueing data as described in Section 5.1.

	MH	HMC	SS
Total time (seconds)	0.51	0.61	1.05
Recorded samples	10,000	800	400
ESS	48/200	107/200	198/200
% rejected	44.6 %	18.0 %	0.0%

5.2 Optimization-based approach

Because of the complexity in the setup and execution of the above samples, we also propose a new method for posterior inference using convex optimization to develop confidence bounds for the quantity of interest. This approach is inspired from the literature of robust optimization (e.g., Ben-Tal and Nemirovski 2002, Ben-Tal et al. 2009, Bertsimas et al. 2011) that represents uncertain or ambiguous parameters in an optimization problem by imposing constraints, forming the so-called ambiguity or uncertainty sets. When the uncertain parameters follow probability distributions, the uncertainty sets are constructed as a predictive set that contains the truth with a prescribed confidence. This is particularly useful in approximating chance-constrained programs (e.g., Ben-Tal and Nemirovski 2002, Chapter 2), and performance measures from complex stochastic models (e.g., Bandi and Bertsimas 2012, Bandi and Bertsimas 2014). Recently, Gupta (2015) considered obtaining Bayesian statistical guarantees for uncertainty sets, particularly for the case of distributionally robust optimization where the uncertain parameters are themselves probability distributions. Beyond that, to our best knowledge, there has been no attempt to use robust optimization as a tool for computing Bayesian posterior. Thanks to the probabilistic structure and the high dimensionality of the parameter space (qm) induced in our setting, we advocate robust optimization as an efficient computational tool for inferring on the posterior distribution.

To begin, consider an uncertainty set of d based on the posterior. This uncertainty set replaces the probabilistic posterior with a deterministic set-based statement of the feasibility of the discrepancy. That is, given some constant c that controls the "size" of the uncertainty set, let

$$\mathscr{U} = \{d = g + G\theta \,| p(\theta|\text{data}) \geq c, \theta \in \mathbb{R}^{qm}\}.$$

Clearly, as c gets smaller, $\mathscr{U}$ is nondecreasing in size. Motivated by some theoretical results excluded from this article, a reasonable approach would be to set

$$c = \frac{1}{2}\Phi^{-1}(w)^2 + \max_{\theta \in \mathbb{R}^{qm}} p(\theta|\text{data}).$$

The benefits of this choice are two fold. First, this guarantees that $\mathscr{U}$ will be nonempty. Second, this choice leads to asymptotic bounds very close to 95% confidence bounds, which are popular in diverse areas of science and business.

Our optimization problem to find the bounds is stated as

$$\begin{array}{ll} \text{max or min} & \tilde{z}^{\mathsf{T}} d, \\ \text{subject to} & d \in \mathscr{U}. \end{array} \tag{6}$$

which can be solved with most convex programming approaches if the feasible region is convex (such as penalty methods and interior point methods; see Bertsekas (1999)). Because of the structure of the problem,

Proposition $\mathscr{U}$ is a strictly convex set on the interior of $\mathbb{R}_+^{qm}$ for all $\varepsilon > 0$ and $\lambda > 0$.

Note that the assumption of the prior plays a critical role. If ε was not included in the prior, we would not be guaranteed that $\mathscr{U}$ is a bounded set on the interior of valid discrepancies (by the coercivity of $p(\cdot|\text{data})$). The prior for θ was chosen to be log-concave. If a different choice was made for the prior on θ, $\mathscr{U}$ may not be convex.

Pedagogical example (continued) We now offer a graphical explanation of the differences in the methods described in Sections 5.1 and 5.2. Say a user is concerned with staffing exactly 6 servers. What is an upper bound on the probability that the average wait time is below 2 minutes? Using the discretization introduced for this example, this would correspond to

$$\pi_2(\gamma_1) + \pi_2(\gamma_2) = \tilde{\pi}_2(\gamma_1)\delta_2(\gamma_1) + \tilde{\pi}_2(\gamma_2)\delta_2(\gamma_2) = 0.251\delta_2(\gamma_1) + 0.294\delta_2(\gamma_2).$$

i.e., $\zeta(x_2,\gamma_1) = \zeta(x_2,\gamma_2) = 1$, and all other ζ's are 0.

Consider the sampling based approach. If 200 outputs are sampled from $\theta^{(1)},\ldots,\theta^{(200)}$, the upper credible interval would be determined by finding the hyperplane parallel to $0.251\delta_2(\gamma_1) + 0.294\delta_2(\gamma_2)$ such that exactly 5 samples are to the right and above it. The robust optimization approach works similarly, with the major difference that the hyperplane is maximized over the convex set $\mathscr{U}$ described in Section 5.2. This is illustrated in Figure 3, where it should be emphasized that these are both the samples and the true uncertainty set computed from the data, not merely a cartoon. Here, the convexity (and nonellipsoidal) shape of the projection of the uncertainty set is verified visually. The samples appear to roughly agree with the uncertainty set despite them being generated from an approximate sampler.

6 Conclusion

We have studied inference on model discrepancies between surrogate stochastic or simulation models and real response data under the availability of multiple design points. This probabilistic constraints of this case is unique to stochastic settings and deviates from previous work in the deterministic setting where no such constraint is considered, and in other stochastic calibration problems where specific model structures are assumed known or perfectly specified. This articles has also compared two ways of computing posterior predictive intervals, typical samplers and a new optimization approach. We anticipate that learning stochastic model discrepancy can be extended to higher-dimensional problems (such as continuous, and thus infinite-dimensional, distributions). The major complexity here is that the number of positivity constraints is on the order of the space space, thus any proposed method must be careful designed to be well-defined and computable.

ACKNOWLEDGMENTS

We thank Ilan Guedj for the data organization and for Avi Mandelbaum for continuing to place the data on the website http://ie.technion.ac.il/serveng/. We also gratefully acknowledge support from the the National Science Foundation under grants CMMI-1400391/1542020 and CMMI-1436247/1523453.

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. Springer Science & Business Media.

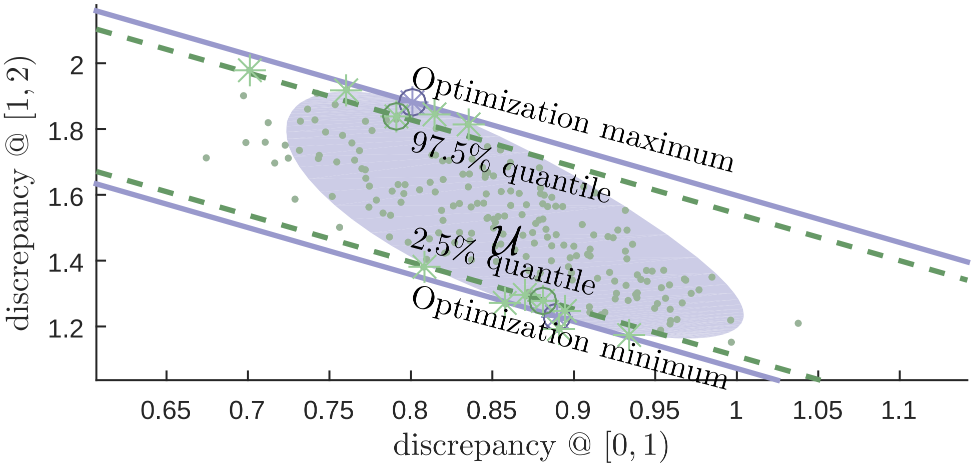

Figure 3: Graphical description of the differences between optimization- and sampling-based approaches for the stochastic discrepancy problem. The bottom axis corresponds to $\delta_2(\gamma_1)$ and the vertical axis corresponds to $\delta_2(\gamma_2)$. The light arrows describe the gradient of the objective function. The 200 dots are samples from a well-tuned slice sampler. The upper limit of the 2.5% and 97.5% quantile is indicated by the dashed lines line. The region labeled $\mathscr{U}$ is the projection of the convex set $\mathscr{U}$ onto this two-dimensional plane. The maximum of the optimization is determined by the pictured line.

Balci, O., and R. G. Sargent. 1982. “Some examples of simulation model validation using hypothesis testing”. In *Proceedings of the 1982 Winter Simulation Conference*, edited by Highland, Chao, and Madrigal, 621–629. Piscataway, NJ: IEEE.

Bandi, C., and D. Bertsimas. 2012. “Tractable Stochastic analysis in high dimensions via robust optimization”. *Mathematical Programming* 134 (1): 23–70.

Bandi, C., and D. Bertsimas. 2014. “Robust option pricing”. *European Journal of Operational Research* 239 (3): 842–853.

Basawa, I. V., U. N. Bhat, and R. Lund. 1996. “Maximum likelihood estimation for single server queues from waiting time data”. *Queueing Systems* 24 (1-4): 155–167.

Bélisle, C. J., H. E. Romeijn, and R. L. Smith. 1993. “Hit-and-run algorithms for generating multivariate distributions”. *Mathematics of Operations Research* 18 (2): 255–266.

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski. 2009. *Robust Optimization*. Princeton University Press.

Ben-Tal, A., and A. Nemirovski. 2001. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM.

Ben-Tal, A., and A. Nemirovski. 2002. “Robust optimization–methodology and applications”. *Mathematical Programming* 92 (3): 453–480.

Bertsekas, D. P. 1999. *Nonlinear Programming*. Athena Scientific Belmont.

Bertsimas, D., D. B. Brown, and C. Caramanis. 2011. “Theory and applications of robust optimization”. *SIAM Review* 53 (3): 464–501.

Bingham, N., and S. M. Pitts. 1999. “Non-parametric estimation for the M/G/∞ queue”. *Annals of the Institute of Statistical Mathematics* 51 (1): 71–97.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. “Statistical analysis of a telephone call center: A queueing-science perspective”. *Journal of the American Statistical Association* 100 (469): 36–50.

Chen, M.-H., and B. Schmeiser. 1998. “Toward black-box sampling: A random-direction interior-point Markov chain approach”. *Journal of Computational and Graphical Statistics* 7 (1): 1–22.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2014. *Bayesian Data Analysis*. 3 ed. Taylor & Francis.

Glasserman, P. 2003. *Monte Carlo Methods in Financial Engineering*. Springer Science & Business Media.

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

Goeva, A., H. Lam, and B. Zhang. 2014. "Reconstructing input models via simulation optimization". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 698–709. Piscataway, NJ: IEEE.

Gupta, V. 2015. "Near-Optimal Ambiguity Sets for Distributionally Robust Optimization". Technical report.

Hall, P., and J. Park. 2004. "Nonparametric inference about service time distribution from indirect measurements". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66 (4): 861–875.

Helm, J. E., and M. P. Van Oyen. 2014. "Design and optimization methods for elective hospital admissions". *Operations Research* 62 (6): 1265–1282.

Higdon, D., J. Gattiker, B. Williams, and M. Rightley. 2008. "Computer Model Calibration Using High-Dimensional Output". *Journal of the American Statistical Association* 103 (482): 570–583.

Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. 2004. "Combining field data and computer simulations for calibration and prediction". *SIAM Journal on Scientific Computing* 26 (2): 448–466.

Hu, Z., J. Cao, and L. J. Hong. 2012. "Robust simulation of global warming policies using the DICE model". *Management Science* 58 (12): 2190–2206.

Kennedy, M. C., and A. O'Hagan. 2001. "Bayesian calibration of computer models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (3): 425–464.

Larson, R. C. 1990. "The queue inference engine: Deducing queue statistics from transactional data". *Management Science* 36 (5): 586–601.

Li, J. Q., P. Rusmevichientong, D. Simester, J. N. Tsitsiklis, and S. I. Zoumpoulis. 2015. "The value of field experiments". *Management Science* 61 (7): 1722–1740.

Neal, R. M. 2003. "Slice sampling". *Ann. Statist.* 31 (3): 705–767.

Neal, R. M. 2011. "MCMC using Hamiltonian dynamics". *Handbook of Markov Chain Monte Carlo*:113–162.

Owen, A. B. 2013. *Monte Carlo Theory, Methods and Examples*.

Pickands III, J., and R. A. Stine. 1997. "Estimation for an M/G/∞ queue with incomplete information". *Biometrika* 84 (2): 295–308.

Sargent, R. G. 2005. "Verification and validation of simulation models". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 130–143. Piscataway, NJ: IEEE.

Schruben, L. W. 1980. "Establishing the credibility of simulations". *Simulation* 34 (3): 101–105.

Storlie, C. B., W. A. Lane, E. M. Ryan, J. R. Gattiker, and D. M. Higdon. 2015. "Calibration of computational models with categorical parameters and correlated outputs via Bayesian smoothing spline ANOVA". *Journal of the American Statistical Association* 110 (509): 68–82.

Whitt, W. 1981. "Approximating a point process by a renewal process: The view through a queue, an indirect approach". *Management Science* 27 (6): 619–636.

AUTHOR BIOGRAPHIES

MATTHEW PLUMLEE (mplumlee@umich.edu) is an Assistant Professor researching simulation experiments and calibration at the Department of Industrial and Operations Engineering at the University of Michigan.

HENRY LAM (khlam@umich.edu) is an Assistant Professor researching stochastic simulation, risk analysis, and simulation optimization at the Department of Industrial and Operations Engineering at the University of Michigan.