Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

STATISTICAL INFERENCE FOR APPROXIMATE BAYESIAN OPTIMAL DESIGN

Prateek Jaiswal
Harsha Honnappa

School of Industrial Engineering
Purdue University
West Lafayette, IN 47906, USA

ABSTRACT

This paper studies a generic Bayesian optimal design formulation with chance constraints, where the decision variable lies in a separable, reflexive Banach space. This setting covers a gamut of simulation and modeling problems that we illustrate through two example problem formulations. The posterior objective cannot be computed, in general, and it is necessary to use approximate Bayesian inference. Sampling-based approximate inference, however, introduces significant variance and, in general, leads to non-convex approximate feasible sets, even when the original problem is convex. In this paper, we use variational Bayesian approximations that introduce no variance and retain the convexity of the feasibility set, subject to easily satisfied regularity conditions on the approximate posterior, albeit at the expense of a much larger bias. Our main results, therefore, establish large sample asymptotic consistency of the optimal solutions and optimal value of this approximate Bayesian optimal design formulation.

1 INTRODUCTION

This paper concerns chance constrained stochastic optimization problems over Banach spaces of the form

$$\begin{aligned} \underset{x\in\mathscr{X}}{\text{minimize}} \quad & F(x) := \int f(x,\xi)\pi(d\xi|\mathbf{X}_n) \\ \text{s.t.} \quad & \pi\left(g(x,\xi)\leq 0|\mathbf{X}_n\right) \geq \beta, \forall x\in\mathscr{X} \end{aligned} \tag{TP}$$

where $\beta\in(0,1)$ is the pre-specified confidence level desired by the decision-maker (DM), $\mathscr{X}$ is a convex subset (defined with respect to a suitable metric) of a separable reflexive Banach space, the function $f:\mathscr{X}\times\Theta\to\mathbb{R}$ encodes the cost/risk associated with the given values of the parameters $\xi\in\Theta\subseteq\mathbb{R}^k$ (for $d>0$) and $x\in\mathscr{X}$ (respectively), $g:\mathscr{X}\times\Theta\to\mathbb{R}$ define the constraint on the problem, $\pi(\cdot|\mathbf{X}_n)$ is a probability measure on the parameter space Θ and $\mathbf{X}_n$ is a $\mathbb{R}^{d\times n}$-valued random variable. While π can be quite general, in this paper we assume that it is a Bayesian posterior measure with support in Θ and $\mathbf{X}_n$ represents samples from an underlying stochastic model that is parameterized in some way by $\xi\in\Theta$. For simplicity we will assume that $\mathbf{X}_n$ consists of n independent samples of an $\mathbb{R}^d$-valued random variable X_0 with measure P_0.

Any solution of (TP) is a function of the sample size, which raises the question of statistical inference on the optimizers and the corresponding optimal value. In this paper, specifically, we address the *frequentist consistency* of this Bayesian procedure under the data-generating measure (corresponding to X_0). In simple terms, frequentist consistency refers to the phenomenon that the posterior measure "concentrates" on the data generating model in the large sample limit. Standard results on this question include Le Cam's results on Bayesian consistency as well as the Bernstein-von Mises theorem; see van der Vaart (1998).

These results assume that the posterior is computable in closed form, which it is not in general. It is, therefore, typical to use approximations to the posterior, computed using either Markov chain Monte

Carlo (MCMC) sampling or variational Bayesian (VB) procedures, as surrogates to the 'true' posterior distribution. In this paper, we consider approximations to the posterior (and the objective in (TP)) computed using VB. Where MCMC uses sampling/simulation methods for generating samples from (an approximation to) the posterior, VB directly computes approximations to the posterior distribution by solving a variational optimization problem (different from (TP)) over a family or subset of candidate approximating measures. There are several computational advantages to VB over MCMC, and we direct the reader to the review paper (Blei et al. 2017) for details. Specifically in the context of the the Bayesian optimal design problem, using MCMC (or just Monte Carlo) is rather problematic due to the presence of chance constraints – sampling is a significant source of variance in these settings in addition to the bias introduced by the Monte Carlo estimator (see Peña-Ordieres et al. (2020)) . VB approximations, on the other hand, trade variance for a higher bias. However, the bias can be controlled by carefully choosing the candidate measures.

Nonetheless, while VB has significant computational advantages, it should be acknowledged that the bias cannot be fully eliminated; as the reader will discern below, finding a no-bias approximation is equivalent to computing the posterior. Consequently, statistical inference is of paramount importance, and asymptotic frequentist consistency is the most basic statistical property one should demand from any statistical procedure. Our main results in Theorem 1 (and 2) establish this for the solutions (and the corresponding values) of (TP). An added advantage of our analysis is that it can be adapted to prove asymptotic consistency of the 'true' posterior distribution.

The problem (TP) models a gamut of problems in engineering and science, including optimal system design, inverse and uncertainty quantification problems within a Bayesian statistical framework. Collectively, we call these as *Bayesian optimal design* (BOD) problems, and $x \in \mathscr{X}$ is a 'design' or 'treatment' in an experiment. For consistency we will use 'design' as our terminology in this paper. In the next section, we start by illustrating (TP) on system design and experiment design problems.

2 ILLUSTRATIVE EXAMPLE PROBLEMS

We consider two disparate problem settings, highlighting the generality of (TP) and our forthcoming results.

2.1 Stochastic System Design

To illustrate the system design problem in (TP) with an example, we model a staffing problem in a multi-server Markovian queueing system. In this setting the decision maker (DM) has to decide the optimal number of servers, c, after observing inter-arrival and service time data. We assume that the parameters of the inter-arrival and service time distributions, denoted as λ and μ respectively, are unknown. Observe that λ and μ together form the system parameter $\xi = \{\lambda, \mu\}$ and the number of servers c is the decision variable. The goal of the DM is to find the optimal number of servers, such that the steady state probability that the customer waits for service is at most $\alpha \in (0,1)$. The DM is assumed to collect n realizations of the random vector $\mathscr{V} := \{T, S, E\}$, denoted $\mathbf{X}_n := \{\mathscr{V}_1, \ldots \mathscr{V}_n\}$, where T, S, and E are the random variables denoting the arrival epochs, service-start, and service-end times of each customer $i \in \{1, 2, \ldots n\}$ respectively. We also assume that the inter-arrival and service times are independent; that is, $T_i - T_{i-1}$ is independent of $E_i - S_i$ for each $i \geq 1$. The joint likelihood of the inter-arrival and service times for the n jobs is

$$P_\xi(\mathbf{X}_n) := \prod_{i=1}^{n} \lambda e^{-\lambda(T_i - T_{i-1})} \mu e^{-\mu(E_i - S_i)}.$$

Constraint functions: The DM chooses the number of servers c to maintain a constant measure of congestion, measured using the steady state probability that a typical customer waits for service: $1 - W_q(c, \lambda, \mu)$, where $W_q(c, \lambda, \mu)$ is the probability that the customer did not wait. A closed-form expression for $1 - W_q(c, \lambda, \mu)$ is known (Gross et al. 2008):

$$1 - W_q(c, \lambda, \mu) = \frac{r^c}{c!(1-\rho)} \Bigg/ \left(\frac{r^c}{c!(1-\rho)} + \sum_{t=0}^{c-1} \frac{r^t}{t!} \right),$$

where $r = \frac{\lambda}{\mu}$ and $\rho = \frac{r}{c}$ and $\rho < 1$ is the *traffic intensity*. For the queue to be stable it is necessary and sufficient that $\rho < 1$. Now, the smallest c that satisfies the conditions

$$(\alpha - \{1 - W_q(c,\lambda,\mu)\}) > 0 \text{ and } (c\mu - \lambda) > 0$$

is chosen. The corresponding constraint optimization problem is

$$\begin{aligned} \text{minimize} \quad & c \\ \text{s.t.} \quad & \pi((\alpha - \{1 - W_q(c,\lambda,\mu)\}) > 0, (c\mu - \lambda) > 0 | \mathbf{X}_n) \geq \beta. \end{aligned} \tag{TP-Q1}$$

Moreover, we can also consider minimizing weighted sum of number of servers and expected waiting time in queue $T_q(c,\lambda,\mu) := \frac{1}{\lambda} \frac{r^c \rho}{c!(1-\rho)^2} \Big/ \left(\frac{r^c}{c!(1-\rho)} + \sum_{t=0}^{c-1} \frac{r^t}{t!} \right)$, that is for $\{\zeta_1, \zeta_2\} \in (0,1) \times (0,1)$ and $\zeta_1 + \zeta_2 = 1$

$$\begin{aligned} \underset{c \geq 1}{\text{minimize}} \quad & \zeta_1 c + \zeta_2 \mathbb{E}_{\pi(\{\lambda,\mu\}|\mathbf{X}_n)}[T_q(c,\lambda,\mu)] \\ \text{s.t.} \quad & \pi((\alpha - \{1 - W_q(c,\lambda,\mu)\}) > 0, (c\mu - \lambda) > 0 | \mathbf{X}_n) \geq \beta. \end{aligned} \tag{TP-Q2}$$

This staffing problem and its variations have been well studied in the queueing literature; interested reader may refer to (Jaiswal et al. 2020; Gans et al. 2003; Aksin et al. 2009). Note that while this problem is particularly simple, we are also interested in significantly more general settings where the decision variable can be Banach space-valued.

2.2 Bayesian Inverse Problems

In inverse problems, the aim is to recover input ξ to any mathematical model $\mathscr{G}$ using observations X corrupted by noise. Mathematically, the observation X is modeled as $X = \mathscr{G}(\xi) + \eta$, where η is the observational noise, typically assumed to be Gaussian and independent of ξ. Inverse problems with complex mathematical models are prevalent in weather forecasting, oceanography, subsurface geophysics, and molecular dynamics (Stuart 2010). In most complex problems, ξ is a high (or possibly infinite) dimensional parameter and $\mathscr{G}$ is a linear but high (or infinite) dimensional operator such as a partial differential equation operator. The Bayesian approach is a popular method for solving inverse problems. Following Bayesian statistics, a prior is posited on ξ and using a sequence of observations $\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\}$, a posterior distribution on ξ, $\pi(\xi|\mathbf{X}_n)$, is computed using Bayes rule. Thereafter, the posterior distribution is used for statistical inference and/or parameter (model input) estimation.

We are interested in settings where the solution to the inverse problem is used to make a decision or to suggest an optimal design for a system. For instance, in a inverse problem for a weather forecasting model, a posterior distribution capturing the uncertainty in the weather predictions, could be used for optimally solving a unit-commitment decision making problem in power generation (Staffell and Pfenninger 2018). To be specific, let $\mathscr{G}(u)$ model the availability of renewable energy resources (solar/wind) over a spatial domain $\mathscr{D}$ and let $x \in \mathscr{X}$ be the the unit-commitment allocation decision variable over domain $\mathscr{D}$. Furthermore, denote the cost/risk associated with $\mathscr{G}(u)$ for a given x as $f(x, \mathscr{G}(u))$. The constraint function $g(x, \mathscr{G}(u))$ can be thought of as a measure of power generation capacity of a plant, and it must be bounded by the respective limiting capacity of a plant. Now, it is straightforward to observe that the above unit-commitment problem using the solution to a weather forecasting inverse problem can be expressed as (TP).

In the next section we first present a variational Bayesian approximation to (TP).

3 VARIATIONAL BAYES FOR CHANCE-CONSTRAINED OPTIMAL DESIGN

Bayesian statistics delineates natural principles to model uncertainty in parameter estimation, using observed data combined with prior knowledge. Let $\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\}$, be n independent and identically distributed

samples from the $\mathscr{F}$ measurable random vector X_0 with support in $\mathbb{R}^d$ on probability space $(\Omega, \mathscr{F}, P_\xi)$, with P_ξ as the associated probability measure with parameter $\xi \in \Theta$; let P_0 represent the measure of X_0. Recall that the posterior measure is defined as

$$\pi(\xi|\mathbf{X}_n) := \frac{\pi(\xi)P_\xi(\mathbf{X}_n)}{\int_\Theta \pi(\xi)P_\xi(\mathbf{X}_n)d\xi},$$

where $P_\xi(\mathbf{X}_n) \equiv \prod_{i=1}^n P_\xi(X_i)$ is the likelihood of observing $\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\}$ and $\pi(\xi)$ is the prior probability measure over the parameters. The prior measure encodes knowledge about the model space, and posterior measure is a succinct synthesis of this prior knowledge with observed data. There are the two significant challenges in solving (TP):

1. *Computing the posterior distribution:* While in some cases conjugate priors can be used, this is not acceptable in most problems, resulting in an intractable computation. As noted before, the posterior intractability is the common motivation for using VB or MCMC methods for approximate Bayesian inference.
2. *Convexity of the feasibility set:* Ideally, one should expect (TP) to be a convex program to take advantage of well established convex solvers. However, even if the posterior distribution is computable, to qualify (TP) as a convex program the feasibility set,

   $$\{x \in \mathscr{X} : \pi(g(x,\xi) \leq 0|\mathbf{X}_n) \geq \beta\},$$

   must be convex. It might be possible that this set is not convex even when the underlying constraint function $g(x,\xi)$ is (in x) and, therefore, finding a global optimum becomes challenging.

MCMC methods offer one way to do approximate Bayesian inference with asymptotic statistical guarantees. However, these guarantees are offset by issues like poor mixing, large variance and complex diagnostics in practical settings with finite computational budgets. Furthermore, in the MCMC setting, one must construct an empirical approximation to the chance constraint feasibility set in (TP). This approximation suffers from high variance and statistical bias.

To illustrate this, consider the following simple example of a chance-constraint feasibility set adapted from Peña-Ordieres et al. (2020). We plot in Figure 1 the following chance-constraint feasibility set

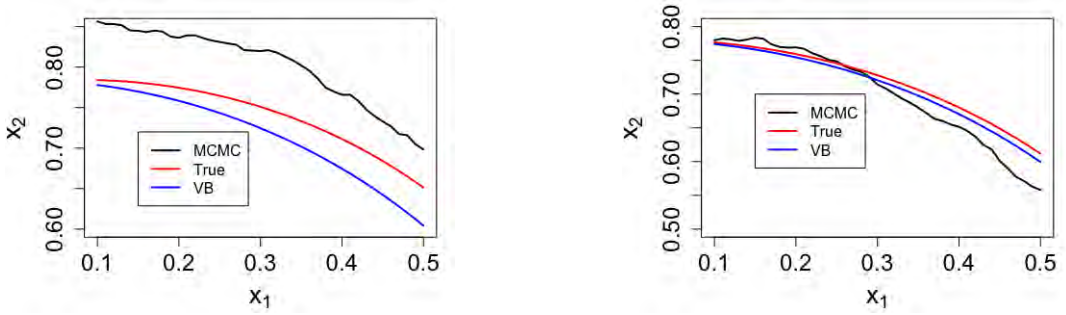
$$\left\{x \in R^2 : \mathscr{N}\left(\xi^T x - 1 \leq 0 | \mu = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma_A = [1, -0.1; -0.1, 1]\right) > \beta\right\},$$

$$\left\{x \in R^2 : \mathscr{N}\left(\xi^T x - 1 \leq 0 \middle| \mu = \begin{bmatrix}0\\0\end{bmatrix}, \Sigma_B = \begin{bmatrix}1 & -0.025\\-0.025 & 1\end{bmatrix}\right) > \beta\right\},$$

and its empirical approximator using 8000 MCMC samples (Metropolis-Hastings with a burn-in of 3000 samples) generated from the underlying correlated multivariate Gaussian distribution. Observe that the resulting MCMC approximate feasibility set is non-convex, while the VB approximation is convex.

Therefore, due to the posterior intractability of the feasible region when using MCMC sampling approaches, we propose to use Variational Bayes (VB) methods. Loosely put, VB methods trade bias for variance, and yield low (or zero) variance approximations to the chance constraint feasibility set. The crucial insight underlying VB is to approximate the intractable posterior $\pi(\xi|\mathbf{X}_n)$ with an element $q^*(\xi|\mathbf{X}_n)$ of a simpler *variational family* $\mathscr{Q}$ of candidate measures. The variational solution q^* is the element of $\mathscr{Q}$ that is 'closest' to $\pi(\xi|\mathbf{X}_n)$, where closeness is usually measured in the Kullback-Leibler (KL) sense. Thus,

$$q^*(\xi|\mathbf{X}_n) := \text{argmin}_{q\in\mathscr{Q}} \text{KL}(q(\theta)\|\pi(\xi|\mathbf{X}_n)). \tag{1}$$



(a) Σ_A (b) Σ_B

Figure 1: Feasible Region : True Distribution vs Monte Carlo Approximation (5000 samples) vs. VB (mean field approximation).

Using this, we approximate (TP) with,

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}[f(x,\xi)] \\ \text{s.t.} \quad & q^*\left(g(x,\xi) \leq 0 | \mathbf{X}_n\right) \geq \beta . \forall x \in \mathscr{X}, \end{aligned} \tag{VTP}$$

Choosing the approximation to the posterior distribution from a class of 'simple' distributions adresses the two critical problems associated with (TP). Besides the tractability of the posterior distribution, for instance, using the results in Proposition 4 of Farshbaf-Shaker et al. (2017) the choice of a log-concave family of distributions as the approximating family retains the convexity of the feasibility set, if the constraint functions are sufficiently regular; see Jaiswal et al. (2020). Note that the solution to (VTP) is necessarily biased. In the next section we rigorously prove that this solution is, nevertheless, asymptotically consistent in the large sample limit.

4 FREQUENTIST CONSISTENCY OF (VTP)

In this section, we establish the the frequentist consistency of the optimal solution set $\mathscr{S}^*_{VB}(\mathbf{X}_n)$ of (VTP) and show that it converges to the optimal solution set $\mathscr{S}^*$ of (TP) almost surely in P_0, the data-generating measure corresponding to parameter $\xi_0 \in \Theta$. We also show that the corresponding optimal values $V^*_{VB}(\mathbf{X}_n)$ converge to the optimal V^* of (TP). Our proof of asymptotic consistency uses techniques from the variational calculus in Banach spaces (Lucchetti and Wets 1993) and the consistency of the VB-approximate posterior distribution, which is proved under certain conditions on the prior distribution, likelihood model, and the variational approximation in (Wang and Blei 2018).

4.1 Definitions and Assumptions

We first collect a number of requisite definitions and assumptions.

Definition 1 (Weak and strong convergence in Banach Space) Let $\{x_k\}$ be a sequence in Banach space $\mathscr{X}$ equipped with norm denoted as $\|\cdot\|$. Denote $\mathscr{X}^*$ as the dual space of $\mathscr{X}$.

1. A sequence $\{x_k\}$ strongly converges to $x \in \mathscr{X}$, if

$$\lim_{k\to\infty} \|x_k - x\| = 0.$$

2. A sequence $\{x_k\}$ weakly converges to $x \in \mathscr{X}$, if

$$\forall \mu \in \mathscr{X}^*, \lim_{k\to\infty} \langle x_k, \mu \rangle = \langle x, \mu \rangle.$$

We denote δ_{ξ_0} as the Dirac delta measure concentrated at the parameter $\xi_0 \in \Theta$. Next, we first list down structural regularity conditions on the objective function $f(\cdot,\cdot)$.

Assumption 4.1. *We impose the following conditions on $f(x,\xi)$.*

1. *$f(x,\xi) : \mathscr{X} \times \Theta \mapsto \mathbb{R} \cup \{\infty\}$ and $f(\cdot,\cdot) > -\infty$. Also,*

$$dom(f) := \{(x,\xi) \in \mathscr{X} \times \Theta : f(x,\xi) < \infty\} \neq \emptyset.$$

   and $dom(f) = D \times \Theta$, with D a non-empty weakly closed (closed with respect to the weak topology) subset of $\mathscr{X}$.
2. *f is a random weakly lower semi-continuous function,*
   (a) for all $\xi \in \Theta$, the function $x \mapsto f(x,\xi)$ is weakly lower semi continuous, that is for every x_k converging weakly to x, $\liminf_{k\to\infty} f(x_k,\xi) \geq f(x,\xi)$.
   (b) for all $x \in \mathscr{X}$, $\xi \mapsto f(x,\xi)$ is $\mathscr{A}$ measurable, where $\mathscr{A}$ is the Borel σ-algebra generated by Θ.
3. *The function $x \mapsto f(x,\xi_0)$ is weakly continuous, that is for every x_k converging weakly to x, $\lim_{k\to\infty} f(x_k,\xi_0) = f(x,\xi_0)$.*
4. *For all $\alpha \in \mathbb{R}$, the (inf-)level sets of f:*

$$lev_\alpha(f) := \{(x,\xi) : f(x,\xi) \leq \alpha\} \subset D \times \Theta,$$

   are sequentially closed with respect to the product of the weak topology on $\mathscr{X}$ and the metric topology on Θ.
5. *f is bounded on bounded subsets of $D \times \Theta$.*
6. *Define $B_{\mathscr{X}}$ be a unit ball in $\mathscr{X}$ and for all $r \in \mathbb{R}^+$ and $\xi \in \Theta$, $w_r(\xi) := \inf\{f(x,\xi) : x \in rB_{\mathscr{X}}\}$. For $r \in \mathbb{R}^+$ there exists a family of functions $u_r : \Theta \to \mathbb{R} \cup \{\infty\}$ such that u_r is upper semi-continuous and $u_r \leq w_r$. There also exists a measurable function $h : \Theta \mapsto \mathbb{R} \cup \{\infty\}$ such that for all ξ, $h \leq u_r$ and $h(\xi_0) > -\infty$.*
7. *The strict (inf-)level sets of the functions $\{f(x,\cdot), x \in D\}$*

$$lev^<_\alpha f(x,\cdot) := \{\xi \in \Theta : f(x,\xi) < \alpha\}$$

   are $\delta_{\xi_0}-$continuity sets, that is there boundaries are null sets.
8. *Let $\mathscr{W} := \{f(x,\cdot), x \in D\}$. We assume that the probability measures $\mathscr{M} := \{\delta_{\xi_0}, q^*(\xi|\mathbf{X}_n), n \in \mathbb{N}\}$, are $\mathscr{W}$-tight, that is given any function $w \in \mathscr{W}$, to every $\varepsilon > 0$, there corresponds a bounded set B_ε such that for all $Q \in \mathscr{M}$,*

$$\int_{S \setminus B_\varepsilon} |w(s)| Q(ds) < \varepsilon.$$

The conditions listed above are adapted from (Lucchetti and Wets 1993). Notice that the first six assumptions above imposes structural conditions on $f(\cdot,\cdot)$ and are independent of $q^*(\xi|\mathbf{X}_n)$ (as defined in (1)) and its limit δ_{ξ_0} unlike last two conditions. Furthermore, we require that the constraint function $g(\cdot,\cdot)$ in (VTP) satisfies the following conditions.

Assumption 4.2. *We impose the following conditions on $g(x,\xi)$ for each $i \leq m$.*

1. *$g(x,\xi) : \mathscr{X} \times \Theta \mapsto \mathbb{R} \cup \{\infty\}$ and $f(,) > -\infty$. Also,*

$$dom(g) := \{(x,\xi) \in \mathscr{X} \times \Theta : f(x,\xi) < \infty\} \neq \emptyset.$$

   and $dom(f) = D \times \Theta$, with D a non-empty weakly closed (closed with respect to the weak topology) subset of $\mathscr{X}$.

2. *g is a Carathéodory function, that is*
   (a) for all $\xi \in \Theta$, the function $x \mapsto g(x,\xi)$ is weakly continuous, that is for every x_k converging weakly to x, $\lim_{k\to\infty} g(x_k,\xi) = g(x,\xi)$.
   (b) for all $x \in \mathscr{X}$, $\xi \mapsto g(x,\xi)$ is $\mathscr{A}$ measurable.
3. *for a given $x \in \mathscr{X}$, $\xi \mapsto g(x,\xi)$ is continuous with respect to the metric topology on Θ and ξ_0 lies in the interior of the set $\{\xi : g(x,\xi) \leq 0\}$.*

We also assume that the variational family $\mathscr{Q}$ satisfy the following property.

Assumption 4.3. *We assume that the probability measures $Q \in \mathscr{Q}$ are tight, that is to every $\varepsilon > 0$, there corresponds a bounded set B_ε such that for all $Q \in \mathscr{Q}$,*

$$\int_{S\setminus B_\varepsilon} Q(ds) < \varepsilon.$$

Notice that the above assumption can be easily satisfied by any family of light-tailed distributions. Some of the common examples of variational family are the family of Gaussian distributions, exponential family of distributions, or the family of factorized mean field distributions, that discards the correlation between components of parameter ξ (Blei et al. 2017).

Next we define hypo-convergence and epi-convergence of a sequence of functions $\{h_k(x)\}$ to $h(x)$ defined on Banach space.

Definition 2 (Mosco-Epi-convergence) A sequence of functions $\{h_k(x)\}$ Mosco-epi-converges to $h(x)$; that is M-ep$-\lim_{n\to\infty} h_k(x) = h(x)$, if

1. for every x_k converging weakly to x, $\liminf_{k\to\infty} h_k(x_k) \geq h(x)$, and
2. there exists a sequence x_k converging strongly to x, such that $\limsup_{k\to\infty} h_k(x_k) \leq h(x)$.

Definition 3 (Mosco-Hypo-convergence) A sequence of functions $\{h_k(x)\}$ Mosco-hypo-converges to $h(x)$; that is M-hypo$-\lim_{k\to\infty} h_k(x) = h(x)$, if

1. for every x_k converging weakly to x, $\limsup_{k\to\infty} h_k(x_k) \leq h(x)$, and
2. there exists a sequence x_k converging strongly to x, such that $\liminf_{k\to\infty} h_k(x_k) \geq h(x)$.

Finally, we define an indicator function as $\mathbb{I}_{(-\infty,0]}(t) := 1$ if $t \leq 0$ and 0 if $t > 0$.

4.2 Theoretical Results

Our first result is a straightforward application of the main result in Wang and Blei (2018). We show point-wise and Mosco-epi convergence of the expected unconstrained risk function $f(\cdot,\cdot)$ to the true risk/cost function as the number of samples increases, where the expectation is taken with respect to the VB posterior as defined in (1).

Lemma 4.1. *Under Assumption 4.1, we show that,*

1. *for each $x \in \mathscr{X}$, $\lim_{n\to\infty} \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}[f(x,\xi)] = f(x,\xi_0)$ $in-P_0$*
2. *and, M-ep$-\lim_{n\to\infty} \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}[f(x,\xi)] = f(x,\xi_0)$ $in-P_0$*

Proof. Recall from (Wang and Blei 2018) that the VB approximate posterior $q^*(\xi|\mathbf{X}_n)$ is consistent; that is for every $\eta > 0$,

$$\int_{\|\xi-\xi_0\|>\eta} q^*(\xi|\mathbf{X}_n)d\xi \to 0 \; P_0 - a.s. \text{ as } n \to \infty.$$

Therefore, $q^*(\xi|\mathbf{X}_n)$ converges weakly to δ_{ξ_0} in-P_0 probability (Ghosal and Van der Vaart 2017, Proposition 6.2). Now due to Assumption 4.1, both the claims are a direct consequence of the result in Lucchetti and Wets 1993, Theorem 13 and 14 . □

In the next lemma, we show that the indicator function defined using the constraint $g(\cdot,\cdot)$ satisfies almost similar structural properties as the cost/risk function $f(\cdot,\cdot)$.

Lemma 4.2. *Under Assumption 4.2, we show that $G(x,\xi) := -\mathbb{I}_{(-\infty,0]}(g(x,\xi))$ satisfies the following properties:*

1. *$G(,) > -\infty$ and $dom(G) := \{(x,\xi) \in \mathscr{X} \times \Theta : G(x,\xi) < \infty\} \neq \emptyset$. Also, $dom(G) = D \times \Theta$, with D a non-empty weakly closed (closed with respect to the weak topology) subset of $\mathscr{X}$.*
2. *$G(,)$ is a random weakly lower semi-continuous function.*
3. *For all $\alpha \in \mathbb{R}$, the (inf-)level sets of G:*

$$lev_\alpha(G) := \{(x,\xi) : G(x,\xi) \leq \alpha\} \subset D \times \Theta,$$

   are sequentially closed with respect to the product of the weak topology on $\mathscr{X}$ and the metric topology on Θ.
4. *$G(,)$ is bounded on the bounded subsets of $D \times \Theta$.*
5. *Define $B_{\mathscr{X}}$ be a unit ball in $\mathscr{X}$ and for all $r \in \mathbb{R}^+$ and $\xi \in \Theta$, $w_r(\xi) := \inf\{G(x,\xi) : x \in rB_{\mathscr{X}}\}$. For $r \in \mathbb{R}^+$ there exists a family of functions $u_r : \Theta \to \mathbb{R} \cup \{\infty\}$ such that u_r is upper semi-continuous and $u_r \leq w_r$. There also exists a measurable function $h : \Theta \mapsto \mathbb{R} \cup \{\infty\}$ such that for all ξ, $h \leq u_r$ and $h(\xi_0) > -\infty$.*
6. *The strict (inf-)level sets of the functions $\{G(x,\cdot), x \in D\}$*

$$lev^<_\alpha G(x,\cdot) := \{\xi \in \Theta : G(x,\xi) < \alpha\}$$

   are $\delta_{\xi_0}-$continuity sets, that is there boundaries are null sets.
7. *Let $\mathscr{W}_m := \{G(x,\cdot), x \in D\}$. We assume that the probability measures $\mathscr{M} := \{\delta_{\xi_0}, q^*(\xi|\mathbf{X}_n), n \in N\}$, are $\mathscr{W}_m$-tight, that is given any function $w \in \mathscr{W}_m$, to every $\varepsilon > 0$, there corresponds a bounded set B_ε such that for all $Q \in \mathscr{M}$,*

$$\int_{S \setminus B_\varepsilon} |w(s)| Q(ds) < \varepsilon.$$

Proof. We prove the properties of G in sequence:

1. The first property follows directly from Assumption 4.2 (1) and the definition of the indicator function. Also, note that $D = \mathscr{X}$, therefore it is non-empty and weakly closed.
2. Fix $\xi \in \Theta$. Since by Assumption 4.2 (a) each $g(x,\xi)$ is weakly continuous in x, therefore $\mathbb{I}_{(-\infty,0]}(g(x,\xi))$ is weakly upper-semicontinuous(USC) in x because $\mathbb{I}_{(-\infty,0]}(\cdot)$ is USC. Also, since the product of non-negative weakly USC functions are also weakly USC, it follows that $G(x,\xi) := -\mathbb{I}_{(-\infty,0]}(g(x,\xi))$ is weakly LSC. Moreover, for each $x \in \mathscr{X}$, using by assumption 4.2 (2-b) and the fact that $\mathbb{I}_{(-\infty,0]}(\cdot)$ is an indicator function, it also follows that $\xi \mapsto G(x,\xi) = -\mathbb{I}_{(-\infty,0]}(g(x,\xi))$ is $\mathscr{A}$ measurable. Therefore, $G(x,\xi)$ is a random weakly lower-semi continuous function.
3. For $\alpha \notin [-1,0)$, the result follows straightforwardly. Now fix $\alpha \in [-1,0)$. Since $D = \mathscr{X}$, it follows immediately that $lev_\alpha(G) \subset \mathscr{X} \times \Theta$. Next, fix $\xi \in \Theta$ and a sequence $\{x_k\} \in \mathscr{X}$ converging weakly to $x \in \mathscr{X}$. Since, for each $\xi \in \Theta$, $x \mapsto G(x,\xi)$ is weakly lower semi-continuous, it follows that if $x_k \in lev_\alpha(G)$ then,

$$G(x,\xi) \leq \liminf_{k\to\infty} G(x_k,\xi) \leq \alpha.$$

   Therefore, $x \in lev_\alpha(G)$. Using similar technique, we can use Assumption 4.2 (3) to first show that for a given $x \in \mathscr{X}$, $\xi \mapsto G(x,\xi)$ is lower-semi continuous and then show that it level set defined above is closed with respect to metric topology on Θ.
4. Since $G(x,\xi)$ is bounded on $\mathscr{X} \times \Theta$ therefore it also bounded on any of its subset and thus the fourth property follows immediately.

5. Since by definition $G(x,\xi)$ is bounded, it is easy to construct such functions. In particular, fix $u_r(\xi) = -1, \forall \xi \in \Theta$ and $h(\xi) = c, \forall \xi \in \Theta$ where $c \leq -1$, is some constant.
6. For $\alpha \notin (-1,0]$, the result follows immediately since for any $\alpha > 0$, $lev_\alpha^{\leq} G(x,\cdot) = \Theta$ and when $\alpha \leq -1$, $lev_\alpha^{\leq} G(x,\cdot) = \emptyset$. For $\alpha \in (-1,0]$, Assumption 4.2 (3) ensures that ξ_0 lie in the interior of the strict level set $lev_\alpha^{\leq} G(x,\cdot)$, therefore the boundary of this set is of measure zero with respect to δ_{ξ_0}, hence it is a $\delta_{\xi_0}-$ continuity set.
7. Since $G \in \{-1,0\}$, this property follows directly from Assumption 4.3.

□

Furthermore, in the next lemma, we show that the expectation of $G(\cdot,\cdot)$ with respect to the VB posterior converges both in pointwise and Mosco-hypo sense to the true as the number of samples increases.

Proposition 4.1. *We show that under Assumption 4.1, the Mosco-hypo and point-wise convergence of* $q^*\left(\mathbb{I}_{(-\infty,0]}(g(x,\xi))|\mathbf{X}_n\right)$ *to* $\mathbb{I}_{(-\infty,0]}(g(x,\xi_0))$ *in*$-P_0$ *as* $n \to \infty$*; that is*

$$1)\ \textit{M-hypo} - \lim_{n\to\infty} \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(x,\xi))\right] = \mathbb{I}_{(-\infty,0]}(g(x,\xi_0))\ \textit{in} - P_0,$$
$$2)\ \textit{for each } x \in \mathscr{X}, lim_{n\to\infty}\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(x,\xi))\right] = \mathbb{I}_{(-\infty,0]}(g(x,\xi_0))\ \textit{in} - P_0.$$

Proof. Recall $G(x,\xi) = -\mathbb{I}_{(-\infty,0]}(g(x,\xi))$. It follows directly from Lemma 4.2 that G satisfies all the necessary conditions to invoke Theorem 14 of Lucchetti and Wets (1993). Therefore,

$$\text{M-ep} - \lim_{n\to\infty} \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[G(x,\xi)\right] = G(x,\xi_0))\ \text{in} - P_0.$$

Since, by definition, for any $\xi \in \Theta$, $G(x,\xi) = -\mathbb{I}_{(-\infty,0]}(g(x,\xi))$, therefore the result follows immediately. Similarly, the second result follows immediately from Theorem 13 of Lucchetti and Wets (1993) □

Finally, as consequence of Lemma 4.2 and Proposition 4.1 we prove our main result. We show that the optimal value $V_{VB}^*(\mathbf{X}_n)$ and the optimal solution set $\mathscr{S}_{VB}^*(\mathbf{X}_n)$ computed by solving (VTP) converges (in P_0-probability) to the true optimal value and solution set respectively as the number of samples increases to infinity.

Theorem 1 We show that $V_{VB}^*(\mathbf{X}_n) \to V^*$ in $-P_0$. and $\mathbb{D}(\mathscr{S}_{VB}^*(\mathbf{X}_n), \mathscr{S}^*) \to 0$ in $-P_0$. as $n \to \infty$, where $\mathbb{D}(A,B) := \sup_{x\in A}\inf_{y\in B}\sup_{\|\mu\|_*\leq 1}\langle x-y,\mu\rangle, \forall \mu \in \mathscr{X}^*$ is the distance between two sets A and B in $\mathscr{X}$.

Proof. Recall $\mathscr{S}_{VB}^*(\mathbf{X}_n)$ is the solution of (VTP) and $\mathscr{S}^*$ is the solution of (TP).

Now observe that, since both $\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(x,\xi))\right]$ and $\mathbb{I}_{(-\infty,0]}(g(x,\xi_0))$ are weakly upper- semi-continuous their corresponding super-level sets are weakly closed; and if $\mathscr{X}$ is weakly bounded than the corresponding feasible sets are also weakly compact. Also, if the the corresponding feasibility sets are non-empty then the corresponding optimal sets $\mathscr{S}_{VB}^*(\mathbf{X}_n)$ and $\mathscr{S}^*$ are also non-empty.

Next let us assume that there exists a true solution x^* of (TP) which lies in the interior of $\mathscr{X}$, that is for any $\varepsilon > 0$, there is $x \in \mathscr{X}$ such that for any $\mu \in \mathscr{X}^*$ $\sup_{\|\mu\|_*\leq 1}\langle x-x^*,\mu\rangle < \varepsilon$ and $g(x,\xi_0) \leq 0$. It implies that there exists a sequence $\{x_k\} \subset \mathscr{X}$ such that x_k converges weakly x^* as $k \to \infty$ and $g(x_k,\xi_0) \leq 0$ for all $k \geq 1$. Now fix $x \in \mathscr{X}$ such that $g(x,\xi_0) \leq 0$. Since, due to Lemma 4.1 (2) $\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(x,\xi))\right]$ converges pointwise to $\mathbb{I}_{(-\infty,0]}(g(x,\xi_0))$ in $-P_0$, therefore there exists an n_0 such that for all $n \geq n_0$, we have $\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(x,\xi))\right] \geq \beta$, since $\beta \in (0,1)$. Hence for all $n \geq n_0$, x is a feasible solution of (VTP) and therefore $\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}[f(x,\xi)] \geq V_{VB}^*(\mathbf{X}_n)$. Taking lim sup on either sides, we obtain

$$\limsup_{n\to\infty}\ V_{VB}^*(\mathbf{X}_n) \leq \limsup_{n\to\infty}\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}[f(x,\xi)] = f(x,\xi_0)\ \text{in} - P_0,$$

where the last inequality follows from Lemma 4.1 (1). Now, since x can be chosen arbitrarily close to x^* in weak topology, it follows from weak continuity (see Assumption 4.1 (3)) of $x \mapsto f(x,\xi_0)$ that

$$\limsup_{n\to\infty} V^*_{VB}(\mathbf{X}_n) \leq f(x^*,\xi_0) = V^* \text{ in} - P_0. \tag{2}$$

Next, let $\hat{x}_n \in \mathscr{S}^*_{VB}$; that is $\hat{x}_n \in \mathscr{X}$, $\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(\hat{x}_n,\xi))\right] \geq \beta$ and $V^*_{VB}(\mathbf{X}_n) = \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}[f(\hat{x}_n,\xi)]$. Since $\mathscr{X}$ is weakly compact, we assume that $\hat{x}_n$ converges weakly to x^* in $-P_0$. Due to Lemma 4.1 (1), $\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(x,\xi))\right]$ Mosco-hypo-converges to $\mathbb{I}_{(-\infty,0]}(g(x,\xi_0))$ in$-P_0$ as $n\to\infty$, therefore we have

$$\limsup_{n\to\infty} \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(\hat{x}_n,\xi))\right] \leq \mathbb{I}_{(-\infty,0]}(g(x^*,\xi_0)). \tag{3}$$

Now using the fact that $\mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(\hat{x}_n,\xi))\right] \geq \beta$ for every $n \geq 1$, it follows from (3) that x^* is a feasible point of (TP), since $\limsup_{n\to\infty} \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}\left[\mathbb{I}_{(-\infty,0]}(g(\hat{x}_n,\xi))\right] \geq \beta$ implies $\mathbb{I}_{(-\infty,0]}(g(x^*,\xi_0)) \geq \beta$ and $\beta \in (0,1)$. Therefore, it follows that $f(x^*,\xi_0) \geq V^*$. Since, due to Lemma 4.1 (2), $\liminf_{n\to\infty} \mathbb{E}_{q^*(\xi|\mathbf{X}_n)}[f(\hat{x}_n,\xi)] \geq f(x^*,\xi_0)$ in $-P_0$, it follows that

$$\liminf_{n\to\infty} V^*_{VB}(\mathbf{X}_n) \geq V^* \text{ in} - P_0. \tag{4}$$

Hence, it follows from (2) and (4) that $V^*_{VB}(\mathbf{X}_n) \to V^*$ in $-P_0$ and it also follows that x^* is the true solution of (TP), therefore $\mathbb{D}(\mathscr{S}^*_{VB}(\mathbf{X}_n),\mathscr{S}^*) \to 0$ in $-P_0$. □

In the next result, we show that the solution obtained in (VTP) are feasible with high probability. Let us define the set where the true constraint is satisfied as $F_0 := \{x \in \mathscr{X} : \{g(x,\xi_0) \leq 0\},\}$, and VB-approximate feasibility set is denoted as $\hat{F}_{VB}(\mathbf{X}_n) := \{x \in \mathscr{X} : q^*(g(x,\xi) \leq 0|\mathbf{X}_n) \geq \beta\}$. We prove the next result using the convergence rate results for VB approximation in (Zhang and Gao 2020).

Theorem 2 We show that if $x \in \mathscr{X}\backslash F_0$, then there exists a constant $C > 0$, such that $P_0[x \in \hat{F}_{VB}(\mathbf{X}_n)] \leq \frac{C}{\beta}(\varepsilon_n^2 + \eta_n^2)$, where $\varepsilon_n^2 \to 0$ as $n\to\infty$ and $\eta_n^2 := \frac{1}{n}\inf_{q\in\mathscr{Q}} \mathbb{E}_{P_0}\left[\int_\Theta q(\xi)\log\frac{q(\xi)}{\pi(\xi|\mathbf{X}_n)}d\xi\right]$.

Proof. Using Markov's inequality observe that for any $x \in \mathscr{X}$,

$$P_0[q^*(g(x,\xi) \leq 0|\mathbf{X}_n) \geq \beta] \leq \frac{1}{\beta}\mathbb{E}_0[q^*(\{g(x,\xi) \leq 0\}|\mathbf{X}_n)]. \tag{5}$$

Since $x \in \mathscr{X}\backslash F_0$ implies that $x \in \{g(x,\xi_0) > 0\}$, it follows that

$$\{g(x,\xi) \leq 0\} \subseteq \{g(x,\xi) < g(x,\xi_0)\}.$$

Therefore, for all $x \in \mathscr{X}\backslash F_0$ it follows from (5) that

$$P_0[q^*(g(x,\xi) \leq 0|\mathbf{X}_n) \geq \beta] \leq \frac{1}{\beta}\mathbb{E}_0[q^*(\{g(x,\xi) < g(x,\xi_0)\}|\mathbf{X}_n)]. \tag{6}$$

Now using Theorem 2.1 in (Zhang and Gao 2020), it follows that if

$$L_n(\xi,\xi_0) := n\sup_{x\in\mathscr{X}} \mathbb{I}_{(0,\infty)}(g(x,\xi_0) - g(x,\xi))$$

satisfies assumption (C1) in Zhang and Gao (2020), then there exists a constant C such that

$$\mathbb{E}_0[q^*(\{g(x,\xi) < g(x,\xi_0)\}|\mathbf{X}_n)] \leq C(\varepsilon_n^2 + \eta_n^2),$$

where $\eta_n^2 := \frac{1}{n}\inf_{q\in\mathscr{Q}} \mathbb{E}_{P_0}\left[\int_\Theta q(\xi)\log\frac{q(\xi)}{\pi(\xi|\mathbf{X}_n)}d\xi\right]$. Now observe that, using the above result in (6), the result follows immediately. □

REFERENCES

Aksin, Z., M. Armony, and V. Mehrotra. 2009, January. "The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research". *Production and Operations Management* 16(6):665–688.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. 2017, April. "Variational Inference: A Review for Statisticians". *Journal of the American Statistical Association* 112(518):859–877.

Farshbaf-Shaker, M. H., R. Henrion, and D. Hömberg. 2017, October. "Properties of Chance Constraints in Infinite Dimensions with an Application to PDE Constrained Optimization". *Set-Valued and Variational Analysis* 26(4):821–841.

Gans, N., G. Koole, and A. Mandelbaum. 2003, April. "Telephone Call Centers: Tutorial, Review, and Research Prospects". *Manufacturing & Service Operations Management* 5(2):79–141.

Ghosal, S., and A. Van der Vaart. 2017. *Fundamentals of Nonparametric Bayesian Inference*, Volume 44. Cambridge: Cambridge University Press.

Gross, D., J. F. Shortie, J. M. Thompson, and C. M. Harris. 2008, July. *Simple Markovian Queueing Models*. 4th ed. New Jersey: Wiley.

Jaiswal, P., H. Honnappa, and V. A. Rao. 2020, December. "Variational Bayesian Methods for Stochastically Constrained System Design Problems". In *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, edited by C. Zhang, F. Ruiz, T. Bui, A. B. Dieng, and D. Liang, Volume 118, 1–12: Proceedings of Machine Learning Research.

Lucchetti, R., and R. J.-B. Wets. 1993, January. "Convergence of Minima of Integral Functionals, with Applications to Optimal Control and Stochastic Optimization". *Statistics & Risk Modeling* 11(1):69–84.

Peña-Ordieres, A., J. R. Luedtke, and A. Wächter. 2020, January. "Solving Chance-Constrained Problems via a Smooth Sample-Based Nonlinear Approximation". *SIAM Journal on Optimization* 30(3):2221–2250.

Staffell, I., and S. Pfenninger. 2018, February. "The Increasing Impact of Weather on Electricity Supply and Demand". *Energy* 145:65–78.

Stuart, A. M. 2010, May. "Inverse Problems: A Bayesian Perspective". *Acta Numerica* 19:451–559.

van der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Wang, Y., and D. M. Blei. 2018, August. "Frequentist Consistency of Variational Bayes". *Journal of the American Statistical Association* 114(527):1147–1161.

Zhang, F., and C. Gao. 2020, August. "Convergence Rates of Variational Posterior Distributions". *Annals of Statistics* 48(4):2180–2207.

AUTHOR BIOGRAPHIES

PRATEEK JAISWAL is a Ph.D. candidate in the School of Industrial Engineering at Purdue University. His research interests are in machine learning and stochastic optimization. His e-mail address is jaiswalp@purdue.edu.

HARSHA HONNAPPA is an assistant professor in the School of Industrial Engineering at Purdue University. His research interests are in applied probability, game theory, and machine learning. He is a member of INFORMS, IEEE, and SIAM, and serves as an associate editor for Operations Research and Operations Research Letters. His email address is honnappa@purdue.edu. His website is https://engineering.purdue.edu/SSL.