2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–12

Variational Bayesian Methods for Stochastically Constrained System Design Problems

Prateek Jaiswal* JAISWALP@PURDUE.EDU
Harsh Honnappa HONNAPPA@PURDUE.EDU
School of Industrial Engineering, Purdue University, West Lafayette, IN, USA
Vinayak A. Rao VARAO@PURDUE.EDU
Department of Statisitcs, Purdue University, West Lafayette, IN, USA

Abstract

We study system design problems stated as parameterized stochastic programs with a chance-constraint set. We adopt a Bayesian approach that requires the computation of a posterior predictive integral which is usually intractable. In addition, for the problem to be a well-defined convex program, we must retain the convexity of the feasible set. Consequently, we propose a variational Bayes-based method to approximately compute the posterior predictive integral that ensures tractability and retains the convexity of the feasible set. Under certain regularity conditions, we also show that the solution set obtained using variational Bayes converges to the true solution set as the number of observations tends to infinity. We also provide bounds on the probability of qualifying a true infeasible point (with respect to the true constraints) as feasible under the VB approximation for a given number of samples.

1. Introduction

A general system design problem can be formally stated as the following constraint optimization problem

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}, \boldsymbol{\xi}) \\ \text{s.t.} \quad & g_i(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0, \ i \in \{1, 2, 3, \ldots, m\}, \end{aligned} \tag{TP}$$

where $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^p$ is the input/control vector in a convex set $\mathcal{X}$ and $\boldsymbol{\xi} \in \Theta \subseteq \mathbb{R}^q$ is the system parameter vector. The function $f(\boldsymbol{x}, \boldsymbol{\xi}) : \mathcal{X} \times \Theta \mapsto \mathbb{R}$ encodes the cost/risk associated with the given values of parameter and control variable $\boldsymbol{\xi}$ and $\boldsymbol{x}$ respectively. Similarly, the functions $g_i(\boldsymbol{x}, \boldsymbol{\xi}) : \mathcal{X} \times \Theta \mapsto \mathbb{R}$ define the constraints on $\boldsymbol{\xi}$ and $\boldsymbol{x}$. Under certain regularity conditions on the cost and the constraint functions, and for a given value of the true system parameter $\boldsymbol{\xi}_0$, the problem (TP) at $\boldsymbol{\xi} = \boldsymbol{\xi}_0$ can be solved to obtain the optimal control vector $\boldsymbol{x}^*$. In practice, the true system parameters are unknown and these parameters must be estimated using observed data.

In this paper, we take a Bayesian approach and model the uncertainty over the parameters $\boldsymbol{\xi}$ by computing a posterior distribution $\pi(\boldsymbol{\xi}|\mathbf{X}_n)$ for a given prior distribution $\pi(\boldsymbol{\xi})$ and the likelihood $P_{\boldsymbol{\xi}}(\mathbf{X}_n)$ of observing data $\mathbf{X}_n$. We approximate the true problem (TP),

* Corresponding author

© P. Jaiswal, H. Honnappa & V.A. Rao.

using the posterior distribution, with the following joint chance-constrained problem:

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{\pi(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\boldsymbol{x},\boldsymbol{\xi})] \\ \text{s.t.} \quad & \pi\left(g_i(\boldsymbol{x},\boldsymbol{\xi}) \leq 0,\ i \in \{1,2,3,\ldots,m\}|\mathbf{X}_n\right) \geq \beta, \forall \boldsymbol{x} \in \mathcal{X}, \end{aligned} \tag{BJCCP}$$

where $\beta \in (0,1)$ is the specified confidence level desired by the decision maker (DM) based on the requirement, usually $\beta > \frac{1}{2}$. We provide a supporting example (see Appendix B) to motivate the chance-constraint formulation as opposed to using expectations, in which case the constraints are only satisfied on an average. The unconstrained version of the above problem has been studied as a special case in Jaiswal et al. (2019).

In practice, computing posterior distributions is challenging and mostly intractable, and is typically approximated using Markov Chain Monte Carlo (MCMC) or Variational Bayesian(VB) methods. MCMC methods has its own drawbacks like poor mixing, large variance, and complex diagnostics, which have been the usual motivation for using VB (Blei et al., 2017). Here, we provide another important motivation for using VB in the chance-constrained Bayesian inference setting. In particular, we present an example (motivated from Pena-Ordieres et al. (2019)) where a sampling based approach to approximate the chance-constraint convex feasibility set (constraint set) in (BJCCP), results in a non-convex approximation; whereas an appropriate VB approximation *retains its convexity*. Therefore, we approximate (BJCCP) using a VB approximate posterior $q^*(\boldsymbol{\xi}|\mathbf{X}_n)$ to $\pi(\boldsymbol{\xi}|\mathbf{X}_n)$ as:

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{q^*(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\boldsymbol{x},\boldsymbol{\xi})] \\ \text{s.t.} \quad & q^*\left(g_i(\boldsymbol{x},\boldsymbol{\xi}) \leq 0,\ i \in \{1,2,3,\ldots,m\}|\mathbf{X}_n\right) \geq \beta, \forall \boldsymbol{x} \in \mathcal{X}. \end{aligned} \tag{VBJCCP}$$

Under certain regularity conditions, we also show that the optimizers of (VBJCCP) are consistent with those of (TP). More precisely, we show that the solution set obtained in (VBJCCP) converges to the true solution set as the number of observations, n tends to infinity. We also provide bounds on the probability of qualifying a true infeasible point (with respect to the true constraints) as feasible under the VB approximation for a given number of samples. As part of the future work, we want to analyze the risk-sensitive VB approximation of the (BJCCP), where the risk is quantified as the deviation of the approximate feasibility set from the true.

2. Variational Bayes for Chance-Constrained System 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 $\mathcal{F}$ measurable random vector $X(\omega)$ with support $\Omega \subset \mathbb{R}^d$ on probability space $(\Omega, \mathcal{F}, P_{\boldsymbol{\xi}})$, with $P_{\boldsymbol{\xi}}$ as the associated probability measure, with parameter $\boldsymbol{\xi}$.

Using the posterior distribution $\pi(\boldsymbol{\xi}|\mathbf{X}_n)$, we approximate (TP) as a data-driven joint chance-constrained problem, stated formally as:

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{\pi(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\boldsymbol{x},\boldsymbol{\xi})] \\ \text{s.t.} \quad & \pi\left(g_i(\boldsymbol{x},\boldsymbol{\xi}) \leq 0,\ i \in \{1,2,3,\ldots,m\}|\mathbf{X}_n\right) \geq \beta, \forall \boldsymbol{x} \in \mathcal{X}. \end{aligned} \tag{BJCCP}$$

and $\beta \in (0,1)$ is the specified confidence level desired by the decision maker (DM) based on the requirement. These are the two significant challenges in solving (BJCCP):

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. The posterior intractability is the common motivation for using VB (Blei et al., 2017) and MCMC techniques for approximate Bayesian inference.

2. *Convexity of the feasibility set:* Ideally, one should expect (BJCCP) to be a convex program to take advantage of the well established convex solvers. But, even if the posterior distribution is computable, to qualify (BJCCP) as a convex program, the feasibility set,

$$\{\boldsymbol{x} \in \mathcal{X} : \pi\left(g_i(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0,\ i \in \{1, 2, 3, \ldots, m\} | \mathbf{X}_n\right) \geq \beta\} \tag{1}$$

must be convex. It might be possible that the above set is not convex even when the underlying constraint functions $g_i(\boldsymbol{x}, \boldsymbol{\xi}), i \in \{1, 2, \ldots m\}$ are so (in $\boldsymbol{x}$) and thus finding a global optimum becomes challenging (Prékopa, 1995).

Note that, if the constraint function has some specified structural regularity and the posterior distribution belongs to a certain class of distributions, then it can be shown that the feasibility set in (1) is convex. For instance, it can be shown that if the constraint functions $g_i(\boldsymbol{x}, \boldsymbol{\xi}), i \in \{1, 2, \ldots m\}$ are quasi-convex in $(\boldsymbol{x}, \boldsymbol{\xi})$ and the distribution is log-concave then the feasibility set in (BJCCP) is convex ((Shapiro et al., 2009, Chapter 4) and Prékopa (2003)). Also, Lagoa et al. (2005) showed that if the constraint function $g_i(\boldsymbol{x}, \boldsymbol{\xi})$ is of the form $\{\boldsymbol{a}^T\boldsymbol{x} \leq \boldsymbol{b}\}$, where $\boldsymbol{\xi} = (\boldsymbol{a}^T, \boldsymbol{b})^T$ and has a symmetric log-concave density then with $\beta > \frac{1}{2}$ the feasibility set in (BJCCP) is convex.

To address the posterior intractability, Monte Carlo (MC) methods offer one way to do approximate Bayesian inference with asymptotic guarantees. However, their asymptotic guarantees are offset by issues like poor mixing, large variance and complex diagnostics in practical settings with finite computational budgets. Apart from these common issues, there is another important reason due to which any sampling-based method can not be used directly to solve (BJCCP). Using the empirical approximation to the posterior distribution (constructed using the samples generated from MCMC algorithm) to approximate the chance-constraint feasibility set in (BJCCP), results in a non-convex feasibility set (Pena-Ordieres et al., 2019). To illustrate this, consider the following simple example (modified slightly) of a chance-constraint feasibility set from Pena-Ordieres et al. (2019). We plot in Figure 1(a) the following chance-constraint feasibility set

$$\left\{\boldsymbol{x} \in R^2 : \mathcal{N}\left(\boldsymbol{\xi}^T\boldsymbol{x} - 1 \leq 0 | \boldsymbol{\mu} = [0,0]^T, \boldsymbol{\Sigma}_A = [1, -0.1; -0.1, 1]\right) > \beta\right\}, \tag{2}$$

and its empirical approximator using 8000 MCMC (Metropolis-Hastings with 3000 burn-in samples) samples generated from the underlying correlated multivariate Gaussian distribution. We observe that the resulting MC approximate feasibility set is non-convex.

Therefore, due to the posterior intractability and the non-convexity of the feasible region when using sampling approaches, as an alternative, we propose to use Variational Bayes (VB) methods. The idea behind VB is to approximate the intractable posterior $\pi(\boldsymbol{\xi}|\mathbf{X}_n)$ with an element $q^*(\boldsymbol{\xi}|\mathbf{X}_n)$ of a simpler *variational family* $\mathcal{Q}$. Examples of $\mathcal{Q}$ include the

family of Gaussian distributions, delta functions, or the family of factorized 'mean-field' distributions that discard correlations between components of $\boldsymbol{\xi}$. The variational solution q^* is the element of $\mathcal{Q}$ that is 'closest' to $\pi(\boldsymbol{\xi}|\mathbf{X}_n)$, where closeness is usually measured in the Kullback-Leibler (KL) sense. Thus,

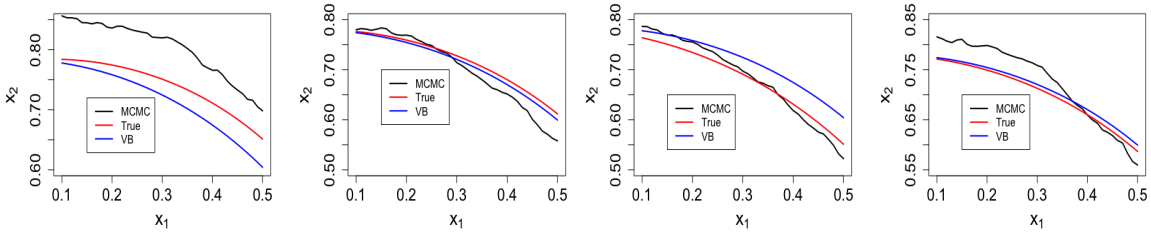
$$q^*(\boldsymbol{\xi}|\mathbf{X}_n) := \text{argmin}_{q\in\mathcal{Q}}\text{KL}(q(\theta)\|\pi(\boldsymbol{\xi}|\mathbf{X}_n)). \tag{3}$$

Using this, we approximate (BJCCP) with,

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{q^*(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\boldsymbol{x},\boldsymbol{\xi})] \\ \text{s.t.} \quad & q^*\left(g_i(\boldsymbol{x},\boldsymbol{\xi}) \le 0,\ i \in \{1,2,3,\ldots,m\}|\mathbf{X}_n\right) \ge \beta, \forall \boldsymbol{x} \in \mathcal{X}, \end{aligned} \tag{VBJCCP}$$

where β is the confidence level. Choosing the approximation to the posterior distribution from a class of 'simple' distributions would facilitate in addressing the two critical problems associated with (BJCCP). Besides the tractability of the posterior distribution, for instance, using the results in Prékopa (2003) and Lagoa et al. (2005) the choice of a log-concave family of distributions as the approximating family could retain the convexity of the feasibility set, if the constraint functions have certain structural regularity.

Next, we show that using the popular mean-field variational family to approximate the correlated multivariate Gaussian distribution in the same example in (2), we obtain a smooth and convex approximation to the (BJCCP) feasibility set. First, we compute mean-field approximation $q_A(\boldsymbol{\xi})$ and $q_B(\boldsymbol{\xi})$ of $\mathcal{N}\left(\boldsymbol{\xi}|\boldsymbol{\mu} = [0,0]^T, \boldsymbol{\Sigma}\right)$ for fours different covariance matrices $\boldsymbol{\Sigma}$, with fixed variance $\sigma_{11} = \sigma_{22} = 1$ but varying covariance $\sigma_{12} = \{-0.1, -0.025, 0.025, 0.1\}$. Then, we plot the respective approximate VB chance-constraint feasibility region in Figure 1. We observe that VB approximation provides a smooth convex approximation to the true feasibility set, but it could be outside the true feasibility region if the ξ_1 and ξ_2 are positively correlated.



(a) $\sigma_{12}^A = -0.1$ (b) $\sigma_{12}^B = -0.025$ (c) $\sigma_{12}^C = 0.025$ (d) $\sigma_{12}^D = 0.1$

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

2.1. Theoretical properties of (VBJCCP)

In this section, we establish theoretical guarantees on the approximate optimal solution set $\mathcal{S}^*_{VB}(\mathbf{X}_n)$ obtained using the VB approximation and show that it converges to the optimal solution set $\mathcal{S}^*$ of (TP) almost surely in P_0. We show similar result for their corresponding optimal values $V^*_{VB}(\mathbf{X}_n)$ and V^*. The consistency of the approximate solution follows using techniques from the variational calculus 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). For brevity, we state the following results without any assumptions and proofs; it will be stated formally in Appendix C.

Proposition 1 *We show that* $V^*_{VB}(\mathbf{X}_n) \to V^*$ $P_0 - a.s.$ *and* $\mathbb{D}(\mathcal{S}^*_{VB}(\mathbf{X}_n), \mathcal{S}^*) \to 0$ $P_0 - a.s.$ *as* $n \to \infty$*, where* $\mathbb{D}(A, B) := \sup_{\boldsymbol{x}\in A} \inf_{\boldsymbol{y}\in B} \|\boldsymbol{x} - \boldsymbol{y}\|$*, is the distance between two sets* A *and* B*.*

In the next result, we show that the solution obtained in (VBJCCP) are feasible with high probability. Let us define the set where the true constraint $i \in \{1, 2, \ldots m\}$ is satisfied as $F^i_0 := \{\boldsymbol{x} \in \mathcal{X} : \{g_i(\boldsymbol{x}, \boldsymbol{\xi}_0) \leq 0\},\}$, and VB-approximate feasibility set is denoted as $\hat{F}_{VB}(\mathbf{X}_n) := \{\boldsymbol{x} \in \mathcal{X} : q^*\left(g_i(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0,\ i \in \{1, 2, 3, \ldots, m\}|\mathbf{X}_n\right) \geq \beta\}$. We prove the next result using the convergence rate results for VB approximation in Zhang and Gao (2019).

Proposition 2 *We show that if* $\boldsymbol{x} \in \mathcal{X}\backslash F^i_0$*, then there exists constant* $C_i > 0$ *for each* $i \in \{1, 2, \ldots m\}$*, such that* $\mathbb{P}_0[\boldsymbol{x} \in \hat{F}_{VB}(\mathbf{X}_n)] \leq \frac{C_i}{\beta}(\epsilon_n^2 + \eta_n^2)$*, where* $\epsilon_n^2 \to 0$ *as* $n \to \infty$ *and* $\eta_n^2 := \frac{1}{n} \inf_{q\in\mathcal{Q}} \mathbb{E}_{P_0}\left[\int_\Theta q(\boldsymbol{\xi}) \log \frac{q(\boldsymbol{\xi})}{\pi(\boldsymbol{\xi}|\mathbf{X}_n)} d\boldsymbol{\xi} d\boldsymbol{\xi}\right]$.

References

Zeynep Aksin, Mor Armony, and Vijay Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, January 2009. doi: 10.1111/j.1937-5956.2007.tb00288.x. URL https://doi.org/10.1111/j.1937-5956.2007.tb00288.x.

Tevfik Aktekin and Tahir Ekin. Stochastic call center staffing with uncertain arrival, service and abandonment rates: A bayesian perspective. *Naval Research Logistics (NRL)*, 63(6): 460–478, September 2016. doi: 10.1002/nav.21716. URL https://doi.org/10.1002/nav.21716.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 2017.

Jitka Dupacova and Roger Wets. Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *The Annals of Statistics*, 16(4): 1517–1549, December 1988. doi: 10.1214/aos/1176351052. URL https://doi.org/10.1214/aos/1176351052.

Noah Gans, Ger Koole, and Avishai Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, April 2003. doi: 10.1287/msom.5.2.79.16071. URL https://doi.org/10.1287/msom.5.2.79.16071.

Donald Gross, John F. Shortie, James M. Thompson, and Carl M. Harris. *Simple Markovian Queueing Models*. Wiley, July 2008. doi: 10.1002/9781118625651.ch2. URL https://doi.org/10.1002/9781118625651.ch2.

L. Jeff Hong, Yi Yang, and Liwei Zhang. Sequential convex approximations to joint chance constrained programs: A monte carlo approach. *Operations Research*, 59(3):617–630, June 2011. doi: 10.1287/opre.1100.0910. URL https://doi.org/10.1287/opre.1100.0910.

Prateek Jaiswal, Harsha Honnappa, and Vinayak A Rao. Risk-sensitive variational bayes: Formulations and bounds. *arXiv preprint arXiv:1903.05220v3*, 2019.

Constantino M. Lagoa, Xiang Li, and Mario Sznaier. Probabilistically constrained linear programs and risk-adjusted controller design. *SIAM Journal on Optimization*, 15(3):938–951, January 2005. doi: 10.1137/s1052623403430099. URL https://doi.org/10.1137/s1052623403430099.

Alejandra Pena-Ordieres, James R. Luedtke, and Andreas Wachter. Solving chance-constrained problems via a smooth sample-based nonlinear approximation. 2019.

András Prékopa. *Stochastic Programming*. Springer Netherlands, 1995. doi: 10.1007/978-94-017-3087-7. URL https://doi.org/10.1007/978-94-017-3087-7.

András Prékopa. Probabilistic programming. *Handbooks in operations research and management science*, 10:267–351, 2003.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

Yixin Wang and David M. Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 0(0):1–15, 2018. doi: 10.1080/01621459.2018.1473776. URL https://doi.org/10.1080/01621459.2018.1473776.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *Annals of Statistics*, 2019. URL https://www.e-publications.org/ims/submission/AOS/user/submissionFile/34901?confirm=c41c6a61.

Appendix A. A System Design Problem

To illustrate the system design problem in (TP) with an example, we model a queueing system and show that the optimal staffing problem aptly fits into the above framework. Consider a staffing problem where the decision maker (DM) has to decide the optimal number of servers, c, after observing the arrival and service data in an $M/M/c$ queueing system; a queueing system where the inter-arrival times and service times are exponentially distributed (Markovian) with c number of servers, is denoted as $M/M/c$ queueing system. We assume that the rate parameters of the exponentially distributed inter-arrival and service times distribution are unknown and denoted as λ and μ respectively. Note that λ and μ, combined together form the system parameter $\xi = \{\lambda, \mu\}$ and the number of servers c is the control/input variable. The DM first uses a single server and collects data after the system reaches its 'steady state'. Since the DM observed congestion in the queues, he/she decides to employ more servers. The DM collects n realizations of the random vector $\mathcal{V} := \{T, S, E\}$,

denoted as $\mathbf{X}_n := \{\mathcal{V}_1, \dots \mathcal{V}_n\}$ where T, S, and E are the random variables denoting the arrival, service-start, and service-end time of each customer $i \in \{1, 2, \dots n\}$ respectively. We also assume that there is no time lag between the two successive states for any customer and 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 arrival and departure times for n customers is

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

Constraint functions: The DM chooses the number of servers c to maintain a constant measure of congestion. Congestion is usually measured as $1 - W_q(c, \lambda, \mu)$, where $W_q(c, \lambda, \mu)$ is the probability that the customer did not wait in the queue. The closed-form expression for $1 - W_q(c, \lambda, \mu)$ is known to be (see 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 $\rho = \frac{r}{c}$ with $\rho < 1$. ρ is also known as *traffic intensity* and for a stable queue $\rho < 1$. DM fixes α, which is maximum desired fraction of customers delayed in queue and the smallest c is chosen that satisfies:

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

Referring to the queueing literature, we will use the term the Quality of Service(QoS) constraint for the first constraint. The corresponding constraint optimization problem is

$$\begin{aligned} \text{minimize} \quad & c \\ \text{s.t.} \quad & (\alpha - \{1 - W_q(c, \xi)\}) > 0 \quad \text{(QoS)}, \\ & (c\mu - \lambda) > 0. \end{aligned} \tag{TP-Q}$$

The above staffing problem and its variations are well studied in the queueing literature; interested reader may refer to Gans et al. (2003) and Aksin et al. (2009).

Appendix B. Other Data-Driven Approaches to solve (TP)

Since the system parameters are unknown in practice, these are usually estimated using the observed data $\mathbf{X}_n$. The simplest approach could be to substitute the maximum likelihood estimates(MLE) $\hat{\xi}(\mathbf{X}_n)$ of the parameters ξ in the (TP) and solve the following approximate problem:

$$\begin{aligned} \text{minimize} \quad & f(x, \hat{\xi}) \\ \text{s.t.} \quad & g_i(x, \hat{\xi}) \leq 0, \ i \in \{1, 2, 3, \dots, m\}. \end{aligned} \tag{TP-MLE}$$

We solved the queueing staffing problem in Appendix A using the MLE approach on simulated data (n observations, with n in 50-400 in increments of 50) and computed the approximate optimal number of servers denoted as C^n_{MLE}. We repeated this experiment over

n	50	100	150	200	250	300	350	400
$\phi(C^n_{MLE})$	0.52	0.56	0.57	0.57	0.61	0.62	0.58	0.56

Table 1: Fraction of times C^n_{MLE} violates QoS constraint. $\lambda_0 = 16, \mu_0 = 4, \alpha = 0.37$.

100 sample paths and computed $\phi(C^n_{MLE})$, the fraction of experiments C^n_{MLE} violates the QoS constraint. Table 1 shows that the QoS constraint is violated in over 50% of the experiments.

It is anticipated that the MLE approach is unable to capture the uncertainty in parameter estimation therefore an alternative method is proposed using forecasting techniques. In this approach first, the uncertainty over the parameter estimation is captured by forecasting a probability distribution $P(\xi)$ over the system parameters and then the forecast distribution is used to solve the (TP) problem using one of the following two formulations:

- Average-Constraint(AC)

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_P[f(x,\xi)] \\ \text{s.t.} \quad & \mathbb{E}_P[g_i(x,\xi)] \leq 0, \ i \in \{1,2,3,\ldots,m\}, \end{aligned} \tag{TP-FAC}$$

- Chance-Constraint(CC)

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_P[f(x,\xi)] \\ \text{s.t.} \quad & P\left\{g_i(x,\xi) \leq 0, \ i \in \{1,2,3,\ldots,m\}\right\} > \beta, \end{aligned} \tag{TP-FCC}$$

  where $\beta \in (0,1)$ is the *confidence level*.

Now consider a simple example from Hong et al. (2011), where the true problem is to find $c^* = \min\{c : \xi - c \leq 0\}$. Since, ξ is unknown the DM uses data to forecast that $\xi \sim \mathcal{N}(\cdot|0,1)$ is normally distributed. Using the AC formulation, notice that the approximate optimal solution $c^*_A = \min\{c : \mathbb{E}_\xi[\xi] - c \leq 0\} = 0$ and $\mathbb{P}_{\boldsymbol{\xi}}\{\xi \geq c^*_A\} = 0.5$. The above simple example shows that AC optimal solution could violate the constraint 50% of the times. On the other hand, CC formulation enables the DM to ensure that the approximate optimal solution satisfy the constraints with higher confidence by setting a higher confidence level(β). In forecasting approach, the DM needs to forecast each time the new data is collected. We propose a principled data-driven approach using Bayesian methods, wherein we combine forecasting and optimization. A similar approach has also been discussed in Aktekin and Ekin (2016) to solve the $M/M/c$ staffing problem with abandonment, but crucially relies on the availability of conjugate priors.

Appendix C. Proofs

C.1. Proof of Proposition 1

Assumption C.1 *We assume that the function $f(\boldsymbol{x},\boldsymbol{\xi})$ and $g_i(\boldsymbol{x},\boldsymbol{\xi}), \forall i \in \{1,2,\ldots m\}$ are Carathéodory functions; that is $f(\boldsymbol{x},\cdot)$ and $g_i(\boldsymbol{x},\cdot)$ are measurable for every $\boldsymbol{x} \in \mathcal{X}$, and $f(\cdot,\boldsymbol{\xi})$ and $g(\cdot,\boldsymbol{\xi})$ are continuous for almost every $\boldsymbol{\xi} \in \Theta$. We also assume that $f(\cdot,\boldsymbol{\xi})$ is locally Lipschitz continuous in $\boldsymbol{x}$ for almost every $\boldsymbol{\xi} \in \Theta$ and $f(\boldsymbol{x},\cdot)$ is uniformly integrable with respect to any $q \in \mathcal{Q}$, the variational family.*

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

Lemma 3 *We show that for each $\boldsymbol{x} \in \mathcal{X}$*

$$\lim_{n\to\infty} q^*\left(\prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))|\mathbf{X}_n\right) = \prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}_0)) \ P_0 - a.s.$$

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

$$\lim_{n\to\infty} \int_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|>\eta} q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi} = 0 \ P_0 - a.s. \tag{4}$$

Observe that for any $\boldsymbol{x} \in \mathcal{X}$ and $\eta > 0$,

$$\begin{aligned} q^*\left(\prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))|\mathbf{X}_n\right) &= \int_{\Theta} \prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi} \\ &= \int_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|>\eta} \prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi} + \int_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\leq\eta} \prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi}. \end{aligned} \tag{5}$$

Observe that, the result in (4) combined with the fact that the first term in (5) is always positive and bounded, implies that $\lim_{n\to\infty} \int_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|>\eta} \prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi} = 0 \ P_0 - a.s.$ Now taking limits on either side of (5), we have

$$\begin{aligned} \lim_{n\to\infty} q^*\left(\prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))|\mathbf{X}_n\right) &= \lim_{n\to\infty} \int_{\|\boldsymbol{\xi}-\boldsymbol{\xi}_0\|\leq\eta} \prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi} \ P_0 - a.s, \\ &= \prod_{i=1}^{m} \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}_0)) \ P_0 - a.s \end{aligned} \tag{6}$$

and the lemma follows. ■

Next we define hypo-convergence and epi-convergence of a sequence of function $\{h_k(\boldsymbol{x})\}$ to $h(\boldsymbol{x})$.

Definition 4 (Hypo-convergence) *A sequence of functions $\{h_k(\boldsymbol{x})\}$ hypo-converges to $h(\boldsymbol{x})$; that is $hypo - \lim_{n\to\infty} h_k(\boldsymbol{x}) = h(\boldsymbol{x})$, if*

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

Definition 5 (Epi-convergence) *A sequence of functions $\{h_k(\boldsymbol{x})\}$ epi-converges to $h(\boldsymbol{x})$; that is $ep - \lim_{n\to\infty} h_k(\boldsymbol{x}) = h(\boldsymbol{x})$, if*

1. *for every $\boldsymbol{x}_k \to \boldsymbol{x}$, $\liminf_{k\to\infty} h_k(\boldsymbol{x}_k) \geq h(\boldsymbol{x})$, and*

2. *there exists a sequence* $\boldsymbol{x}_k \to \boldsymbol{x}$*, such that* $\limsup_{k\to\infty} h_k(\boldsymbol{x}_k) \leq h(\boldsymbol{x})$.

Lemma 6 *Under Assumption C.1, we show that,*

1. *for each* $\boldsymbol{x} \in \mathcal{X}$, $\lim_{n\to\infty} \mathbb{E}_{q^*(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\boldsymbol{x},\boldsymbol{\xi})] = f(\boldsymbol{x},\boldsymbol{\xi}_0)\ P_0 - a.s.$
2. *and,* $ep - \lim_{n\to\infty} \mathbb{E}_{q^*(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\boldsymbol{x}_n,\boldsymbol{\xi})] = f(\boldsymbol{x}_0,\boldsymbol{\xi}_0)\ P_0 - a.s.$

Proof Due to Assumption C.1, both the results above are a direct consequence of the result in Dupacova and Wets (1988, Theorem 3.7). ■

Lemma 7 *We show that under Assumption C.1,* $q^*\left(\prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))|\mathbf{X}_n\right)$ *hypo-converges to* $\prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}_0))\ P_0 - a.s$ *as* $n \to \infty$*; that is*

$$hypo - \lim_{n\to\infty} q^*\left(\prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))|\mathbf{X}_n\right) = \prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}_0))\ P_0 - a.s. \tag{7}$$

Proof Since by Assumption C.1 each $g_i(\boldsymbol{x},\boldsymbol{\xi}_0)$ is continuous in $\boldsymbol{x}$, therefore $\mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}_0))$ is upper-semicontinuous(USC) in $\boldsymbol{x}$ because $\mathbb{I}_{(-\infty,0]}(\cdot)$ is USC. Also, since the product of non-negative USC functions are also USC, it follows that $\prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}_0))$ is USC. Similarly, since by assumption $g_i(\boldsymbol{x},\boldsymbol{\xi})$ is *Carathéodory* function, therefore $\prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))$ is a random upper-semicontinuous function (Dupacova and Wets, 1988). Now, using the reverse Fatou's Lemma, for any $\boldsymbol{x}_k \to \boldsymbol{x}_0$

$$\begin{aligned}\limsup_{\boldsymbol{x}_k\to\boldsymbol{x}_0} \int_\Theta \prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi} &\leq \int_\Theta \limsup_{\boldsymbol{x}_k\to\boldsymbol{x}_0} \prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi} \\ &\leq \int_\Theta \prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x}_0,\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi},\end{aligned} \tag{8}$$

therefore $\int_\Theta \prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))q^*(\boldsymbol{\xi}|\mathbf{X}_n)d\boldsymbol{\xi}$ is also upper-semicontinuous in $\boldsymbol{x}$. Also, since $-\prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi})) \in \{-1,0\}$ is a bounded random lower-semicontinuous function and $q^*(\boldsymbol{\xi}|\mathbf{X}_n) \Rightarrow \delta_{\boldsymbol{\xi}_0}\ P_0 - a.s$ (Wang and Blei, 2018), Theorem 3.7 in Dupacova and Wets (1988) implies that

$$\text{hypo} - \lim_{n\to\infty} q^*\left(\prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}))|\mathbf{X}_n\right) = \prod_{i=1}^m \mathbb{I}_{(-\infty,0]}(g_i(\boldsymbol{x},\boldsymbol{\xi}_0))\ P_0 - a.s; \tag{9}$$

and the result follows. ■

Proof [Proof of Proposition 1] We will first show that the assertion of the theorem is true for $m = 1$. Recall $\mathcal{S}^*_{VB}(\mathbf{X}_n)$ is the solution of (VBJCCP): and $\mathcal{S}^*$ is the solution of (TP).

Now observe that, since both $q^*\left(g(\boldsymbol{x},\boldsymbol{\xi}) \leq 0|\mathbf{X}_n\right)$ and $\mathbb{I}_{(-\infty,0]}(g(\boldsymbol{x},\boldsymbol{\xi}_0))$ are upper- semi-continuous their corresponding super-level sets are closed; and if $\mathcal{X}$ is bounded than the

corresponding feasible sets are also compact. Also, if the the corresponding feasibility sets are non-empty then the corresponding optimal sets $\mathcal{S}^*_{VB}(\mathbf{X}_n)$ and $\mathcal{S}^*$ are also non-empty.

Next let us assume that there exists a true solution $\boldsymbol{x}^*$ of (TP) which lies in the interior of $\mathcal{X}$, that is for any $\epsilon > 0$, there is $\boldsymbol{x} \in \mathcal{X}$ such that $\|\boldsymbol{x}-\boldsymbol{x}^*\| < \epsilon$ and $g(\boldsymbol{x}, \boldsymbol{\xi}_0) \leq 0$. It implies that there exists a sequence $\{\boldsymbol{x}_k\} \subset \mathcal{X}$ such that $\boldsymbol{x}_k \to \boldsymbol{x}^*$ as $k \to \infty$ and $g(\boldsymbol{x}_k, \boldsymbol{\xi}_0) \leq 0$ for all $k \geq 1$. Now fix $\boldsymbol{x} \in \mathcal{X}$ such that $g(\boldsymbol{x}, \boldsymbol{\xi}_0) \leq 0$. Since, due to our result in Lemma 3 $q^*\left(g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0 | \mathbf{X}_n\right)$ converges pointwise to $\mathbb{I}_{(-\infty,0]}(g(\boldsymbol{x}, \boldsymbol{\xi}_0))$ $P_0 - a.s$, therefore there exists an n_0 such that for all $n \geq n_0$, we have $q^*\left(g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0 | \mathbf{X}_n\right) \geq \beta$. Hence for all $n \geq n_0$, $\boldsymbol{x}$ is a feasible solution of (VBJCCP) and therefore $\mathbb{E}_{q^*(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\boldsymbol{x}, \boldsymbol{\xi})] \geq V^*_{VB}(\mathbf{X}_n)$. Taking lim sup on either sides, we obtain

$$\limsup_{n\to\infty} V^*_{VB}(\mathbf{X}_n) \leq \limsup_{n\to\infty} \mathbb{E}_{q^*(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\boldsymbol{x}, \boldsymbol{\xi})] = f(\boldsymbol{x}, \boldsymbol{\xi}_0) \ P_0 - a.s,$$

where the last inequality follows from Lemma 6 (1). Now, since $\boldsymbol{x}$ can be chosen arbitrarily close to $\boldsymbol{x}^*$, it follows that

$$\limsup_{n\to\infty} V^*_{VB}(\mathbf{X}_n) \leq f(\boldsymbol{x}^*, \boldsymbol{\xi}_0) = V^* \ P_0 - a.s. \tag{10}$$

Next, let $\hat{\boldsymbol{x}}_n \in \mathcal{S}^*_{VB}$; that is $\hat{\boldsymbol{x}}_n \in \mathcal{X}$, $q^*\left(g(\hat{\boldsymbol{x}}_n, \boldsymbol{\xi}) \leq 0 | \mathbf{X}_n\right) \geq \beta$ and $V^*_{VB}(\mathbf{X}_n) = \mathbb{E}_{q^*(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\hat{\boldsymbol{x}}_n, \boldsymbol{\xi})]$. Since $\mathcal{X}$ is compact, we assume that $\hat{\boldsymbol{x}}_n \to \boldsymbol{x}^*$ $P_0 - a.s.$ Due to Lemma 7, $q^*\left(g(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0 | \mathbf{X}_n\right)$ hypo-converges to $\mathbb{I}_{(-\infty,0]}(g(\boldsymbol{x}, \boldsymbol{\xi}_0))$ $P_0 - a.s$ as $n \to \infty$, therefore we have

$$\limsup_{n\to\infty} q^*\left(g(\hat{\boldsymbol{x}}_n, \boldsymbol{\xi}) \leq 0 | \mathbf{X}_n\right) \leq \mathbb{I}_{(-\infty,0]}(g(\boldsymbol{x}^*, \boldsymbol{\xi}_0)). \tag{11}$$

Now using the fact that $q^*\left(g(\hat{\boldsymbol{x}}_n, \boldsymbol{\xi}) \leq 0 | \mathbf{X}_n\right) \geq \beta$ for every $n \geq 1$, it follows from (11) that $\boldsymbol{x}^*$ is a feasible point of (TP), since $\limsup_{n\to\infty} q^*\left(g(\hat{\boldsymbol{x}}_n, \boldsymbol{\xi}) \leq 0 | \mathbf{X}_n\right) \geq \beta$ implies $\mathbb{I}_{(-\infty,0]}(g(\boldsymbol{x}^*, \boldsymbol{\xi}_0)) \geq \beta$ and $\beta \in (0, 1)$. Therefore, it follows that $f(\boldsymbol{x}^*, \boldsymbol{\xi}_0) \geq V^*$. Since, due to Lemma 6 (2), $\liminf_{n\to\infty} \mathbb{E}_{q^*(\boldsymbol{\xi}|\mathbf{X}_n)}[f(\hat{\boldsymbol{x}}_n, \boldsymbol{\xi})] \geq f(\boldsymbol{x}^*, \boldsymbol{\xi}_0)$ $P_0 - a.s$, it follows that

$$\liminf_{n\to\infty} V^*_{VB}(\mathbf{X}_n) \geq V^* \ P_0 - a.s. \tag{12}$$

Hence, it follows from (10) and (12) that $V^*_{VB}(\mathbf{X}_n) \to V^*$ $P_0 - a.s$ and it also follows that x^* is the true solution of (TP), therefore $\mathbb{D}(\mathcal{S}^*_{VB}(\mathbf{X}_n), \mathcal{S}^*) \to 0$ $P_0 - a.s.$ The above arguments can be easily generalized for the general case with m number of constraints. ■

C.2. Proof of Proposition 2

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

$$\begin{aligned}\mathbb{P}_0[q^*\left(g_i(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0, \ i \in \{1, 2, 3, \ldots, m\} | \mathbf{X}_n\right) \geq \beta] &\leq \frac{1}{\beta}\mathbb{E}_0[q^*\left(\cap_{i=1}^m \{g_i(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} | \mathbf{X}_n\right)] \\ &\leq \frac{1}{\beta}\mathbb{E}_0[q^*\left(\{g_i(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} | \mathbf{X}_n\right)] \ \forall i \in \{1, \ldots, m\}.\end{aligned} \tag{13}$$

Fix $i \in \{1, 2 \ldots, m\}$. Since $\boldsymbol{x} \in \mathcal{X} \backslash F_0^i$ implies that $\boldsymbol{x} \in \{g_i(\boldsymbol{x}, \boldsymbol{\xi}_0) > 0\}$, it follows that

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

Therefore, for all $\boldsymbol{x} \in \mathcal{X} \backslash F_0^i$ and any $i \in \{1, \ldots, m\}$, it follows from (13) that

$$\begin{aligned} \mathbb{P}_0[q^* \left(g_i(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0,\ i \in \{1, 2, 3, \ldots, m\} | \mathbf{X}_n\right) \geq \beta] &\leq \frac{1}{\beta} \mathbb{E}_0[q^* \left(\cap_{i=1}^m \{g_i(\boldsymbol{x}, \boldsymbol{\xi}) \leq 0\} | \mathbf{X}_n\right)] \\ &\leq \frac{1}{\beta} \mathbb{E}_0[q^* \left(\{g_i(\boldsymbol{x}, \boldsymbol{\xi}) < g_i(\boldsymbol{x}, \boldsymbol{\xi}_0)\} | \mathbf{X}_n\right)]. \end{aligned} \tag{14}$$

Now using Theorem 2.1 in Zhang and Gao (2019), it follows that for each $i \in \{1, 2, 3, \ldots, m\}$ if

$$L_n^i(\theta, \theta_0) := n \sup_{\boldsymbol{x} \in \mathcal{X}} \mathbb{I}_{(0, \infty)}(g_i(\boldsymbol{x}, \boldsymbol{\xi}_0) - g_i(\boldsymbol{x}, \boldsymbol{\xi}))$$

satisfies assumption (C1), then there exists a constant C_i for each $i \in \{1, 2, 3, \ldots, m\}$ such that

$$\mathbb{E}_0[q^* \left(\{g_i(\boldsymbol{x}, \boldsymbol{\xi}) < g_i(\boldsymbol{x}, \boldsymbol{\xi}_0)\} | \mathbf{X}_n\right)] \leq C_i(\epsilon_n^2 + \eta_n^2),$$

where $\eta_n^2 := \frac{1}{n} \inf_{q \in \mathcal{Q}} \mathbb{E}_{P_0} \left[\int_\Theta q(\boldsymbol{\xi}) \log \frac{q(\boldsymbol{\xi})}{\pi(\boldsymbol{\xi} | \mathbf{X}_n)} d\boldsymbol{\xi} d\boldsymbol{\xi}\right]$. Now observe that, using the above result in (14) directly proves the assertion of the proposition. ■