VARIATIONAL INFERENCE FOR DATA-DRIVEN STOCHASTIC PROGRAMMING

by

Prateek Jaiswal

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Industrial Engineering

West Lafayette, Indiana

August 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Harsha Honnappa, Chair

School of Industrial Engineering

Dr. Vinayak A. Rao

Department of Statistics

Dr. Raghu Pasupathy

Department of Statistics

Dr. Gesualdo Scutari

School of Industrial Engineering

Dr. J. George Shanthikumar

Krannert School of Management

Approved by:

Dr. Abhijit Deshmukh

To my parents, wife

and

to the memory of my beloved mother-in-law

ACKNOWLEDGMENTS

This dissertation would not have been possible without constant encouragement and guidance from my advisor Prof. Harsha Honnappa. I am immensely grateful to him for being an excellent mentor, advisor, and teacher. This thesis is an outcome of his trust in assigning me challenging problems and thus giving me the confidence to push my limits. In addition to the work in this dissertation, he also encouraged me to work on different projects that enhanced my learning experience significantly. I am also thankful to him for introducing me to my co-advisor Prof. Vinayak A. Rao. Working with Prof. Rao was one of the most enjoyable experiences of my graduate studies. He shaped my thesis constantly by asking relevant and fundamental questions underlying a given problem. I am also grateful to him for patiently answering my questions and carefully evaluating my ideas. Prof. Honnappa also introduced me to Prof. Raghu Pasupathy. We together worked on various interesting problems that helped me in broadening my research interest significantly. I am thankful to Prof. Pasupathy for his constant encouragement during graduate studies and for being an excellent teacher. I would also like to express my sincere gratitude to Prof. J. George Shanthikumar and Prof. Gesualdo Scutari for serving on my doctoral advisory committee and providing invaluable comments and suggestions.

I also want to thank my colleagues Imon, Prakash, Ruixin, Maithilee, Viplove, Ye, Monica, Cansu, Rahul, Brayan, Eric, Arnob, Andrew, Sagar, Aniket and many others for numerous reinvigorating over-the-lunch and corridor talks, for attending lectures and seminars together, and for many useful discussions and tips. I would also like to thank Anita Park, Cheryl Barnhart, and members of the administrative team in the School of Industrial Engineering for their many non-academic but necessary support.

I am fortunate to have another family away from home in West Lafayette: Gaurav, Romila, Debapriya, Surya, Manali, Amit, Salil, Richa, Ashish, Shraddha, Bhavana, and Vishal. I am thankful to them for their support and for always being there whenever needed the most. It was indeed an enjoyable five years spent with you all!

I am also incredibly grateful to my parents, parents-in-law, brother, brother-in-law for believing in me and always supporting my decisions. In particular, I will be forever indebted

to my parents for teaching me the importance of education and going beyond their reach to provide me the best education. Last but most important of all, I would like to thank my wife Aparajita for being selfless in changing her career path and coming along thousands of miles away to fulfill my dreams and for her unconditional support in all my endeavors. I will always be indebted to her for being with me on this roller-coaster ride at every moment and being so considerate when I was unable to give her enough time. I would also like thank my month-old adorable daughter Suchi for instilling joy and happiness in my family when needed the most.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Stochastic programs are standard models for decision-making under uncertainty and have been extensively studied in the operations research literature. In general, stochastic programming involves minimizing an expected cost function, where the expectation is with respect to fully specified stochastic models that quantify the aleatoric or 'inherent' uncertainty in the decision-making problem. In practice, however, the stochastic models are unknown but can be estimated from data, introducing an additional epistemic uncertainty into the decision-making problem. The Bayesian framework provides a coherent way to quantify the epistemic uncertainty through the posterior distribution by combining prior beliefs of the decision-makers with the observed data. Bayesian methods have been used for data-driven decision-making in various applications such as inventory management, portfolio design, machine learning, optimal scheduling, and staffing, etc.

Bayesian methods are challenging to implement, mainly due to the fact that the posterior is computationally intractable, necessitating the computation of approximate posteriors. Broadly speaking, there are two methods in the literature implementing approximate posterior inference. First are sampling-based methods such as Markov Chain Monte Carlo. Sampling-based methods are theoretically well understood, but they suffer from various issues like high variance, poor scalability to high-dimensional problems, and have complex diagnostics. Consequently, we propose to use optimization-based methods collectively known as variational inference (VI) that use information projections to compute an approximation to the posterior. Empirical studies have shown that VI methods are computationally faster and easily scalable to higher-dimensional problems and large datasets. However, the theoretical guarantees of these methods are not well understood. Moreover, VI methods are empirically and theoretically less explored in the decision-theoretic setting.

In this thesis, we first propose a novel VI framework for risk-sensitive data-driven decision-making, which we call risk-sensitive variational Bayes (RSVB). In RSVB, we jointly compute a risk-sensitive approximation to the 'true' posterior and the optimal decision by solving a minimax optimization problem. The RSVB framework includes the naive approach of first computing a VI approximation to the true posterior and then using it in place of the true

posterior for decision-making. We show that the RSVB approximate posterior and the corresponding optimal value and decision rules are asymptotically consistent, and we also compute their rate of convergence. We illustrate our theoretical findings in both parametric as well as nonparametric setting with the help of three examples: the single and multi-product newsvendor model and Gaussian process classification. Second, we present the Bayesian joint chance-constrained stochastic program (BJCCP) for modeling decision-making problems with epistemically uncertain constraints. We discover that using VI methods for posterior approximation can ensure the convexity of the feasible set in (BJCCP) unlike any sampling-based methods and thus propose a VI approximation for (BJCCP). We also show that the optimal value computed using the VI approximation of (BJCCP) are statistically consistent. Moreover, we derive the rate of convergence of the optimal value and compute the rate at which a VI approximate solution of (BJCCP) is feasible under the true constraints. We demonstrate the utility of our approach on an optimal staffing problem for an M/M/c queue. Finally, this thesis also contributes to the growing literature in understanding statistical performance of VI methods. In particular, we establish the frequentist consistency of an approximate posterior computed using a well known VI method that computes an approximation to the posterior distribution by minimizing the Rényi divergence from the 'true' posterior.

1. INTRODUCTION

Consider the following parameterized optimization problem:

$$\begin{aligned} \text{minimize} \quad & R(a,\theta) \\ \text{s.t.} \quad & g_{\mathrm{i}}(a,\theta) \leq 0,\ \mathrm{i} \in \{1,2,3,\ldots,m\}, \end{aligned} \tag{TP}$$

where $a \in \mathcal{A}$ is the decision variable and $\theta \in \Theta$ is the model parameter. The function $R(a,\theta) : \mathcal{A} \times \Theta \mapsto \mathbb{R}$ encodes the cost/risk and the functions $g_{\mathrm{i}}(a,\theta) : \mathcal{A} \times \Theta \mapsto \mathbb{R}$ define the constraints. Under certain regularity conditions on the cost and the constraint functions, and for a given value of parameter θ, we assume that (TP) can be solved to compute a set of optimal decisions a^*. These types of problems are studied in the OR/MS community as stochastic programs [1], [2]. In particular, (TP) can be represented as the following general stochastic program [3],

$$\begin{aligned} \text{minimize} \quad & \mathcal{R}^0_{P_\theta}[\ell(a,\xi)] \\ \text{s.t.} \quad & \mathcal{R}^{\mathrm{i}}_{P_\theta}[\bar{g}_{\mathrm{i}}(a,\xi)] \leq 0,\ \mathrm{i} \in \{1,2,3,\ldots,m\}, \end{aligned} \tag{SPP}$$

where $\mathcal{R}^{\mathrm{i}}_{P_\theta}(\cdot)$, $\mathrm{i} \in \{0,1,2,\ldots,m\}$ are well-defined risk measures for a given distribution P_θ that measures the aleatoric uncertainty in the random variable ξ. Compare (TP) and (SPP) to note that $R(a,\theta) = \mathcal{R}^0_{P_\theta}[\ell(a,\xi)]$ and $g_{\mathrm{i}}(a,\theta) = \mathcal{R}^{\mathrm{i}}_{P_\theta}[\bar{g}_{\mathrm{i}}(a,\xi)]$ for appropriate measurable functions $\ell(\cdot,\cdot)$ and $\bar{g}_{\mathrm{i}}(\cdot,\cdot)$. Some of the popular risk measures are *expectation*, *Value-at-risk* (VaR), *Conditional Value-at-risk* (CVaR), and *entropic risk-measure* [3].

To illustrate the (TP) with an example, consider the problem of optimally staffing a stochastic system modeled as an M/M/c queue. A queueing system with exponentially distributed inter-arrival times (T) and service times (S) with c number of servers is denoted as an M/M/c queue. We assume that the arrival and service rates are λ and μ respectively, and they together constitute the model parameter $\theta = \{\lambda, \mu\}$. The DM chooses the number

of servers c to ensure that the steady-state probability that the customer waits in the queue, denoted as $1 - W_q(c, \lambda, \mu)$, is no more than α. The corresponding staffing problem is to

$$\begin{aligned} \text{minimize} \quad & c \\ \text{subject to} \quad & (\{1 - W_q(c, \lambda, \mu)\} - \alpha) < 0 \quad \text{(Quality of Service)}, \\ & (\lambda - c\mu) < 0 \quad \text{(Stability)}, \end{aligned} \tag{TP-Q}$$

where $\alpha \in (0, 1/2)$ is the desired efficiency. The second constraint is to ensure that the queue is stable. This staffing problem and its variations are well studied in the queueing literature [4]–[6].

In many applications, the constraint set defined by functions g_{i} in (TP) does not depend on the model parameter θ. As an example, consider the newsvendor problem, a canonical data-driven decision-making problem, which has been extensively studied in the stochastic programming literature [1], [7]–[9]. Recall that the newsvendor loss function is defined as $\ell(a, \xi) := h \max(a - \xi, 0) + b \max(\xi - a, 0)$, where h (underage cost) and b (overage cost) are given positive constants, $\xi \in [0, \infty)$ the random demand, and a the inventory or decision variable. The optimal decision problem in this case is to

$$\text{minimize}_{a \in \mathcal{A}} R(a, \theta) = \mathbb{E}_{P_\theta}[\ell(a, \xi)], \tag{1.1}$$

where P_θ is the distribution over future random demands ξ. Stochastic programs with deterministic constraints are also common in machine learning applications such as Bayesian regression and classification, and their variations incorporating regularization [10], [11]. For instance, consider the problem of classifying an input pattern or features $Y \in [0, 1]^d$ into one of the two classes $\{-1, 1\}$, where $\xi \in \{-1, 1\}$ denote the class of Y. For a given Y, the classifier is modelled using a Bernoulli distribution $p(\xi|Y, \theta) = \Psi_\xi(\theta(Y))$, where $\theta : [0, 1]^d \to \mathbb{R}$ is a non-parametric model parameter in a separable Banach space Θ and measurable functions $\Psi_1(x) = (1 + \mathrm{e}^{-x})^{-1}$ and $\Psi_{-1}(x) = 1 - \Psi_1(x)$. Assuming Y is independent of ξ and has distribution $\nu(\cdot)$, the sequence of independent observations $\{\tilde{Y}_n, \tilde{X}_n\} = \{(Y_1, \xi_1), (Y_2, \xi_2), \ldots, (Y_n, \xi_n)\}$ are generated from model $P_\theta(\xi, Y) = p(\xi|Y, \theta)\nu(Y)$. The loss function $\ell(a, \xi)$ is

0 if $a = \xi$, c_+ if $\{a = +1, \xi = -1\}$, and c_- if $\{a = -1, \xi = +1\}$, where c_+ and c_- are known positive constants. Now the objective is to minimize the following model risk

$$R(a,\theta) = \mathbb{E}_{P_\theta}[\ell(a,\xi)] = \begin{cases} c_+ \mathbb{E}_\nu[\Psi_{-1}(\theta(y))], & a = +1, \\ c_- \mathbb{E}_\nu[\Psi_1(\theta(y))], & a = -1. \end{cases} \tag{1.2}$$

Notice that if the parameter θ is known, then the optimal staffing, newsvendor, and binary classification problems are deterministic optimization problems, assuming the risk functions can be computed. We denote true model parameters by θ_0. In practice, however, the true model parameters are unknown which further introduces epistemic uncertainty into the formulation of the stochastic programs. Epistemic uncertainties can be reduced by estimating the true model parameters from data. Plug-in estimators such as maximum likelihood estimates typically produce sub-optimal or counterintuitive solutions in decision-making problems [12, Section 2]. This has spurred a significant body of research on decision-making and optimization methodology that accounts for this epistemic uncertainty; for instance, distributionally robust optimization (DRO) methods [13], [14] optimize worst case deviations from the empirical objective, while Bayesian decision-theoretic methods [12], [15] penalize the predictive uncertainty from the Bayes posterior distribution over the parameters. The authors in [16], [17] discuss other methods of modeling epistemic uncertainty. In this thesis, we are interested in the Bayesian framework, which provides a coherent way to quantify the epistemic uncertainty through the posterior distribution by combining prior belief of the decision-maker (DM) with data. Bayesian methods have been used for data-driven decision-making in various applications such as inventory management [18], [19], portfolio design [12], machine learning [10], [11], engineering [20], simulation optimization [21]–[23], etc.

In general, the Bayesian inferential problem is hard owing to the intractable posterior computation. Consequently, DM's typically make restrictive modeling choices, such as assuming that the likelihood model has a conjugate prior. However, conjugate priors are not available for many interesting and practical likelihood models, thereby limiting the use and utility of the Bayesian framework. To mitigate this limitation of computational intractability, there is a substantial body of work on *approximate Bayesian computation* focused on

the question of efficiently and accurately approximating the posterior distribution. Broadly, these posterior approximation techniques are categorized into sampling and optimization-based approaches. Markov Chain Monte Carlo (MCMC) is the canonical sampling method, where the objective is to design an ergodic Markov chain whose invariant or stationary distribution is precisely the posterior distribution. MCMC, however, is known to suffer from high variance, complex diagnostics, and has poor scaling properties with the problem dimension and size of the dataset [24].

Variational Inference (VI) or Variational Bayesian (VB) methods, in contrast, use optimization to compute an approximation to the posterior distribution from a class of 'simpler' distribution functions (that does not, necessarily, contain the posterior) called the *variational family*, by minimizing the divergence of distributions in the variational family from the posterior distribution. Importantly, the posterior distribution being intractable, VI methods optimize a surrogate objective that upper bounds the divergence measure, and the optimizer of the surrogate is precisely the posterior distribution when the variational family includes it. The Kullback-Leibler (KL) divergence is a standard choice in VI methods [25], though there is increasing interest in other choices such as α-Rényi divergence as well [26] which yield approximations that have better support coverage. Moreover, VI methods have been empirically demonstrated to be a faster and easier-to-scale alternative to sampling-based methods for Bayesian inference in various high-dimensional and complex hierarchical probabilistic models with large datasets [25] . Despite their popularity in machine learning and statistics community, the statistical performance of these methods were only studied recently [27]–[30]. Nonetheless, VI methods are empirically and theoretically less explored in the decision-theoretic setting.

Through this thesis, we develop VI methods to compute tractable approximations to stochastic programs with epistemic uncertainty. More crucially, we analyze the statistical inferential properties of these approximations and establish theoretical guarantees on the predictive performance of inferred decision rules and values. We would also like to note that the theoretical results established in this thesis also extends to the decision rules and values computed using the true posterior, since VI posterior is identical to true posterior when the variational family includes it. Furthermore, this thesis also contributes to the growing

literature in understanding statistical performance of VI methods. In particular, we establish the frequentist consistency of an approximate posterior computed using a well known VI method that computes an approximation to the posterior distribution by minimizing the Rényi divergence from the 'true' posterior.

In the next two sections, we first consider the class of stochastic programs where the constraint set is known *a priori* or deterministic like newsvendor problem and then we study stochastic programs with epistemically uncertain constraint set like optimal staffing problem.

1.1 Stochastic Programs with Deterministic Constraints

The stochastic programs with deterministic constraint set have been an active topic of research among statisticians, economists, and engineers since early 20th century. Among various theories to model epistemic uncertainty in such stochastic programs, a seminal unifying framework was proposed by Abraham Wald [31], [32]. In Wald's general decision theory, the unknown model parameter θ is treated as a random variable, defined on a probability space $(\Theta, \mathcal{T}, \Pi)$, where the set of possible values of a parameter is denoted by Θ with $\sigma-$algebra $\mathcal{T}$ and probability distribution $\Pi(\cdot)$. In essence, the distribution $\Pi(\cdot)$ measures the epistemic uncertainty in θ. The decision space (known *a priori*) is denoted as $\mathcal{A}$. Then for a fixed $\Pi(\cdot)$ and a given cost/risk function $R(\cdot, \cdot)$ the Wald's decision making problem is defined as

$$\text{minimize}_{a \in \mathcal{A}} \ \mathbb{E}_{\Pi}[R(a, \theta)]. \tag{1.3}$$

Recall that the model parameter is unknown but the DM has access to data which can be used to quantify the epistemic uncertainty in θ. The DM can use data to *forecast* a distribution over it. A natural question that arises in context of this formulation is, *what is an appropriate distribution* Π *that facilitates data-driven decision making?* Bayesian statistics delineates natural principles to forecast distributions over parameters in a coherent way by combining prior knowledge with the observed data. Now let us suppose that the DM observes n samples from the stochastic model $P_{\theta_0}^{\infty}$ denoted as $\tilde{X}_n := \{\xi_1, \xi_2, \ldots, \xi_n\}$, with unknown

θ_0. In the Bayesian framework, we posit a prior distribution $\Pi(\theta)$ and use the likelihood of observing $\tilde{X}_n$, that is $p_\theta^n(\tilde{X}_n)$, to define the posterior distribution as

$$d\Pi_n \equiv d\Pi(\theta|\tilde{X}_n) := \frac{d\Pi(\theta)p_\theta^n(\tilde{X}_n)}{\int_\Theta d\Pi(\theta)p_\theta^n(\tilde{X}_n)}. \tag{1.4}$$

The objective in 1.3 is popularly known as *Bayes risk* when Π is the posterior distribution in 1.3.

A second crucial question in relation to (1.3) is: *Is expectation with respect to the posterior distribution Π_n the most appropriate way of capturing uncertainty over the model parameters, θ?* To answer this, recall that the expectation *is* a risk measure and it averages the risk of uncertainty in the future costs in (1.3). In various applications, like portfolio design, the decision-maker (DM) could be risk-averse and would like to use a risk measure that reflects this risk attitude, unlike the risk-neutral expectation. Recently, there has also been significant interest in the machine learning and operations research community in studying models that are sensitive to tail and subgroups effects [12], [15], [33]–[35]. Therefore, to facilitate the DM in making risk sensitive decisions, we investigate the use of a risk sensitive measure known as the *entropic risk measure*, defined as

$$\varrho_{\Pi_n}^\gamma(R(a,\theta)) := \frac{1}{\gamma}\log \mathbb{E}_{\Pi_n}[\exp(\gamma R(a,\theta))], \tag{1.5}$$

where $\gamma \in \mathbb{R}$. The entropic risk models a range of risk-averse or risk-seeking behaviors in a succinct manner through the parameter γ. Suppose $\gamma \to 0$, then observe that

$$\lim_{\gamma \downarrow 0} \varrho_{\Pi_n}^\gamma(R(a,\theta)) = \mathbb{E}_{\Pi_n}(R(a,\theta));$$

that is, there is no sensitivity to potential risks due large tail effects and the DM is risk neutral. On the other hand,

$$\lim_{\gamma \to \infty} \varrho_{\Pi_n}^\gamma(R(a,\theta)) = \operatorname{ess\,sup}_{\Pi_n}(R(a,\theta)),$$

where ess sup is the essential supremum of the random function $R(a,\theta)$ (almost surely). In other words, a decision maker is completely risk averse and anticipates the worst possible realization (almost surely). Similar conclusions can be drawn when $\gamma < 0$, resulting in a risk-seeking behavior.

Observe that (1.5) strictly generalizes the standard Bayesian decision-theoretic formulation of a decision-making problem, where the goal is to solve $\min_{a\in\mathcal{A}} \mathbb{E}_{\Pi_n}[R(a,\theta)]$. Furthermore, it also coincides with other risk-based Bayesian methods, such as the penalized posterior variance method studied in [12] for solving the Markowitz portfolio optimization problem, under certain parameterizations. More precisely, for $R(a,\theta) = \varrho^{\gamma}_{P_\theta}(\ell(a,\xi))$ (for any loss function $\ell(\cdot,\cdot)$) and small, but strictly positive γ, a Taylor expansion of $\varrho^{\gamma}_{\Pi_n}(R(a,\theta))$ straightforwardly shows that (1.5) is equivalent to problem (3.1) in [12], that is $\varrho^{\gamma}_{\Pi_n}(R(a,\theta)) = \varrho^{\gamma}_{\pi(\xi|\tilde{X}_n)}(\ell(a,\xi)) \overset{\gamma\simeq 0}{\simeq} \gamma \mathrm{Var}_{\pi(\xi|\tilde{X}_n)}[\ell(a,\xi)] + \mathbb{E}_{\pi(\xi|\tilde{X}_n)}[\ell(a,\xi)]$, where $\boldsymbol{\pi}(\xi|\tilde{X}_n) := \int p(\xi|\theta) d\Pi(\theta|\tilde{X}_n)$ is the posterior predictive distribution and $\mathrm{Var}[\,\cdot\,]$ denotes the variance functional.

Our setting is most related to recent work on *Bayesian risk optimization* (BRO) in [15], [36]. In BRO, the authors consider optimal decision-making using various risk measures (other than the entropic risk measure) computed under the posterior distribution. The authors establish several important results, including that the optimal value and decisions are asymptotically consistent as the sample size tends to infinity, and central limit type theorem for the optimal values. Moreover, they also assume that $\tilde{X}_n$ are independent and identically distributed (i.i.d.) samples from $P^{\infty}_{\theta_0}, \theta_0 \in \Theta \subset \mathbb{R}^d$.

In practice, Bayesian inference is challenging owing to the fact that computing the posterior distribution is intractable. More precisely, computing the integral in the denominator in (1.4) poses severe computational challenges, thereby rendering the inference problem in (1.3) (with Π as Π_n) and (1.5) intractable in general. The works in [12], [15], [36] presumes that the posterior distribution is actually computable. The authors do not address the critical computational questions surrounding Bayesian methods or the impact of (inevitable) computational approximations. As motivated before, in this thesis we address the computational intractability of the posterior distribution by using an optimization-based posterior approximation technique VI instead of sampling-based methods.

In particular, in Chapter 2 we introduce a novel computationally tractable framework which we call *risk-sensitive variational Bayes* (RSVB) to approximate (1.5) when the distribution over parameter θ is the posterior distribution. The proposed general framework can be used to extract computational methods for doing risk-sensitive approximate Bayesian inference. We show that our general framework includes two well known computational algorithms for doing approximate Bayesian inference viz. *naive* VI (NVB) and *loss-calibrated* VI (LCVB) [10]. We also study the impact of RSVB computational approximations on the predictive performance of the inferred decision rules and values. We show that the RSVB approximate posterior and the corresponding optimal value and decision rules are asymptotically consistent, and we also compute their rate of convergence. We establish these result under regularity conditions that do not require $\tilde{X}_n$ to be i.i.d. samples from $P_{\theta_0}^{\infty}, \theta_0 \in \Theta$, and where Θ can be any arbitrary model space with norm. We illustrate our theoretical findings in both parametric as well as nonparametric setting with the help of three examples: the single and multi-product newsvendor model and Gaussian process classification. Furthermore, in Chapter 3 we establish asymptotic guarantees on the decision rules computed using LCVB approximation method under relatively milder set of assumptions when $\Theta \subset \mathbb{R}^d$.

1.2 Stochastic Programs with Epistemically Uncertain Constraints

Recall (TP), stated here again,

$$\begin{aligned} \text{minimize} \quad & R(a, \theta) \\ \text{s.t.} \quad & g_{\mathrm{i}}(a, \theta) \leq 0, \ \mathrm{i} \in \{1, 2, 3, \ldots, m\}. \end{aligned} \tag{TP}$$

Since the model parameters are unknown, we again adopt a Bayesian approach and model the epistemic uncertainty over the parameters θ by computing a posterior distribution $\Pi(\theta|\tilde{X}_n)$.

We approximate the true problem (TP), using the posterior distribution, with the following joint chance-constrained problem:

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{\Pi(\theta|\tilde{X}_n)}[R(a,\theta)] \\ \text{s.t.} \quad & \Pi\left(g_{\mathrm{i}}(a,\theta) \leq 0,\ \mathrm{i} \in \{1,2,3,\ldots,m\}|\tilde{X}_n\right) \geq \beta, \forall a \in \mathcal{A}, \end{aligned} \tag{BJCCP}$$

where $\beta \in (0,1)$ is the specified confidence level desired by the decision DM based on the requirement, usually $\beta > \frac{1}{2}$. We provide a supporting example in Chapter 4 to motivate the chance-constrained formulation as opposed to using expectations, in which case the constraints are only satisfied on an average. To the best of our knowledge, Bayesian models of data-driven chance constrained optimization have not been considered before in the literature. On the other hand, we note that there is precedence for Bayesian formulations of data-driven stochastic optimization problems studied in [15], [37]–[39].

We would also like note that (BJCCP) is fundamentally different from the usual stochastic programs with chance constraints [2]. In particular, the objective there is to solve the following problem using samples from the unknown data generating distribution P_{θ_0}

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{P_{\theta_0}}[\ell(a,\xi)] \\ \text{s.t.} \quad & P_{\theta_0}\left(\bar{g}_{\mathrm{i}}(a,\xi) \leq 0,\ \mathrm{i} \in \{1,2,3,\ldots,m\}|\tilde{X}_n\right) \geq \alpha, \forall a \in \mathcal{A}, \end{aligned} \tag{JCCP}$$

for some fixed mappings $\bar{g}_{\mathrm{i}}(\cdot,\cdot)$ and $\alpha \in (0,1)$. There is an extensive literature on data-driven methods for solving (JCCP), specifically scenario-based (SB) approaches [40]–[42], distributionally robust optimization (DRO) [40], [43]–[45] and sample average approximation (SAA) [46], [47]. We direct the reader to the excellent recent review paper [48] for a comprehensive overview of the literature on data-driven chance constrained optimization to solve (JCCP). In particular, we observe that the ambiguity set in DRO quantifies the epistemic uncertainty when 'centered' (defined, for instance, through the Wasserstein metric) around the empirical measure, which converges to the data-generating measure in the large sample limit; see [49] which establishes the consistency of chance-constrained DRO with Wasserstein ambiguity sets. This highlights an important difference with our current

setting, where the posterior distribution (or its approximation) is used as a quantification of the epistemic uncertainty about the 'true' parameter θ_0, and is shown to weakly converge to a Dirac delta distribution concentrated at θ_0 as the number of samples n tends to infinity.

Recall that computing posterior distributions is challenging and mostly intractable, and is typically approximated using MCMC or VI methods. As noted before, MCMC methods have their own drawbacks like poor mixing, large variance, and complex diagnostics, which have been the usual motivation for using VI [50]. Here, we provide another important motivation for using VI in the chance-constrained Bayesian inference setting. In particular, we present an example (motivated from [51]) where a sampling based approach to approximate the chance-constrained convex feasibility set (constraint set) in (BJCCP), results in a non-convex approximation; whereas an appropriate VI approximation *retains its convexity* (for an appropriate choice of variational family). Therefore, we approximate (BJCCP) using a VI approximate posterior $Q^*(\theta|\tilde{X}_n)$ to $\Pi(\theta|\tilde{X}_n)$ as:

$$\begin{aligned} \text{minimize} \quad & \mathbb{E}_{Q^*(\theta|\tilde{X}_n)}[R(a,\theta)] \\ \text{s.t.} \quad & Q^*\left(g_{\mathrm{i}}(a,\theta) \leq 0,\ \mathrm{i} \in \{1,2,3,\ldots,m\}|\tilde{X}_n\right) \geq \beta, \forall a \in \mathcal{A}. \end{aligned} \tag{VBJCCP}$$

However, since VI posterior can be a biased approximation to the true posterior and thus may result into an approximate feasible set which could include infeasible points. Therefore, it is important to study the consistency properties of the feasible set, the optimal values and solutions of (VBJCCP) with respect to the sample size n. Consequently, in Chapter 4 we study (VBJCCP) approximation of the true stochastic programming problem when the model parameters are unknown. We show that the optimal value of (VBJCCP) are consistent with the optimal value of (TP). More precisely, we show that the optimal value computed in (VBJCCP) converges to the true optimal value as the number of samples tends to infinity. We augment this by also establishing a probabilistic rate of convergence of the optimal value. We also provide bounds on the probability of qualifying a true infeasible point (with respect to the true constraints) as feasible under the VI approximation for a given number of samples. Finally, we demonstrate the utility of our approach on an optimal staffing problem for an M/M/c queueing model.

1.3 Variational Bayesian Inference - Beyond KL divergence

Recall that the high level idea behind VI is to approximate the intractable posterior $\Pi(\theta|\tilde{X}_n)$ with an element $Q(\theta)$ of some simpler class of distributions $\mathcal{Q}$ known as variational family. The variational solution $Q^*(\theta|\tilde{X}_n)$ is the element of $\mathcal{Q}$ that is closest to $\Pi(\theta|\tilde{X}_n)$, where closeness is measured in terms of the KL divergence. Thus, $Q^*(\theta|\tilde{X}_n)$ is the solution to:

$$Q^*(\theta|\tilde{X}_n) = \text{argmin}_{\tilde{Q}\in\mathcal{Q}}\text{KL}(\tilde{Q}(\theta)\|\Pi(\theta|\tilde{X}_n)). \tag{1.6}$$

Despite its popularity, classical VI (KLVI) has a number of well-documented limitations. An important one is its tendency to produce approximations that under estimate the spread of the posterior distribution [52]–[55]: in essence, the KLVI solution tends to match closely with the dominant mode of the posterior. This arises from the choice of the divergence measure $\text{KL}(Q(\theta)\|\Pi(\theta|\tilde{X}_n)) := \mathbb{E}_Q[\log(dQ(\theta)/d\Pi(\theta|\tilde{X}_n))]$, which does not penalize solutions where $dQ(\theta)$ is small while $d\Pi(\theta|\tilde{X}_n)$ is large. While many statistical applications only focus on the mode of the distribution, definite calculations of the variance and higher moments are critical in predictive and decision-making problems.

A natural solution is to consider different divergence measures than those used in variational Bayes. Expectation propagation (EP) [54] was developed to minimize $\mathbb{E}_{\Pi_n}[\log(\Pi_n/Q)]$ instead, though this requires an expectation with respect to the intractable posterior. Consequently, EP can only minimize an approximation of this objective. Moreover, there are some extensions of EP with alternate divergence measures [56], [57].The authors in [58] replaces KL divergence in EP to χ^2-divergence to compute variational approximations that significantly improve upon the KLVI and EP in accurately approximating the posterior variance.

More recently, Rényi's α-divergence [59] has been used as a family of parameterized divergence measures for variational inference [52], [60]. The α-Rényi ($\alpha > 1$) approximate posterior $Q_r^*(\theta|\tilde{X}_n)$ is defined as

$$Q_r^*(\theta|\tilde{X}_n) := \text{argmin}_{\tilde{Q}\in\mathcal{Q}} D_\alpha\left(\Pi(\theta|\mathbf{X}_n)\|\tilde{Q}(\theta)\right)$$

where $D_\alpha\left(P(y)\|Q(y)\right) = \frac{1}{\alpha-1}\log\int dP(y)\left(\frac{dP(y)}{dQ(y)}\right)^\alpha$. Unlike KLVI approximate posterior, the $\alpha-$Rényi approximate posterior does not underestimate the posterior variance, resulting in predictions that captures the high-risk regions in the support of the posterior [52]. Since, a DM is ultimately interested in using $\alpha-$Rényi approximate posterior for approximate Bayesian inference, establishing its large sample properties will help in analyzing the predictive performance of the inferred decision rules. In fact, our statistical consistency results in Chapter 4, for the KLVI approximate posterior can be easily extended to the $\alpha-$Rényi approximate posterior. Moreover, α-Rényi divergence minimization has empirically demonstrated very promising results for a number of machine learning applications [52], [60].

In recent work, Zhang and Gao [61] have shown conditions under which α-Rényi variational methods are consistent when α is less than one. The setting with α greater than 1 is qualitatively different from both KL and Rényi divergence with $\alpha < 1$ and the results in Zhang and Gao [61] does not extend to this setting. Consequently, in Chapter 5, we address the question of asymptotic consistency of the approximate posterior distribution obtained by minimizing the $\alpha-$Rényi divergence for $\alpha > 1$. Our primary result identifies sufficient conditions under which consistency holds, centering around the existence of a 'good' sequence of distributions in the approximating family. Furthermore, since $D_\alpha\left(\Pi(\theta|\mathbf{X}_n)\|\tilde{Q}(\theta)\right) \to \mathrm{KL}\left(\Pi(\theta|\mathbf{X}_n)\|\tilde{Q}(\theta)\right)$, as $\alpha \to 1$, we recover the asymptotic consistency of the EP approximate posterior from our results on the consistency of α-Rényi approximate posterior.

2. RISK-SENSITIVE VARIATIONAL BAYES

This chapter focuses on a *risk-sensitive* Bayesian formulation of the data-driven decision-making problem of the form

$$\min_{a\in\mathcal{A}} \varrho^{\gamma}_{\Pi_n}(R(a,\theta)) := \frac{1}{\gamma}\log \mathbb{E}_{\Pi_n}[\exp(\gamma R(a,\theta))], \tag{SO}$$

where $\mathcal{A} \subset \mathbb{R}^s$ $(s \geq 1)$ is the decision/action space, θ is a random model parameter lying in an arbitrary measurable space $(\Theta, \mathcal{T})$ distributed according to Π_n the Bayesian posterior distribution over the parameters $\Pi_n(\theta) := \Pi(\theta|\tilde{X}_n)$, and $R(a,\theta) : \mathcal{A}\times\Theta \mapsto \mathbb{R}$ is a problem-specific model risk function. The scalar $\gamma \in \mathbb{R}$ is user-specified and characterizes the sensitivity of the decision-maker (DM) to the distribution Π_n. A prior probability distribution $\Pi(\theta)$ capturing the subjective belief of the decision maker is posited over θ, and that belief is updated according to Bayes rule to compute a posterior distribution $\Pi_n(\theta)$ over the parameters using a set of n observations $\tilde{X}_n = \{\xi_1, \ldots, \xi_n\}$ sampled from a data-generating distribution P^n_θ with density p^n_θ. Mathematically, the posterior distribution is defined as

$$d\Pi(\theta|\tilde{X}_n) = \frac{d\Pi(\theta)p^n_\theta(\tilde{X}_n)}{\int_\Theta d\Pi(\theta)p^n_\theta(\tilde{X}_n)}, \tag{2.1}$$

where $p^n_\theta(\tilde{X}_n)$ is the likelihood of observing $\tilde{X}_n$.

The functional $\varrho^{\gamma}_{\cdot}$ is also known as the *entropic risk measure*, and models a range of risk-averse or risk-seeking behavior in a succinct manner through the parameter γ. Consider only strictly positive γ, and observe that

$$\lim_{\gamma\downarrow 0}\frac{1}{\gamma}\log \mathbb{E}_{\Pi_n}[\exp(\gamma R(a,\theta))] = \mathbb{E}_{\Pi_n}(R(a,\theta));$$

that is, there is no sensitivity to potential risks due to large tail effects and the decision-maker is risk neutral. On the other hand,

$$\lim_{\gamma\to+\infty} \varrho^{\gamma}_{\Pi_n}(R(a,\theta)) = \operatorname{ess\,sup}_{\Pi_n}(R(a,\theta)),$$

where ess sup is the essential supremum of the model risk $R(a, \theta)$. In other words, a decision maker is completely risk averse and anticipates the worst possible realization (Π_n-almost surely). While similar conclusions can be drawn when $\gamma < 0$, resulting in risk-seeking behavior, we restrict ourselves to $\gamma > 0$ in this thesis. Observe that (SO) strictly generalizes the standard Bayesian decision-theoretic formulation of a decision-making problem, where the goal is to solve $\min_{a \in \mathcal{A}} \mathbb{E}_{\Pi_n}[R(a, \theta)]$. Furthermore, it also coincides with other risk-based Bayesian methods, such as the penalized posterior variance method studied in [12] for solving the Markowitz portfolio optimization problem, under certain parameterizations. More precisely, for $R(a, \theta) = \varrho^{\gamma}_{P_\theta}(\ell(a, \xi))$ (for any loss function $\ell(\cdot, \cdot)$) and small, but strictly positive γ, a Taylor expansion of $\varrho^{\gamma}_{\Pi_n}(R(a, \theta))$ straightforwardly shows that (SO) is equivalent to problem (3.1) in [12].

The risk-sensitive formulation (SO) is very general and can be used to model a wide variety of decision-making problems in operations research/ management science [19], [62], [63], simulation optimization [15], [64], and finance [12], [65], [66]. Moreover, it presents a natural way to address epistemic model uncertainty by being Bayesian and risk sensitive. Our approach can be an alternative to distributional robust optimization (DRO) framework [13], where the decision maker models the ambiguity in the choice of distributions by being robust against the unknown data generating distribution (or model).

Although versatile, solving (SO) to compute an optimal decision over $\mathcal{A}$ is challenging. The difficulty mainly lies in computing the denominator in (2.1) for any given prior distribution (except conjugate priors) that makes the posterior distribution intractable. The use of conjugate priors is restrictive and moreover, for many important likelihood models, they often do not exist. Canonically, posterior intractability is addressed using either a sampling- or optimization-based approach. Sampling-based approaches, such as Markov chain Monte Carlo (MCMC), offer a tractable way to compute the integrals and theoretical guarantees of exact inference in the large computational budget limit. However, these asymptotic guarantees are offset by issues like poor mixing, large variance and complex diagnostics in practical settings with finite computational budgets.

In response, optimization-based methods such as variational Bayes (VB) or variational inference (VI) have emerged as a popular alternative [67]. The VB approximation of the true

posterior is a tractable distribution, chosen from a 'simpler' family of distributions (known as variational family) by minimizing the discrepancy between the true posterior and members of that family. Kullback-Liebler (KL) divergence is the most often used measure of the approximation discrepancy, although other divergences (such as the α-Rényi divergence [30], [52], [53]) have been used. The minimizing member (termed the VB approximate posterior) can be used as a proxy for the true posterior. Empirical studies have shown that VB methods are computationally faster and far more scalable to higher-dimensional problems and large datasets. Theoretical guarantees, such as large sample statistical inference, have been a topic of recent interest in theoretical statistics community. Asymptotic properties such as convergence rate and asymptotic normality of the VB approximate posterior have been established recently in [28], [68] and [27] respectively.

Our ultimate goal is not to merely approximate the posterior distribution, but to also make decisions when that posterior is intractable. A naive approach would be to plug in the VB approximation in place of the true posterior in (SO) and compute the optimal decision. However, it has been noted in [10] that such a loss unaware (or 'naive') approach can be 'suboptimal'. In particular, [10] demonstrated, through an example, that a naive posterior approximation only captures the most dominant mode of the true posterior which may not be relevant from decision-making perspective. Consequently, they proposed a loss-calibrated variational Bayesian (LCVB) algorithm for solving Bayesian decision making problems where the underlying risk function is discrete. [11] extended their approach to continuous risk functions. Despite these algorithmic advances in developing decision-centric variational Bayesian methods, their statistical properties such as asymptotic consistency and convergence rates of the loss-aware posterior approximation and the associated decision rule are not well understood. In fact, it is not even clear that the convergence rates of VB approximate posterior established in [28], [68] can be used to establish statistical guarantees on the decision rules learnt using the näive approach. With an aim to address these gaps, we summarize our contribution in this chapter below:

1. We introduce a minimax optimization framework titled 'risk sensitive variational Bayes' (RSVB), extracted from the dual representation of (SO) using the so-called Donsker-

Varadhan variational free-energy principle [69]. The decision-maker computes a risk-sensitive approximation to the true posterior (termed as RSVB posterior) and the decision rule simultaneously by solving a minimax optimization problem. Moreover, for $\gamma \to 0^+$ and $\gamma = 1$, we recover the naive and LCVB approaches as special cases of RSVB.

2. We identify verifiable regularity conditions on the prior, likelihood model and the risk function under which the RSVB posterior enjoys the same rate of convergence as the true posterior to a Dirac delta distribution concentrated at the true model parameter θ_0, as the sample size increases. Using this result, we also prove the rate of convergence of the RSVB decision rule, when the decision space $\mathcal{A}$ is compact. Moreover, our theoretical results directly imply the asymptotic properties of the LCVB posterior and the associated decision rule. It is also worth noting that our results are applicable to non-parametric problems such as Gaussian process classification, where the parameter space is infinite-dimensional, as well as non independent and identically distributed data generating processes. Moreover, our analysis also recovers consistency and rate of convergence of decision-rules under the 'true' posterior distribution as a special case.

3. We demonstrate our theoretical results with help of three applications:

   (a) First, we consider the classic single-product newsvendor problem and verify all the regularity conditions required to establish the convergence rate of the RSVB posterior and the decision rule. We recover the frequentist rate of convergence $\sqrt{n}$ upto logarithmic factor. Moreover, we present simulation results demonstrating the interplay between the risk-sensitive parameter γ and number of samples n.

   (b) Second, we consider the multi-product newsvendor problem and establish the rate of convergence of the corresponding RSVB posterior and decision rule. Here also, we recover the frequentist rate of convergence $\sqrt{n}$ upto logarithmic factor.

   (c) Finally, we consider a binary Gaussian process classification problem, where the model parameter θ lie in a set of continuous functions on a compact subset of $\mathbb{R}^d$. We construct a wavelet prior and prove all the regularity conditions and

compute the rate of convergence of the RSVB posterior (on function space) and the decision rule. The rate of convergence of the RSVB posterior matches to that of the true posterior as established in Vaart and Zanten [70, Theorem 4.5] for the same wavelet prior.

In our theoretical analyses, we mainly establish three important results. First, in Theorem 2.3.1, we compute a bound on the expected distance of a model from the true model, where expectation is taken with respect to the RSVB posterior. The bound depends on the risk sensitivity parameter γ and the number of samples n, and is a sum of two terms: first one quantifies the rate of convergence of the true posterior and the second one is a consequence of the variational approximation. We further establish regularity conditions on the variational family to compute the rate of convergence of the second term in the bound. In the next two results, we use Theorem 2.3.1 to derive high probability bounds on the optimality gaps in values (Theorem 2.3.2) and decisions (Theorem 2.3.3) computed using the RSVB approach. We define optimality gap in decisions as the deviation of the true optimal decision (when true model is known) from the RSVB decision and define optimality gap in values as the absolute difference between oracle risk $R(\cdot, \theta_0)$ evaluated at true and RSVB decision rules. In our simulation results, we first demonstrate the consistency of the RSVB decision with respect to n for various values of γ. We then demonstrate the effect of changing γ on the optimality gaps and the variance of the RSVB posterior for a given n. In particular, we observe that for smaller n, increasing γ (after a certain value) result into a significantly more risk-averse decision, however the effect of increasing γ on risk-averse decision-making reduces as n increases.

Here's a brief roadmap for the rest of the chapter. In the next section we provide a literature survey of relevant results from machine learning, theoretical statistics and operations research, placing our results in appropriate context. In Section 2.2, we present the problem formulation and introduce RSVB framework with relevant notations, definitions and regularity conditions. We develop our theoretical results in Section 2.3. Thereafter, in Section 2.4, we discuss naive and loss-calibrated VB as special cases of RSVB. We then illustrate the bounds obtained in Section 2.3 by specializing the results to the single and multi-product

newsvendor problem and Gaussian process classification problem in Section 2.5 and also present some numerical results. We end with concluding remarks in Section 6.

2.1 Existing literature and our work

Our work fits in with a growing body of work in operations research that lies at the intersection of decision-making under uncertainty and statistical estimation. Our results are also aligned with recent developments of a rigorous theoretical understanding of variational Bayesian methods in statistics and machine learning.

2.1.1 Operations research literature

The primary goal in data-driven decision-making is to learn empirical decision-rules (or *predictive prescriptions* as Bertsimas and Kallus [71] term them) $a^*(\tilde{X}_n)$ that prescribes a decision, given an observation of the covariates $\tilde{X}_n$. Early work in this direction, including classic work by Herbert Scarf on Bayesian solutions to the newsvendor problem [72], focused on two-stage solutions - estimation followed by optimization. Our setting is most related to recent work on *Bayesian risk optimization* (BRO) in [15], [36]. In BRO, the authors consider optimal decision-making using various coherent risk measures computed under the posterior distribution. The authors establish several important results, including that the optimal values and decisions are asymptotically consistent as the sample size tends to infinity, and central limit theorems for these quantities. However, there are substantial differences with our work. First, all of the analysis in Wu, Zhu, and Zhou [15] presumes that the posterior risk measures are actually computable. The authors do not address the critical computational questions surrounding Bayesian methods or the impact of (inevitable) computational approximations on BRO – indeed, this is not their focus. Second, extended coherent risk measures are not considered (in particular, the log-exponential risk measure used here), and it is unclear if the asymptotic results continue hold otherwise. Third, while we use a risk measure to derive the computational framework (RSVB), the focus in Wu, Zhu, and Zhou [15] is purely on the analytical properties of optimal decisions.

More recently, there has been significant interest in methods that use empirical risk minimization (ERM) or sample average approximation (SAA) for directly estimating decision-rules that optimize Monte Carlo or empirical approximations [71], [73]–[78]. The survey by Homem-de-Mello and Bayraksan [79] consolidates recent results on Monte Carlo methods for stochastic optimization. It is important to note that this recent surge of work in data-driven decision-making has largely focused on explicit black-box models. On the other hand, there are many situations where optimal decisions must be made in the presence of a well-defined parametrized stochastic model. Bayesian methods are a natural means for estimating distributions over the parameters of a stochastic model; though, as noted before, the computational complexity of Bayesian algorithms can be high. The interplay between optimization and estimation, in the sense of discovering predictive prescriptions for Bayesian models has largely been ignored. Furthermore, as Liyanage and Shanthikumar [80] show in the newsvendor context, SEO methods can be suboptimal in terms of expected regret and long-term average losses. Liyanage and Shanthikumar [80] introduced *operational statistics* (OS) as an alternative to SEO (see [19], [81] as well), whereby the optimal empirical order quantity is determined as a function of an optimization parameter that can be determined for each sample size. OS has demonstrably better performance, especially on single parameter newsvendor problems (though there is much less known about its statistical properties).

2.1.2 Statistics and machine learning literature

Lacoste-Julien, Huszár, and Ghahramani [10] observe that calibrating a Gaussian process classification algorithm to a fixed loss function can improve classification performance over a loss-insensitive algorithm – indeed, this is the first documented presentation of the LCVB algorithm. Similarly, surrogate loss functions [82], [83] that are regularized upper bounds that depend on the cost function, also implicitly loss-calibrate frequentist classification algorithms. While standard VB methods for posterior estimation have been extensively used in machine learning [67], it is only recently that the theoretical questions surrounding VB have been addressed. In particular, we note [27] who prove asymptotic consistency of VB in the large sample limit, Zhang and Gao [28] and Pati, Bhattacharya, and Yang [68] on the other

hand establish bounds on the rate of convergence of the VB posterior to the 'true' posterior providing a more refined analysis, and [30] where asymptotic consistency of α-Rényi VB was demonstrated. Our analysis in this chapter, extends these results to establish convergence rates of the approximate posterior and learnt decision rules in risk-sensitive variational Bayesian decision-making framework. These bounds, in turn, are complementary to large sample analyses in Jaiswal, Honnappa, and Rao [84].

2.2 Problem Setup

Let $\xi \in \mathcal{X} \subseteq \mathbb{R}^m$ represent an $\mathbb{R}^m$-valued random variable, with density $p(\cdot|\theta)$ associated with the distribution/model P_θ with parameter $\theta \in \Theta$. Let $(\bigotimes_n \mathcal{X}, \mathcal{S}^n, P_\theta^n)$ be a measure space with sigma-algebra $\mathcal{S}^n$ generated by $\bigotimes_n \mathcal{X}$, where, in general, $\bigotimes_n A$ denote the n-fold product of a set A. Let $\tilde{X}_n := \{\xi_1, \ldots, \xi_n\}$ represent a set of n samples from the true model P_0^n with parameter $\theta_0 \in \Theta$. Denoting the likelihood of observing $\tilde{X}_n$ as $p_\theta^n(\tilde{X}_n)$ and the *prior* distribution $\Pi(\theta)$, we define the *posterior distribution* as $d\Pi(\theta|\tilde{X}_n) = \frac{p_\theta^n(\tilde{X}_n)d\Pi(\theta)}{\int_\Theta p_\theta^n(\tilde{X}_n)d\Pi(\theta)}$. We also write $\Pi(\theta|\tilde{X}_n)$ as Π_n for brevity. Moreover, we denote the corresponding prior and posterior density (if they exist) as $\boldsymbol{\pi}(\cdot)$ and $\boldsymbol{\pi}(\cdot|\tilde{X}_n)$.

As noted in the introduction, our objective is to optimize the posterior log-exponential or entropic risk measure of $R(a, \theta)$, that is

$$\min_{a \in \mathcal{A}} \varrho_{\Pi_n}^\gamma(a) = \frac{1}{\gamma} \log \mathbb{E}_{\Pi_n}[\mathrm{e}^{\gamma R(a,\theta)}], \text{ where } \gamma \in \mathbb{R}. \tag{SO}$$

In practical settings, the posterior $\Pi(\theta|\tilde{X}_n)$ typically cannot be easily computed, and decision makers are often led to restrictive modeling choices such as assuming the likelihood function has a conjugate prior. Indeed, one might argue that this is a predominant reason Bayesian methods are not widely used in operations research and engineering. Nonetheless, incorporating non-conjugate priors and complicated likelihood functions is critical for realizing the full utility of decision-theoretic Bayesian methods - however this entails the use of computational approximations. Therefore, in the next paragraph we introduce a framework from which can be extracted computational methods for approximately computing and optimizing posterior decision risk.

2.2.1 Risk-Sensitive Variational Bayes

Our approach exploits the dual representation of the log-exponential risk measure in (SO), which is convex (or extended coherent) [85], [86]. From the Donsker-Varadhan variational free energy principle [69], [87]–[89] we observe that,

$$\varrho^{\gamma}_{\Pi_n}(a) = \begin{cases} \min_{Q\in\mathcal{M}} \left\{ \mathbb{E}_Q[R(a,\theta)] - \frac{1}{\gamma}\mathrm{KL}(Q\|\Pi_n) \right\} & \gamma < 0, \\ \max_{Q\in\mathcal{M}} \left\{ \mathbb{E}_Q[R(a,\theta)] - \frac{1}{\gamma}\mathrm{KL}(Q||\Pi_n) \right\} & \gamma > 0, \end{cases} \tag{DV}$$

where $\mathcal{M}$ is the set of all distribution functions that are absolutely continuous with respect to the posterior distribution Π_n and 'KL' represents the Kullback-Leibler divergence. Formally, for any two distributions P and Q defined on measurable space $(\Theta, \mathcal{T})$, the KL divergence is defined as

$$\mathrm{KL}(Q\|P) = \begin{cases} \int_\Theta dQ(\theta) \log \frac{dQ(\theta)}{dP(\theta)} & \text{if } Q \ll P, \\ \infty & \text{otherwise }, \end{cases} \tag{2.2}$$

where $Q \ll P$ denotes that measure Q is absolutely continuous with respect to P. Notice that this dual formulation exposes the reason we choose to use the log-exponential risk – the right hand side provides a combined assessment of the risk associated with model estimation (computed by the KL divergence $\mathrm{KL}(Q\|\Pi_n)$) and the decision risk under the estimated posterior Q (computed by $\mathbb{E}_Q[R(a,\theta)]$).

In this thesis, we restrict our analyses to the risk-averse case, that is $\gamma > 0$. However, it can be extended easily to the case when $\gamma < 0$ to obtain similar theoretical insights.

As stated above, the reformulation presented in (DV) offers no computational gains. However, restricting ourselves to an appropriately chosen subset $\mathcal{Q} \subset \mathcal{M}$, that consists of distributions where the integral $\mathbb{E}_q[R(a,\theta)]$ can be tractably computed, we immediately obtain a *risk-sensitive variational Bayesian* (RSVB) formulation of (DV):

$$\frac{1}{\gamma}\log \mathbb{E}_{\Pi_n}\left[\mathrm{e}^{\gamma R(a,\theta)}\right] \geq \max_{Q\in\mathcal{Q}} \left\{ \mathbb{E}_Q[R(a,\theta)] - \frac{1}{\gamma}\mathrm{KL}(Q||\Pi_n) \right\} =: \mathcal{F}(a; Q(\cdot), \tilde{X}_n, \gamma), \tag{RSVB}$$

RSVB is our framework for data-driven risk-sensitive decision-making. The family of distributions $\mathcal{Q}$ is popularly known as the *variational family*. The choice of the family $\mathcal{Q}$, disutility/ risk R, and parameter γ encodes specific problem settings. Our analysis in subsequent Section 2.3.1 below reveals general guidelines on how to choose $\mathcal{Q}$ that ensures a small optimality gap (defined below) with high probability.

With an appropriate choice of $\mathcal{Q}$, the optimization on the RHS can yield a good approximation to the log-exponential risk measurement on the left hand side (LHS). For brevity, for a given $a \in \mathcal{A}$ we define the RSVB approximation to the true posterior $\Pi(\theta|\tilde{X}_n)$ as

$$Q^*_{a,\gamma}(\theta|\tilde{X}_n) := \text{argmax}\{Q \in \mathcal{Q} : \mathcal{F}(a; Q(\cdot), \tilde{X}_n, \gamma)\}$$

and the RSVB optimal decision as

$$\mathtt{a}^*_{\mathtt{RS}} := \text{argmin}_{a\in\mathcal{A}}\ \mathcal{F}(a; Q^*_{a,\gamma}(\theta|\tilde{X}_n), \tilde{X}_n, \gamma).$$

Observe that $Q^*_{a,\gamma}(\theta|\tilde{X}_n)$ and $\mathtt{a}^*_{\mathtt{RS}}$ are random quantities, conditional on the data $\tilde{X}_n$. Intuitively, it can be observed that the risk averseness of $\mathtt{a}^*_{\mathtt{RS}}$ increases with increase in γ. To observe this consider the RSVB formulation and note that $\text{KL} > 0$, therefore as γ increases there is more incentive to deviate from the true posterior and choose $Q \in \mathcal{Q}$ that maximizes expected risk for a given $a \in \mathcal{A}$. Consequently as γ increases, the RSVB decision rule becomes more risk-averse.

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 θ. The choice of $\mathcal{Q}$ is decisive in determining the performance of the algorithm. In general, however the requirements on $\mathcal{Q}$ are minimal, and part of the analysis in this chapter is to articulate sufficient conditions on $\mathcal{Q}$ that ensure small optimality gap (defined below) for the optimal decision, $\mathtt{a}^*_{\mathtt{RS}}$. This establishes the "statistical goodness" of the procedure as number of samples increase. In this chapter, we analyze the efficacy of the decision rules obtained using the RSVB approximation, by providing finite sample probabilistic bounds on the optimality gap. We define the *optimality gap* for any $\mathtt{a} \in \mathcal{A}$ with value $V = R(\mathtt{a}, \theta_0)$ as,

Definition 2.2.1 (Optimality Gap). *Let* $V_0^* := \min_{a\in\mathcal{A}} R(a,\theta_0)$ *and* $a_0^* := argmin_{a\in\mathcal{A}} R(a,\theta_0)$ *be the optimal value and decision respectively for the true model parameter* θ_0*. Then, the optimality gap in the value is the difference* $V - V_0^*$*, and the optimality gap in decision variables is* $\|a_0^* - \mathsf{a}\|$*, where* $\|\cdot\|$ *is the Euclidean norm.*

A similar performance measure was used in [11], to measure the effectiveness of loss-calibrated VB (LCVB) approach, which can be obtained by setting $\gamma = 1$, as a special case of our RSVB formulation. Nonetheless, in Section 2.4, we discuss two well-known variational Bayesian algorithms (one of them is LCVB) for decision making, which are special cases of RSVB. Moreover, we establish bounds on their respective optimality gaps as a corollary to the bounds derived for RSVB.

Note that the RSVB algorithm described above is idealized – clearly the objective $\mathcal{F}(a; Q(\cdot), \tilde{X}_n, \gamma)$ cannot be computed since it requires the calculation of the posterior distribution – the very object we are approximating! Note, however that optimizing $\mathcal{F}(a; Q(\cdot), \tilde{X}_n, \gamma)$ is equivalent to optimizing $\{\gamma\mathbb{E}_Q[R(a,\theta)] - \mathrm{KL}(Q(\theta)\|P(\theta,\tilde{X}_n))\}$, where $P(\theta,\tilde{X}_n)$ is known, and for which the optimizers are the same. Since our focus is on bounding the optimality gap, in the remainder of the chapter any reference to the RSVB algorithm is an allusion to the idealized objective $\mathcal{F}(a; Q(\cdot), \tilde{X}_n, \gamma)$.

In the following section, we lay down important assumptions and definitions used throughout the chapter to establish our theoretical results.

2.2.2 Notations and Definitions

We provide the definitions of important terms used throughout the chapter. First, recall the definition of covering numbers:

Definition 2.2.2 (Covering numbers). *Let* $\mathcal{P} := \{P_\theta, \theta \in \Theta\}$ *be a parametric family of distributions and* $d : \mathcal{P}\times\mathcal{P} \mapsto [0,\infty)$ *be a metric. An* $\epsilon-$*cover of a subset* $\mathcal{P}_K := \{P_\theta : \theta \in K \subset \Theta\}$ *of the parametric family of distributions is a set* $K \subset K$ *such that, for each* $\theta \in K$ *there exists a* $\theta \in K$ *that satisfies* $d(P_\theta, P_\theta) \leq \epsilon$*. The* $\epsilon-$*covering number of* $\mathcal{P}_K$ *is* $N(\epsilon, \mathcal{P}_K, d) = \min\{card(K) : K \text{ is an } \epsilon-\text{cover of } K\}$*, where* $card(\cdot)$ *represents the cardinality of the set.*

Next, recall the definition of a test function [32]:

Definition 2.2.3 (Test function)**.** *Let $\tilde{X}_n$ be a sequence of random variables on measurable space $(\mathbb{R}^{q\times n}, \mathcal{S}^n)$. Then any $\mathcal{S}^n$-measurable sequence of functions $\{\phi_n\}$, $\phi_n : \tilde{X}_n \mapsto [0,1]$ $\forall n \in \mathbb{N}$, is a test of a hypothesis that a probability measure on $\mathcal{S}^n$ belongs to a given set against the hypothesis that it belongs to an alternative set. The test ϕ_n is consistent for hypothesis P_0^n against the alternative $P^n \in \{P_\theta^n : \theta \in \Theta\backslash\{\theta_0\}\}$ if $\mathbb{E}_{P^n}[\phi_n] \to \mathbb{1}_{\{\theta\in\Theta\backslash\{\theta_0\}\}}(\theta), \forall\theta \in \Theta$ as $n \to \infty$, where $\mathbb{1}_{\{\cdot\}}$ is an indicator function.*

A classic example of a test function is $\phi_n^{\text{KS}} = \mathbb{1}_{\{\text{KS}_n > K_\nu\}}(\theta)$ that is constructed using the Kolmogorov-Smirnov statistic $\text{KS}_n := \sup_t |\mathbb{F}_n(t) - \mathbb{F}_\theta(t)|$, where $\mathbb{F}_n(t)$ and $\mathbb{F}_\theta(t)$ are the empirical and true distribution respectively, and K_ν is the confidence level. If the null hypothesis is true, the Glivenko-Cantelli theorem [90, Theorem 19.1] shows that the KS statistic converges to zero as the number of samples increases to infinity.

Furthermore, we define the Hellinger distance between two measures P_{θ_1} and P_{θ_2} as

Definition 2.2.4 (Hellinger distance)**.** *The Hellinger distance $h(\theta_1, \theta_2)$ between the two probability distributions P_{θ_1} and P_{θ_2} is defined as $d_H(\theta_1, \theta_2) = \left(\int\left(\sqrt{dP_{\theta_1}} - \sqrt{dP_{\theta_2}}\right)^2\right)^{1/2}$.*

We define the one-sided Hausdorff distance between sets A and B in $\mathbb{R}^s$ as:

Definition 2.2.5 (Hausdorff distance)**.** *The one-sided Hausdorff distance $H(A\|B)$ between sets A and B in a metric space D with distance function d is defined as:*

$$H(A\|B) = \sup_{x\in A} d_h(x, B), \text{ where } d_h(x, B) = \inf_{y\in B} d(x, y).$$

Next, we define an arbitrary loss function $L_n : \Theta \times \Theta \mapsto \mathbb{R}$ that measures the distance between models $(P_{\theta_1}^n, P_{\theta_2}^n)\forall\{\theta_1, \theta_2\} \in \Theta$. At the outset, we assume that $L_n(\theta_1, \theta_2)$ is always positive. We use the following 'control sequence' to establish our probabilistic bounds.

Definition 2.2.6 (Control Sequence)**.** *$\{\epsilon_n\}$ is a sequence such that $\epsilon_n \to 0$ as $n \to \infty$ and $n\epsilon_n^2 \geq 1$.*

We also define

Definition 2.2.7 ($\Gamma-$convergence)**.** *A sequence of functions $F_n : \mathcal{U} \mapsto \mathbb{R}$, for each $n \in \mathbb{N}$, $\Gamma-$converges to $F : \mathcal{U} \mapsto \mathbb{R}$, if*

- *for every $u \in \mathcal{U}$ and every $\{u_n, n \in \mathbb{N}\}$ such that $u_n \to u$, $F(x) \leq \liminf_{n\to\infty} F_n(u_n)$;*
- *for every $u \in \mathcal{U}$, there exists some $\{u_n, n \in \mathbb{N}\}$ such that $u_n \to u$, $F(x) \geq \limsup_{n\to\infty} F_n(u_n)$.*

In addition, we define

Definition 2.2.8 (Primal feasibility)**.** *For any two functions $f : \mathcal{U} \mapsto \mathbb{R}$ and $b : \mathcal{U} \mapsto \mathbb{R}$, a point $u^* \in \mathcal{U}$ is primal feasible to the following constraint optimization problem*

$$\inf_{u\in\mathcal{U}} f(u) \text{ subject to } b(u) \leq c,$$

if $b(u^*) \leq c$, for a given $c \in \mathbb{R}$.

2.2.3 Assumptions

In order to bound the optimality gap, we require some control over how quickly the posterior distribution concentrates at the true parameter θ_0. Our next assumption in terms of a verifiable test condition on the model (sub-)space is one of the conditions required to quantify this rate.

Assumption 2.2.1 (Model indentifiability)**.** *Fix $n \geq 1$. Then, for any $\epsilon > \epsilon_n$ in Definition 2.2.6, $\exists$ a test function $\phi_{n,\epsilon} : \tilde{X}_n \mapsto [0, 1]$ and sieve set $\Theta_n(\epsilon) \subseteq \Theta$ such that*

$$\text{(i) } \mathbb{E}_{P_0^n}[\phi_{n,\epsilon}] \leq C_0 \exp(-Cn\epsilon^2), \text{ and } \text{(ii) } \sup_{\{\theta\in\Theta_n(\epsilon):L_n(\theta,\theta_0)\geq C_1 n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1 - \phi_{n,\epsilon}] \leq \exp(-Cn\epsilon^2).$$

Observe that Assumption 2.2.1(i) quantifies the rate at which a type 1 error diminishes with the sample size, while the condition in Assumption 2.2.1(ii) quantifies that of a type 2 error. Notice that both of these are stated through test functions; indeed, what is required are consistent test functions. Opportunely, [91, Theorem 7.1] (stated below in Lemma 2.3.1 for completeness) roughly implies that a bounded model subspace $\{P_\theta^n, \theta \in \Theta\}$ (the size of which is measured using covering numbers) guarantees the existence of *consistent* test functions,

to test the null hypothesis that the true parameter is θ_0 against an alternate hypothesis – the alternate being defined using the 'distance function' $L_n(\theta_1, \theta_2)$. Subsequently, we will use a specific distance function to obtain finite sample bounds for the optimality gap in decisions and values. In some problem instances, it is also possible to construct consistent test functions directly without recourse to Lemma 2.3.1. We demonstrate this in Section 2.5.1 below.

Next, we assume a condition on the prior distribution that ensures that it provides sufficient mass to the set $\Theta_n(\epsilon) \subseteq \Theta$, as defined above in Assumption 2.2.1.

Assumption 2.2.2. *Fix $n \geq 1$. Then, for any $\epsilon > \epsilon_n$ in Definition 2.2.6 the prior distribution satisfies*

$$\Pi(\Theta_n^c(\epsilon)) \leq \exp(-Cn\epsilon^2).$$

Notice that Assumption 2.2.2 is trivially satisfied if $\Theta_n(\epsilon) = \Theta$. The next assumption ensures that the prior distribution places sufficient mass around a neighborhood – defined using Rényi divergence – of the true parameter θ_0.

Assumption 2.2.3 (Prior thickness)**.** *Fix $n \geq 1$ and a constant $\lambda > 0$. Let $A_n := \left\{\theta \in \Theta : D_{1+\lambda}\left(P_0^n \| P_\theta^n\right) \leq C_3 n\epsilon_n^2\right\}$, where $D_{1+\lambda}\left(P_0^n \| P_\theta^n\right) := \frac{1}{\lambda}\log\int\left(\frac{dP_0^n}{dP_\theta^n}\right)^\lambda dP_0^n$ is the Rényi Divergence between P_0^n and P_θ^n, assuming P_0^n is absolutely continuous with respect to P_θ^n. The prior distribution satisfies*

$$\Pi(A_n) \geq \exp(-nC_2\epsilon_n^2).$$

Notice that the set A_n defines a neighborhood of the distribution corresponding to θ_0 in the model subspace $\{P_\theta^n : \theta \in \Theta\}$. The assumption guarantees that the prior distribution covers this neighborhood with positive mass. This is a standard assumption and if it is violated then the posterior too will place no mass in this neighborhood ensuring asymptotic inconsistency. The above three assumptions are adopted from [91] and has also been used in [28] to prove convergence rates of variational posteriors. Interested readers may refer to [91] and [28] to read more about the above assumptions.

It is apparent by the first term in (RSVB) that in addition to Assumption 2.2.1, 2.2.2, and 2.2.3, we also require regularity conditions on the risk function $R(a,\cdot)$. Thus, the next assumption restricts the prior distribution with respect to $R(a,\theta)$.

Assumption 2.2.4. *Fix $n \geq 1$ and $\gamma > 0$. For any $\epsilon > \epsilon_n$, $a \in \mathcal{A}$,*

$$\mathbb{E}_\Pi[\mathbb{1}_{\{\gamma R(a,\theta) > C_4(\gamma) n\epsilon^2\}} \mathrm{e}^{\gamma R(a,\theta)}] \leq \exp(-C_5(\gamma) n\epsilon^2),$$

where $C_4(\gamma)$ and $C_5(\gamma)$ are scalar positive functions of γ.

Note that the set $\{\gamma R(a,\theta) > C_4(\gamma) n\epsilon^2\}$ represents the subset of the model space where the risk $R(a,\theta)$ (for a fixed decision a) is large, and the prior is assumed to place small mass over such sets. Moreover, using Cauchy-Schwarz inequality observe that

$$\begin{aligned}\mathbb{E}_\Pi[\mathbb{1}_{\{\gamma R(a,\theta) > C_4(\gamma) n\epsilon^2\}} \mathrm{e}^{\gamma R(a,\theta)}] &\leq \left(\mathbb{E}_\Pi[\mathbb{1}_{\{\gamma R(a,\theta) > C_4(\gamma) n\epsilon^2\}}]\right)^{1/2} \left(\mathbb{E}_\Pi[\mathrm{e}^{2\gamma R(a,\theta)}]\right)^{1/2} \\ &\leq \mathrm{e}^{-C_4(\gamma) n\epsilon_n^2} \mathbb{E}_\Pi[\mathrm{e}^{2\gamma R(a,\theta)}],\end{aligned}$$

which implies that if the risk function is bounded in (a,θ), then above condition can be trivially satisfied. Finally, we also require the following condition lower bounding the risk function R.

Assumption 2.2.5. *$R(a,\theta)$ is assumed to satisfy*

$$W := \inf_{\theta \in \Theta} \inf_{a \in \mathcal{A}} \mathrm{e}^{R(a,\theta)} > 0.$$

Note that any risk function which is bounded from below in both the arguments satisfies this condition. Furthermore, following [92] we define a *growth condition* on the 'true' risk function $R(a,\theta_0)$.

Assumption 2.2.6 (Growth condition). *Let $\Psi(d) : [0, \infty) \mapsto [0, \infty)$ be a growth function if it is strictly increasing as $d \to \infty$ and $\lim_{d\to 0} \Psi(d) = 0$. Then for any $A \subset \mathcal{A}$, $R(a, \theta_0)$ satisfies a growth condition with respect to $\Psi(\cdot)$, if*

$$R(a, \theta_0)) \geq \inf_{z\in\mathcal{A}} R(z, \theta_0) + \Psi\left(H\left(A, \arg\min_{z\in\mathcal{A}} R(z, \theta_0)\right)\right). \tag{2.3}$$

The growth condition above is a generalization of strong-convexity. Indeed, if the true risk is strongly convex, then this condition is automatically satisfied.

In the next, section we derive finite sample bounds on the optimality gap in values and decisions, by proving a series of results.

2.3 Asymptotic Analysis of the Optimality Gaps

In this section, we establish high-probability bounds on the optimality gap in values and decision rules computed using RSVB approach for sufficiently large n. Our results in here identify the regularity conditions on the data generating model $\{P^n_\theta, \theta \in \Theta\}$, the prior distribution $\Pi(\theta)$, the variational family $\mathcal{Q}$, the risk function $R(a, \theta)$ to compute the bounds.

We can now state our first result, establishing an upper bound on the expected deviation from the true model P_0, measured using distance function $L_n(\cdot, \theta_0)$, under the RSVB approximate posterior. We also note that the following result generalizes Theorem 2.1 of [28], which is exclusively for the case when $\gamma \to 0^+$. However, the proof techniques are motivated from the proof of Theorem 2.1 in [28].

Theorem 2.3.1. *Fix $a \in \mathcal{A}$ and $\gamma > 0$. For any $L_n(\theta, \theta_0) \geq 0$, under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5, and for $\min(C, C_4(\gamma) + C_5(\gamma)) > C_2 + C_3 + C_4(\gamma) + 2$ and*

$$\eta^R_n(\gamma) := \frac{1}{n} \inf_{Q\in\mathcal{Q}} \mathbb{E}_{P^n_0}\left[\mathrm{KL}(Q(\theta)\|\Pi(\theta|\tilde{X}_n)) - \gamma \inf_{a\in\mathcal{A}} \mathbb{E}_Q[R(a, \theta)]\right],$$

the RSVB approximator of the true posterior $Q^*_{a,\gamma}(\theta|\tilde{X}_n)$ satisfies,

$$\mathbb{E}_{P^n_0}\left[\int_\Theta L_n(\theta, \theta_0) dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \leq n\left(M(\gamma)\epsilon^2_n + M\eta^R_n(\gamma)\right), \tag{2.4}$$

for a positive mapping $M(\gamma) = 2\left(C_1 + MC_4(\gamma)\right)$ *, where* $M = \frac{2C_1}{\min(C,\lambda,1)}$ *, for sufficiently large* n*.*

First recall that ϵ_n is the convergence rate of the true posterior [91, Theorem 7.3]. Notice that the additional term $\eta_n^R(\gamma)$ emerges from the posterior approximation and depends on the choice of the variational family $\mathcal{Q}$, risk function $R(\cdot,\cdot)$, and the parameter γ. The appearance of this term in the bound also signifies that, to minimize expected gap between true model and any other model, defined using $n^{-1}L_n(\theta,\theta_0)$, under the RSVB posterior, the average (with respect to P_0^n) RSVB objective has to be maximized. Later in this section, we specify the conditions on the family of distributions $\{P_\theta^n, \theta \in \Theta\}$, the prior and the variational family $\mathcal{Q}$ that ensure $\eta_n^R(\gamma) \to 0$ as $n \to \infty$. Moreover, we also identify mild regularity conditions on $\mathcal{Q}$ to show that $\eta_n^R(\gamma)$ is $O(\epsilon_n^2)$. Furthermore, we show that as γ increases $\eta_n^R(\gamma)$ decreases. We discuss this result and the bound therein later in the next subsection. Before that, we establish our main result (the bounds on the optimality gap) using the theorem above.

Since the result in Theorem 2.3.1 holds for any positive distance function, we now fix

$$L_n(\theta,\theta_0) = n\left(\sup_{a\in\mathcal{A}} |R(a,\theta) - R(a,\theta_0)|\right)^2. \tag{2.5}$$

Notice that for a given θ, $n^{-1/2}\sqrt{L_n(\theta,\theta_0)}$ is the uniform distance between the $R(a,\theta)$ and $R(a,\theta_0)$. Intuitively, Theorem 2.3.1 implies that the expected uniform difference $\frac{1}{n}L_n(\theta,\theta_0)$ with respect to the RSVB approximate posterior is $O(M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma))$, and if $M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma) \to 0$ as $n \to \infty$ then it converges to zero at that rate.

Also, note that in order to use (2.5) we must demonstrate that it satisfies Assumption 2.2.1. This can be achieved by constructing bespoke test functions for a given $R(a,\theta)$. We demonstrate this approach by an example in Section 2.5.2. Nonetheless, we also provide sufficient conditions for the existence of the test functions later in the section. These conditions are typically easy to verify when the loss function $R(\cdot,\cdot)$ are bounded, for instance.

Now, we first bound the optimality gap between $R(\mathtt{a}^*_{\mathtt{RS}},\theta_0)$ and V_0^*.

Theorem 2.3.2. *Fix $\gamma > 0$. Suppose that the set $\mathcal{A}$ is compact. Then, under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5, for $\min(C, C_4(\gamma) + C_5(\gamma)) > C_2 + C_3 + C_4(\gamma) + 2$ and for any $\tau > 0$, the P_0^n- probability of the following event*

$$\left\{ \tilde{X}_n : R(\mathtt{a}^*_{\mathtt{RS}}, \theta_0) - \inf_{z \in \mathcal{A}} R(z, \theta_0) \leq 2\tau \left[M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma) \right]^{\frac{1}{2}} \right\} \tag{2.6}$$

is at least $1 - \tau^{-1}$, for a positive mapping $M(\gamma) = 2\left(C_1 + MC_4(\gamma)\right)$, where $M = \frac{2C_1}{\min(C, \lambda, 1)}$ for sufficiently large n.

Next, we bound the optimality gap between the approximate optimal decision rule $\mathtt{a}^*_{\mathtt{RS}}$ and the true optimal decision. The bound, in particular, depends on the curvature of $R(a, \theta_0)$ around the true optimal decision, defined using the growth condition in Assumption 2.2.6.

Theorem 2.3.3. *Fix $\gamma > 0$. Suppose that the set $\mathcal{A}$ is compact and $R(a, \theta_0)$ satisfies the growth condition in Assumption 2.2.6, with $\Psi(d)$ such that $\Psi(d)/d^\delta = \kappa$, for any $\delta > 0$. Then, under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5, for $\min(C, C_4(\gamma) + C_5(\gamma)) > C_2 + C_3 + C_4(\gamma) + 2$ and for any $\tau > 0$, the P_0^n- probability of the following event*

$$\left\{ \tilde{X}_n : H\left(\mathtt{a}^*_{\mathtt{RS}}(\tilde{X}_n), \arg\min_{z \in \mathcal{A}} R(z, \theta_0) \right) \leq \left[\frac{2\tau \left[M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma) \right]^{\frac{1}{2}}}{\kappa} \right]^{\frac{1}{\delta}} \right\}$$

is at least $1 - \tau^{-1}$, for a positive mapping $M(\gamma) = 2\left(C_1 + MC_4(\gamma)\right)$, where $M = \frac{2C_1}{\min(C, \lambda, 1)}$ for sufficiently large n.

To fix the intuition, suppose $\delta = 2$ and $\Psi(d) = \frac{\kappa}{2}d^2$, then κ represents the Hessian of the true risk, $R(a, \theta_0)$, near its optimizer. It is easy to see from the above result the rate of convergence of $\mathtt{a}^*_{\mathtt{RS}}$ is scaled by a factor κ^{-1}. That is, higher the curvature near the optimizer, the faster $\mathtt{a}^*_{\mathtt{RS}}$ converges.

Evidently, the bounds obtained in all three results that we have proved so far depends on $\eta_n^R(\gamma)$. Consequently, in the next section, with an aim to understand the properties of the bounds in Theorem 2.3.1, 2.3.2, and 2.3.3, we prove some of the important properties of $\eta_n^R(\gamma)$ with respect to n and γ under some additional regularity conditions.

2.3.1 Properties of $\eta_n^R(\gamma)$

In order to characterize $\eta_n^R(\gamma)$, we specify conditions on variational family $\mathcal{Q}$ such that $\eta_n^R(\gamma) = O(\epsilon_n^2)$, for some $\epsilon_n \geq \frac{1}{\sqrt{n}}$ and $\epsilon_n \to 0$. We impose following condition on the variational family $\mathcal{Q}$ that lets us obtain a bound on $\eta_n^R(\gamma)$ in terms of n and γ.

Assumption 2.3.1. *There exists a sequence of distribution $\{q_n(\cdot)\}$ in the variational family $\mathcal{Q}$ such that for a positive constant C_9,*

$$\frac{1}{n}\left[\mathrm{KL}\left(Q_n(\theta)\|\Pi(\theta)\right) + \mathbb{E}_{Q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n(\tilde{X}_n)\|dP_\theta^n(\tilde{X}_n)\right)\right]\right] \leq C_9\epsilon_n^2. \tag{2.7}$$

If the observations in $\tilde{X}_n$ are i.i.d, then observe that

$$\frac{1}{n}\mathbb{E}_{Q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n(\tilde{X}_n))\|dP_\theta^n(\tilde{X}_n)\right)\right] = \mathbb{E}_{Q_n(\theta)}\left[\mathrm{KL}\left(dP_0)\|dP_\theta(\xi)\right)\right].$$

Intuitively, this assumption implies that the variational family must contain a sequence of distributions that converges weakly to a Dirac delta distribution concentrated at the true parameter θ_0 otherwise the second term in the LHS of (2.7) will be non-zero. Also note that the above assumption does not imply that the minimizing sequence $Q_{a,\gamma}^*(\theta|\tilde{X}_n)$ (automatically) converges weakly to a dirac-delta distribution at the true parameter θ_0. Furthermore, unlike Theorem 2.3 of [28], our condition on $\mathcal{Q}$ in Assumption 2.3.1, to obtain a bound on $\eta_n^R(\ \gamma)$, does not require the support of the distributions in $\mathcal{Q}$ to shrink to the true parameter θ_0 at some appropriate rate, as the numbers of samples increases.

Proposition 2.3.1. *Under Assumption 2.3.1 and for a constant $C_8 = -\inf_{Q\in\mathcal{Q}}\inf_{a\in\mathcal{A}}\mathbb{E}_Q[R(a,\theta)]$ and $C_9 > 0$,*

$$\eta_n^R(\gamma) \leq \gamma n^{-1}C_8 + C_9\epsilon_n^2.$$

In Section 2.5, we present an example where the likelihood is exponentially distributed, the prior is inverse-gamma (non-conjugate), and the variational family is the class of gamma distributions, where we construct a sequence of distributions in the variational family that satisfies Assumption 2.3.1. We also provide another example where the likelihood is multivariate Gaussian with unknown mean and variational family is uncorrelated Gaussian re-

stricted to compact subset of $\mathbb{R}^d$ with an uniform prior on the same compact set satisfy Assumption 2.3.1.

By definition $\epsilon_n^2 \to 0$ and $\epsilon_n \to 0$ as $n \to \infty$, and therefore it follows from Proposition 2.3.1 that $M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma) \to 0$. However, the bound obtained in the last proposition might be loose with respect to γ, when $C_8 < 0$. To see this, we prove the following result.

Proposition 2.3.2. *If the solution to the optimization problem in $\eta_n^R(\gamma)$ is primal feasible then $\eta_n^R(\gamma)$ decreases as γ increases.*

Our next result shows that, under the RSVB approximate posterior distribution $Q_{a,\gamma}^*(\theta|\tilde{X}_n)$, $L_n(\cdot,\cdot)$ as defined in (2.5) converges to zero at the rate $(\epsilon_n^2 + \eta_n^R(\gamma))$ in P_0^n-probability. Here, $Q_{a,\gamma}^*(S|\tilde{X}_n) := \int_S dQ_{a,\gamma}^*(\theta|\tilde{X}_n)$, for any $S \subseteq \Theta$.

Corollary 2.3.1. *For any $a \in \mathcal{A}, \gamma > 0$, and diverging sequence M_n, under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, 2.2.5, and 2.3.1, for $\min(C, C_4(\gamma) + C_5(\gamma)) > C_2 + C_3 + C_4(\gamma) + 2$,*

$$\lim_{n\to\infty} Q_{a,\gamma}^* \left[\left\{ \theta \in \Theta : \left(\sup_{a\in\mathcal{A}} |R(a,\theta) - R(a,\theta_0)| \right)^2 > M_n(M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma)) \right\} \middle| \tilde{X}_n \right] = 0$$

in P_0^n-probability, where $M(\gamma) = 2\left(C_1 + MC_4(\gamma)\right)$, and $M = \frac{2C_1}{\min(C,\lambda,1)}$.

Observe that if $\sum_{n\geq 1} \frac{1}{M_n} < \infty$, then the first Borel-Contelli Lemma [93, Theorem 2.3.1] implies that the sequence converges almost-surely. First, recall from Theorem 2.3.1 that $\epsilon_n \to 0$ as $n \to \infty$ and $n\epsilon_n^2 \geq 1$. The diverging sequence M_n can be chosen in three possible ways. First, $M_n = o\left(\frac{1}{(M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma))^b}\right)$, for some $b < 1$, which ensures that the radius of the ball in Corollary 2.3.1 decreases to 0 as $n \to \infty$. Second, $M_n = \left(\frac{1}{M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma)}\right)$, in this case ball will be of constant radius 1. Also observe that in the last two cases $\sum_{n\geq 1} \frac{1}{M_n} = \infty$, since ϵ_n^2 is not summable, therefore we do not have almost-sure convergence in these cases. In the final case, $M_n = o\left(\frac{1}{(M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma))^b}\right)$ for any $b > 1$ is summable, since, $n\eta_n^R(\gamma) < \infty$ due to Assumption 2.2.5 and Proposition 2.3.1. Note that, in this case the radius of the ball will diverge and hence we obtain almost-sure convergence.

2.3.2 Sufficient conditions on the risk function for existence of tests

To show the existence of test functions, as required in Assumption 2.2.1, we will use the following result from Ghosal, Ghosh, and Vaart [91, Theorem 7.1], that is applicable only to distance measures that are bounded above by the Hellinger distance.

Lemma 2.3.1 (Theorem 7.1 of [91]). *Suppose that for some non-increasing function $D(\epsilon)$, some $\epsilon_n > 0$ and for every $\epsilon > \epsilon_n$,*

$$N\left(\frac{\epsilon}{2}, \{P_\theta : \epsilon \leq m(\theta, \theta_0) \leq 2\epsilon\}, m\right) \leq D(\epsilon),$$

where $m(\cdot, \cdot)$ is any distance measure bounded above by Hellinger distance. Then for every $\epsilon > \epsilon_n$, there exists a test ϕ_n (depending on $\epsilon > 0$) such that, for every $\mathrm{j} \geq 1$,

$$\mathbb{E}_{P_0^n}[\phi_n] \leq D(\epsilon) \exp\left(-\frac{1}{2} n\epsilon^2\right) \frac{1}{1 - \exp\left(-\frac{1}{2} n\epsilon^2\right)}, \text{ and}$$

$$\sup_{\{\theta \in \Theta_n(\epsilon) : m(\theta, \theta_0) > \mathrm{j}\epsilon\}} \mathbb{E}_{P_\theta^n}[1 - \phi_n] \leq \exp\left(-\frac{1}{2} n\epsilon^2 \mathrm{j}\right).$$

For the remaining part of this subsection we assume that $\Theta \subseteq \mathbb{R}^d$. In the subsequent paragraph, we state further assumptions on the risk function to show $L_n(\cdot, \cdot)$ as defined in (2.5) satisfies Assumption 2.2.1. For brevity we denote $n^{-1/2}\sqrt{L_n(\theta, \theta_0)}$ by $d_L(\theta, \theta_0)$, that is

$$d_L(\theta_1, \theta_2) := \sup_{a \in \mathcal{A}} |R(a, \theta_1) - R(a, \theta_2)|, \ \forall \{\theta_1, \theta_2\} \in \Theta \tag{2.8}$$

and the covering number of the set $T(\epsilon) := \{P_\theta : d_L(\theta, \theta_0) < \epsilon\}$ as $N(\delta, T(\epsilon), d_L)$, where $\delta > 0$ is the radius of each ball in the cover. We assume that the risk function $R(a, \cdot)$ satisfies the following bound.

Assumption 2.3.2. *The model risk satisfies*

$$d_L(\theta_1, \theta_2)| \leq K_1 d_H(\theta, \theta_0),$$

where $d_H(\theta_1, \theta_2)$ is the Hellinger distance between two models P_{θ_1} and P_{θ_2}.

For instance, suppose the definition of model risk is $R(a,\theta) = \int_{\mathcal{X}} \ell(x,a)p(y|\theta)dx$, where $\ell(x,a)$ is an underlying loss function. Then, observe that Assumption 2.3.2 is trivially satisfied if $\ell(x,a)$ is bounded in x for a given $a \in \mathcal{A}$ and $\mathcal{A}$ is compact, since $d_L(\theta_1,\theta_2)$ can be bounded by the total variation distance $d_{TV}(\theta_1,\theta_2) = \frac{1}{2}\int |dP_{\theta_1}(x) - dP_{\theta_2}(x)|$ and total variation distance is bounded above by the Hellinger distance [94]. Under the assumption above it also follows that we can apply Lemma 2.3.1 to the metric $d_L(\cdot,\cdot)$ defined in (2.8). Now, we will also assume an additional regularity condition on the risk function.

Assumption 2.3.3. *For every $\{\theta_1,\theta_2\} \in \Theta$, there exists a constant $K_2 > 0$ such that*

$$d_L(\theta_1,\theta_2) \leq K_2\|\theta_1 - \theta_2\|,$$

We can now show that the covering number of the set $T(\epsilon)$ satisfies

Lemma 2.3.2. *Given $\epsilon > \delta > 0$, and under Assumption 2.3.3,*

$$N(\delta, T(\epsilon), d_L) < \left(\frac{2\epsilon}{\delta} + 2\right)^d. \tag{2.9}$$

Observe that the RHS in (2.9) is a decreasing function of δ, infact for $\delta = \epsilon/2$, it is a constant in ϵ. Therefore, using Lemmas 2.3.1 and 2.3.2, we show in the following result that $L_n(\theta,\theta_0)$ in (2.5) satisfies Assumption 2.2.1.

Lemma 2.3.3. *Fix $n \geq 1$. For a given $\epsilon_n > 0$ and every $\epsilon > \epsilon_n$, such that $n\epsilon_n^2 \geq 1$. Under Assumption 2.3.2 and 2.3.3, $L_n(\theta,\theta_0) = n\left(\sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|\right)^2$ satisfies*

$$\mathbb{E}_{P_0^n}[\phi_n] \leq C_0 \exp(-Cn\epsilon^2), \tag{2.10}$$

$$\sup_{\{\theta\in\Theta: L_n(\theta,\theta_0)\geq C_1 n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_n] \leq \exp(-Cn\epsilon^2), \tag{2.11}$$

where $C_0 = 2 * 10^s$ and $C = \frac{C_1}{2K_1^2}$ for a constant $C_1 > 0$.

Since $L_n(\theta,\theta_0) = \frac{1}{n}d_L^2$ satisfies Assumption 2.2.1, Theorem 2.3.1 implies the following finite sample bound.

Corollary 2.3.2. *Fix $a \in \mathcal{A}$ and $\gamma > 0$. Let ϵ_n be a sequence such that $\epsilon_n \to 0$ as $n \to \infty$, $n\epsilon_n^2 \geq 1$ and*

$$L_n(\theta, \theta_0) = n \left(\sup_{a \in \mathcal{A}} |R(a, \theta) - R(a, \theta_0)| \right)^2 .$$

Then under the Assumptions of Theorem 2.3.1 and Lemma 2.3.3 ; for $C = \frac{C_1}{2K_1^2}$, $C_0 = 2 * 10^s$, $C_1 > 0$ such that $\min(C, C_4(\gamma) + C_5(\gamma)) > C_2 + C_3 + C_4(\gamma) + 2$, and for $\eta_n^R(\gamma)$ as defined in Theorem 2.3.1, the RSVB approximator of the true posterior $Q_{a,\gamma}^*(\theta|\tilde{X}_n)$ satisfies,

$$\mathbb{E}_{P_0^n} \left[\int_\Theta L_n(\theta, \theta_0) dQ_{a,\gamma}^*(\theta|\tilde{X}_n) \right] \leq n(M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma)), \tag{2.12}$$

for sufficiently large n and for a function $M(\gamma) = 2\left(C_1 + MC_4(\gamma)\right)$, where $M = \frac{2C_1}{\min(C, \lambda, 1)}$.

2.4 Special Cases of RSVB

Recall from the RSVB formulation that γ encodes the risk sensitivity of the decision maker. In this section, we show that RSVB generalizes two well-known variational Bayesian approaches for decision making, *'naive' VB* (NVB) and *loss-calibrated VB*(LCVB). In particular, the RSVB method is equivalent to NVB when $\gamma \to 0^+$ and LCVB for $\gamma = 1$. In what follows, we discuss NVB and LCVB briefly and demonstrate our theoretical results to these settings.

2.4.1 Naive VB

The naive VB (NVB) method, summarized below in Algorithm 1, is a *"separated estimation and optimization"* method wherein we use the VB approximation to the posterior distribution as a plug-in estimator for computing the posterior predictive loss, and then optimize the resulting approximate posterior predictive loss.

The NVB method completely isolates the statistical estimation problem from the decision-making problem. Observe that as $\gamma \to 0^+$, $Q_{a,\gamma}^*(\theta|\tilde{X}_n)$ and $\mathtt{a}_{\mathtt{RS}}^*$ converges to $Q^*(\theta|\tilde{X}_n)$ and $\mathtt{a}_{\mathtt{NV}}^*$ respectively; that is

Algorithm 1: Naive VB

Input : $R(\cdot,\cdot),\ \tilde{X}_n, \mathcal{Q}$

Output: $\mathtt{a}^*_{\mathtt{NV}}$

Step 1. Compute approximate posterior:
$Q^*(\theta|\tilde{X}_n) := \arg\min_{Q\in\mathcal{Q}} \mathrm{KL}(Q(\cdot)\|\Pi(\cdot|\tilde{X}_n))$;

Step 2. Compute: $\mathtt{a}^*_{\mathtt{NV}} := \arg\min_{a\in\mathcal{A}} \mathbb{E}_{Q^*(\theta|\tilde{X}_n)}[R(a,\theta)]$.

$$\begin{aligned}\lim_{\gamma\to 0^+} Q^*_{a,\gamma}(\theta|\tilde{X}_n) &= \lim_{\gamma\to 0^+} \mathrm{argmax}_{Q\in\mathcal{Q}} \left\{\mathbb{E}_Q[R(a,\theta)] - \frac{1}{\gamma}\mathrm{KL}(Q||\Pi_n)\right\} \\ &= \mathrm{argmin}_{\tilde{Q}\in\mathcal{Q}}\mathrm{KL}(\tilde{Q}(\theta)\|\Pi(\theta|\tilde{X}_n)) := Q^*(\theta|\tilde{X}_n).\end{aligned}$$

To see this, recall the RSVB formulation and multiply by $\gamma > 0$ on either side to obtain:

$$\begin{aligned}\log \mathbb{E}_{\Pi_n}\left[\exp(\gamma R(a,\theta))\right] &\geq \max_{Q\in\mathcal{Q}} \left\{\gamma\mathbb{E}_Q[R(a,\theta)] - \mathrm{KL}(Q||\Pi_n)\right\} \\ &= -\min_{Q\in\mathcal{Q}} \left\{\mathrm{KL}(Q||\Pi_n) - \gamma\mathbb{E}_Q[R(a,\theta)]\right\}.\end{aligned} \tag{2.13}$$

Note that, since $\mathrm{KL}(Q||\Pi_n) - \gamma\mathbb{E}_Q[R(a,\theta)]$ converges uniformly in γ to $\mathrm{KL}(Q||\Pi_n)$ as $\gamma \to 0^+$, therefore former $\Gamma-$converges to the latter and hence their respective minimizers and minimum values [95]. In particular, to prove the uniform convergence, let $\{r_k\}$ be a sequence of rational numbers on $\mathbb{R}^+$, such that $\{r_k\}, k \in \mathbb{N}$ is dense in $\mathbb{R}^+$ and $r_k \to 0^+$ as $k \to \infty$. Now observe that for every $\epsilon > 0$ and given $a \in \mathcal{A}$ and $\theta \in \Theta$, there exists a $K \in \mathbb{N}$, such that for all $k \geq K$ and $Q \in \mathcal{Q}$, $|r_k\mathbb{E}_Q[R(a,\theta)]| < \epsilon$, hence uniform convergence follows.

Now taking limit $\gamma \to 0^+$, the equation (2.13) reduces to the well known *evidence lower bound* [67] , that is

$$0 \geq \max_{Q\in\mathcal{Q}} \left\{-\mathrm{KL}(Q||\Pi_n)\right\} \equiv \log \int_\Theta dP^n_\theta(\tilde{X}_n)\Pi(\theta)d\theta \geq \max_{Q\in\mathcal{Q}} \left\{-\mathrm{KL}(Q(\theta)\|\Pi(\theta)) + \mathbb{E}_Q[dP^n_\theta(\tilde{X}_n)]\right\},$$

where $\Pi(\theta)$ is the prior density. Therefore, it follows that for any $\gamma > 0$

$$\mathrm{argmin}_{a\in A} \left\{\mathbb{E}_{Q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \frac{1}{\gamma}\mathrm{KL}(Q^*(\theta|\tilde{X}_n)\|\Pi(\theta|\tilde{X}_n))\right\} = \mathrm{argmin}_{a\in A}\mathbb{E}_{Q^*(\theta|\tilde{X}_n)}[R(a,\theta)] = \mathtt{a}^*_{\mathtt{NV}}.$$

Since $\lim_{\gamma\to 0^+}\gamma R(\cdot,\cdot)=0$, we do not require Assumption 2.2.4 and 2.2.5 to obtain an analogous result to Theorem 2.3.2 for NVB method. Therefore, the condition on the constants in Theorem 2.3.2 ($\min(C, C_4(\gamma)+C_5(\gamma)) > C_2+C_3+C_4(\gamma)+2$) is simplified to $C > C_2+C_3+2$ by choosing $C_4(\gamma)$ as a small and $C_5(\gamma)$ as a large number.

Theorem 2.4.1. *Let ϵ_n be a sequence such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ as $n\to\infty$ and*

$$L_n(\theta,\theta_0) = n\left(\sup_{a\in\mathcal{A}}|R(a,\theta)-R(a,\theta_0)|\right)^2.$$

Then under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5, and for $C > C_2+C_3+2$ the NVB approximation of the true posterior satisfies,

$$\mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0)dQ^*(\theta|\tilde{X}_n)\right] \le \bar{M}n(\epsilon_n^2+\eta_n(0)), \tag{2.14}$$

where positive constant $\bar{M}$ depends only on C, C_0, C_1, and λ, and

$$\eta_n(0) := \eta_n^R(0) = \frac{1}{n}\inf_{Q\in\mathcal{Q}}\mathbb{E}_{P_0^n}[\mathrm{KL}(Q(\theta)\|\Pi(\theta|\tilde{X}_n))].$$

The next result establishes a bound on the optimality gap of the naive VB estimated optimal value $R(\mathtt{a}^*_{\mathtt{NV}},\theta_0)$ from the true optimal value $V_0=\inf_{z\in\mathcal{A}}R(z,\theta_0)$.

Theorem 2.4.2. *Suppose that the set $\mathcal{A}$ is compact and Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5 are satisfied with $C > C_2+C_3+2$. Then for any $\tau>0$, the P_0^n- probability of the following event*

$$\left\{\tilde{X}_n : R(\mathtt{a}^*_{\mathtt{NV}},\theta_0) - \inf_{z\in\mathcal{A}}R(z,\theta_0) \le 2\tau\left[\bar{M}(\epsilon_n^2+\eta_n(0))\right]^{\frac{1}{2}}\right\} \tag{2.15}$$

is at least $1-\tau^{-1}$, where $\bar{M}$ is a positive constant.

Next, we bound the optimality gap between the approximate optimal decision rule $\mathtt{a}^*_{\mathtt{NV}}$ and the true optimal decision. The bound, in particular, depends on the curvature of $R(a,\theta_0)$ around the true optimal decision. The growth function is denoted as $\Psi(\cdot)$. The following theorem is a special case of the general result for $\mathtt{a}^*_{\mathtt{RS}}$ in Theorem 2.3.3.

Theorem 2.4.3. *Suppose that the set $\mathcal{A}$ is compact and $R(a, \theta_0)$ satisfies the growth condition, with $\Psi^1(d)$ such that $\Psi(d)/d^\delta = \kappa$, for a $\delta > 0$. Moreover, Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5 are satisfied with $C > C_2 + C_3 + 2$. Then for any $\tau > 0$, the P_0^n- probability of the following event*

$$\left\{ \tilde{X}_n : H\left(\mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n), \arg\min_{z \in \mathcal{A}} R(z, \theta_0) \right) \leq \left[\frac{2\tau \left[M(\epsilon_n^2 + \eta_n(0)) \right]^{\frac{1}{2}}}{\kappa} \right]^{\frac{1}{\delta}} \right\}$$

is at least $1 - \tau^{-1}$, where M is the positive constant as defined in Theorem 2.4.1.

2.4.2 Loss Calibrated VB

Algorithm 2 summarizes the *Loss-calibrated VB* (LCVB) method [10]. Observe that

Algorithm 2: Loss-calibrated VB

Input : $R(\cdot, \cdot)$, $\bar{X}_n, \mathcal{Q}$

Output: $\mathtt{a}^*_{\mathtt{LC}}$

Step 1. Compute:

$\mathtt{a}^*_{\mathtt{LC}} := \arg\min_{a \in \mathcal{A}} \max_{Q \in \mathcal{Q}} \left\{ -\mathrm{KL}(Q(\cdot) \| \Pi(\cdot | \bar{X}_n)) + \mathbb{E}_Q[R(a, \theta)] \right\}$.

this method combines the posterior approximation and decision-making problems into one minimax optimization problem. The objective here can be directly contrasted with that in Algorithm 1. Note that the inner maximization will result in an approximate (loss calibrated) posterior distribution at each decision point $a \in \mathcal{A}$.

In this section, we compute a bound on the loss-calibrated optimal decision $\mathtt{a}^*_{\mathtt{LC}}$ and optimal value.

Theorem 2.4.4. *Fix $a_0 \in \mathcal{A}$ and let ϵ_n be a sequence such that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ as $n \to \infty$ and*

$$L_n(\theta, \theta_0) = n \left(\sup_{a \in \mathcal{A}} |R(a, \theta) - R(a, \theta_0)| \right)^2.$$

Then under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5, for some positive constants $C, C_2, C_3, C_4(1)$, *and* $C_5(1)$ *such that* $\min(C, (C_4(1)+C_5(1))) > C_2+C_3+C_4(1)+2$, *and for*

$$\eta_n^R(1) := \frac{1}{n} \inf_{Q\in\mathcal{Q}} \mathbb{E}_{P_0^n}\left[\mathrm{KL}(Q(\theta)\|\Pi(\theta|\tilde{X}_n)) - \inf_{a\in\mathcal{A}} \mathbb{E}_Q[R(a,\theta)]\right],$$

the Loss calibrated VB approximation of the true posterior satisfies,

$$\mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0) dQ_{a_0}^*(\theta|\tilde{X}_n)\right] \leq n(M(1)\epsilon_n^2 + M\eta_n^R(1)), \tag{2.16}$$

where $M(1) = 2\left(C_1 + MC_4(1)\right)$, *and* $M = \frac{2C_1}{\min(C,\lambda,1)}$.

Note that, the second term (inside the expectation) in the definition of $\eta_n^R(1)$ could result in either $\eta_n(0) > \eta_n^R(1)$ or vice versa and therefore could play an important role in comparing the LCVB and naive VB approximations to the true optimal decision.

The next result establishes a bound on the optimality gap of the LCVB estimated optimal value $R(\mathtt{a}_{\mathtt{LC}}^*, \theta_0)$ from the true optimal value $V_0 = \inf_{z\in\mathcal{A}} R(z,\theta_0)$.

Theorem 2.4.5. *Suppose that the set* $\mathcal{A}$ *is compact and Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5 are satisfied with* $\min(C, (C_4(1)+C_5(1))) > C_2+C_3+C_4(1)+2$ *for some positive constants* $C, C_2, C_3, C_4(1)$, *and* $C_5(1)$. *Then, for any* $\tau > 0$, *the* P_0^n- *probability of the following event*

$$\left\{\tilde{X}_n : R(\mathtt{a}_{\mathtt{LC}}^*, \theta_0) - \inf_{z\in\mathcal{A}} R(z,\theta_0) \leq 2\tau\left[(M(1)\epsilon_n^2 + M\eta_n^R(1))\right]^{\frac{1}{2}}\right\} \tag{2.17}$$

is at least $1-\tau^{-1}$, *where* $M(1) = 2\left(C_1 + MC_4(1)\right)$, *and* $M = \frac{2C_1}{\min(C,\lambda,1)}$.

Next, we bound the optimality gap between the approximate LC optimal decision rule $\mathtt{a}_{\mathtt{LC}}^*$ and the true optimal decision.

Theorem 2.4.6. *Suppose that the set* $\mathcal{A}$ *is compact and* $R(a,\theta)$ *has a growth function* $\Psi(d)$ *such that* $\Psi(d)/d^\delta = \kappa$ *for a* $\delta > 0$. *Moreover, Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5 are satisfied with* $\min(C, (C_4(1)+C_5(1))) > C_2+C_3+C_4(1)+2$ *for some positive*

constants $C, C_2, C_3, C_4(1)$, *and* $C_5(1)$. *Then, for any* $\tau > 0$, *the* P_0^n- *probability of the following event*

$$\left\{ H(\mathsf{a}^*_{\mathtt{LC}}, \arg\min_{z\in\mathcal{A}} R(z,\theta_0)) \leq \left[\frac{2\tau \left[(M(1)\epsilon_n^2 + M\eta_n^R(1)) \right]^{\frac{1}{2}}}{\kappa} \right]^{\frac{1}{\delta}} \right\}$$

is at least $1-\tau^{-1}$, *where* $M(1) = 2\left(C_1 + MC_4(1)\right)$, *and* $M = \frac{2C_1}{\min(C,\lambda,1)}$.

2.5 Applications

We illustrate our theoretical findings with the help of three examples: the single and multi-product *newsvendor model* and Gaussian process classification. In the examples, we study the interplay between sample size n and the risk parameter γ, and their effect on the optimality gap in decisions and values.

2.5.1 Single-product Newsvendor Model

In this section, we study a canonical data-driven decision-making problem with a 'well-behaved' risk function $R(a,\theta)$, the data-driven newsvendor model. This problem has received extensive study in the literature, and remains a cornerstone of inventory management [7]–[9]. Recall that the newsvendor loss function is defined as

$$\ell(a,\xi) := h(a-\xi)^+ + b(\xi-a)^+$$

where h (underage cost) and b (overage cost) are given positive constants, $\xi \in [0,\infty)$ the random demand, and a the inventory or decision variable, typically assumed to take values in a compact decision space $\mathcal{A}$ with $\underline{a} := \min\{a : a \in \mathcal{A}\}$ and $\bar{a} := \max\{a : a \in \mathcal{A}\}$, and $\underline{a} > 0$. The distribution over the random demand, P_θ is assumed to be exponential with unknown rate parameter $\theta \in (0,\infty)$. The model risk can easily be derived as

$$R(a,\theta) := \mathbb{E}_{P_\theta}[\ell(a,\xi)] = ha - \frac{h}{\theta} + (b+h)\frac{\mathrm{e}^{-a\theta}}{\theta}, \tag{2.18}$$

which is convex in a. We assume that $\tilde{X}_n := \{\xi_1, \xi_2 \dots \xi_n\}$ be n observations of the random demand, assumed to be i.i.d random samples drawn from P_0.

We fix the model space $\Theta = [T, \infty)$ for some $T > 0$ and assume that θ_0 lies in the interior of Θ. We now assume a non-conjugate truncated inverse-gamma $(\text{Inv} - \Gamma)$ prior distribution restricted to Θ, with shape and rate parameter α and β respectively, that is for a set $A \subseteq \Theta$, we define $\Pi(A) = \text{Inv} - \Gamma_\Theta(A; \alpha, \beta) = \text{Inv} - \Gamma(A \cap \Theta; \alpha, \beta)/\text{Inv} - \Gamma(\Theta; \alpha, \beta)$. We now verify Assumptions 2.2.2, 2.2.1, 2.2.3, 2.2.5 and 2.2.4 (in that order) in this newsvendor setting. The proofs of the lemmas are delayed to the electronic companion for readability.

First, we fix the sieve set $\Theta_n(\epsilon) = \Theta$, which clearly implies that the restricted inverse-gamma prior $\Pi(\theta)$, places no mass on the complement of this set and therefore satisfies Assumption 2.2.2.

Second, under the condition that the true demand distribution is exponential with parameter θ_0 (and $P_0 \equiv P_{\theta_0}$), we demonstrate the existence of test functions satisfying Assumption 2.2.1.

Lemma 2.5.1. *Fix $n \geq 5$. Then, for any $\epsilon > \epsilon_n := \frac{1}{\sqrt{n}}$ with $\epsilon_n \to 0$, and $n\epsilon_n^2 \geq 1$, there exists a test function ϕ_n (depending on ϵ) such that $L_n^{NV}(\theta, \theta_0) = n\left(\sup_{a \in \mathcal{A}} |R(a, \theta) - R(a, \theta_0)|\right)^2$ satisfies*

$$\mathbb{E}_{P_0^n}[\phi_n] \leq C_0 \exp(-Cn\epsilon^2), \tag{2.19}$$

$$\sup_{\{\theta \in \Theta : L_n^{NV}(\theta, \theta_0) \geq C_1 n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1 - \phi_n] \leq \exp(-Cn\epsilon^2), \tag{2.20}$$

where $C_0 = 20$ and $C = \frac{C_1}{2(K_1^{NV})^2}$ for a constant $C_1 > 0$ and $K_1^{NV} = \frac{\left[\left(\frac{h}{\theta_0} - \frac{h}{T}\right)^2 + (b+h)^2\left(\frac{e^{-\underline{a}T}}{T} - \frac{e^{-\underline{a}\theta_0}}{\theta_0}\right)^2\right]^{1/2}}{d_H(T, \theta_0)}$.

The proof of the above result follows by showing that $d_L^{NV} = n^{-1/2}\sqrt{L_n^{NV}(\theta, \theta_0)}$ can be bounded above by the Hellinger distance between two exponential distributions on Θ (under which a test function exists) in Lemma 2.7.8 in the appendix.

Third, we show that there exist appropriate constants such that the inverse-gamma prior satisfies Assumption 2.2.3 when the demand distribution is exponential.

Lemma 2.5.2. *Fix $n_2 \geq 2$ and any $\lambda > 1$. Let $A_n := \{\theta \in \Theta : D_{1+\lambda}(P_0^n \| P_\theta^n) \leq C_3 n\epsilon_n^2\}$, where $D_{1+\lambda}(P_0^n \| P_\theta^n)$ is the Rényi divergence between P_0^n and P_θ^n. Then for $\epsilon_n^2 = \frac{\log n}{n}$ and any $C_3 > 0$ such $C_2 = \alpha C_3 \geq 2$, the truncated inverse-gamma prior $Inv - \Gamma_\Theta(A; \alpha, \beta)$ satisfies*

$$\Pi(A_n) \geq \exp(-nC_2\epsilon_n^2), \forall n \geq n_2.$$

Fourth, it is straightforward to see that the newsvendor model risk $R(a, \theta)$ is bounded below for a given $a \in \mathcal{A}$.

Lemma 2.5.3. *For any $a \in \mathcal{A}$ and positive constants h and b, the newsvendor model risk*

$$R(a, \theta) = \left(ha - \frac{h}{\theta} + (b + h)\frac{\mathrm{e}^{-a\theta}}{\theta} \right) \geq \left(\frac{h\underline{a}^2\theta^*}{(1 + a\theta^*)} \right),$$

where $\underline{a} := \min\{a \in \mathcal{A}\}$ and θ^* satisfies $h - (b + h)\mathrm{e}^{-a\theta^*}(1 + a\theta^*) = 0$.

This implies that $R(a, \theta)$ satisfies Assumption 2.2.5. Finally, we also show that the newsvendor model risk satisfies Assumption 2.2.4.

Lemma 2.5.4. *Fix $n \geq 1$ and $\gamma > 0$. For any $\epsilon > \epsilon_n$ and any $a \in \mathcal{A}$, $R(a, \theta)$ satisfies*

$$\mathbb{E}_\Pi[\mathbb{1}_{\{R(a,\theta)\gamma > C_4(\gamma) n\epsilon^2\}} \mathrm{e}^{\gamma R(a,\theta)}] \leq \exp(-C_5(\gamma) n\epsilon^2),$$

for any $C_4(\gamma) > 2\gamma\left(h\overline{a} + \frac{b}{T}\right)$ and $C_5(\gamma) = C_4(\gamma) - 2\gamma\left(h\overline{a} + \frac{b}{T}\right)$, where $\overline{a} := \max\{a \in \mathcal{A}\}$.

Note that Lemma 2.5.1 implies that $C = \frac{C_1}{2(K_1^{NV})^2}$ for any constant $C_1 > 0$. Fixing $\alpha = 1$ and using Lemma 2.5.2 we can choose $C_2 = C_3 = 2$. Now, C_1 can be chosen large enough such that $C > C_4(\gamma) + C_5(\gamma)$ for a given risk sensitivity $\gamma > 0$. Therefore, the condition on constants in Theorem 2.3.1 reduces to $C_5(\gamma) > 2 + C_2 + C_3 = 5$, and it can be satisfied easily by fixing $C_5(\gamma) = 5.1$(say).

These lemmas show that when the demand distribution is exponential and with a non-conjugate truncated inverse-gamma prior, our results in Theorem 2.3.2 and 2.3.3 can be used for RSVB method to bound the optimality gap in decisions and values for various values of the risk-sensitivity parameter γ. Recall that the bound obtained in Theorem 2.3.3 depends on ϵ_n^2 and $\eta_n^R(\gamma)$.

Lemma 2.5.2 implies that $\epsilon_n^2 = \frac{\log n}{n}$, but in order to get the complete bound we further need to characterize $\eta_n^R(\gamma)$. Recall that, as a consequence of Assumption 2.3.1 in Proposition 2.3.1, for a given $C_8 = -\inf_{Q\in\mathcal{Q}} \inf_{a\in\mathcal{A}} \mathbb{E}_Q[R(a,\theta)]$ that $C_9 > 0$ and $\eta_n^R(\gamma) \leq \gamma n^{-1} C_8 + C_9\epsilon_n^2$.

Therefore, in our next result, we show that in the newsvendor setting, we can construct a sequence $\{Q_n(\theta)\} \subset \mathcal{Q}$ that satisfies Assumption 2.3.1, and thus identify ϵ_n and the constant C_9. We fix $\mathcal{Q}$ to be the family of shifted gamma distributions with support $[T, \infty)$.

Lemma 2.5.5. *Let $\{Q_n(\theta)\}$ be a sequence of shifted gamma distributions with shape parameter $a = n$ and rate parameter $b = \frac{n}{\theta_0}$, then for truncated inverse gamma prior and exponentially distributed likelihood model*

$$\frac{1}{n}\left[\mathrm{KL}\left(Q_n(\theta)\|\Pi(\theta)\right) + \mathbb{E}_{Q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n(\tilde{X}_n))\|dP_\theta^n(\tilde{X}_n)\right)\right]\right] \leq C_9\epsilon_n^2,$$

where $\epsilon_n^2 = \frac{\log n}{n}$ and $C_9 = \frac{1}{2} + \max\left(0, 2 + \frac{2\beta}{\theta_0} - \log\sqrt{2\pi} - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \alpha\log\theta_0\right)$ and prior parameters are chosen such that $C_9 > 0$.

As a specific instance, consider the naive VB case. Since $\gamma \to 0^+$, the term $\eta_n(0)$ in Theorem 2.4.3 is bounded above by $C_9\epsilon_n^2$, where C_9 and ϵ_n^2 are derived in the result above. For the LCVB case, observe that Lemma 2.5.3 implies that $R(\cdot,\cdot)$ is bounded below and therefore $C_8 \leq -\left(\frac{h\underline{a}^2\theta^*}{(1+\bar{a}\theta^*)}\right)$, where $h, \underline{a}, \bar{a}$, and θ^* are given to the modeler or are easily computable. Now since $C_8 < 0$, it is straight forward to observe that $\eta_n^R(\gamma)$ term in Theorem 2.4.6 is bounded above by $C_9\epsilon_n^2$.

Now, using the result established in Lemmas above, we bound the optimality gap in values for the single product newsvendor model risk.

Theorem 2.5.1. *Fix $\gamma > 0$. Suppose that the set $\mathcal{A}$ is compact. Then, for the newsvendor model with exponentially distributed demand with rate $\theta \in \Theta = [T, \infty)$, prior distribution $\Pi(\cdot) = Inv-\Gamma_\Theta(\cdot;\alpha,\beta) = Inv-\Gamma(A\cap\Theta;\alpha,\beta)/Inv-\Gamma(\Theta;\alpha,\beta)$, and the variational family*

fixed to shifted (by $T > 0$) gamma distributions, and for any $\tau > 0$, the P_0^n- probability of the following event

$$\left\{ \tilde{X}_n : R(\mathtt{a}^*_{\mathtt{RS}}, \theta_0) - \inf_{z \in \mathcal{A}} R(z, \theta_0) \leq 2\tau M(\gamma) \left(\frac{\log n}{n} \right)^{1/2} \right\} \tag{2.21}$$

is at least $1 - \tau^{-1}$ for sufficiently large n and for some mapping $M : \mathbb{R}^+ \to \mathbb{R}^+$, where $R(\cdot, \theta)$ is the newsvendor model risk.

Proof. The proof is a direct consequence of Theorem 2.3.2, Lemmas 2.5.1, 2.5.2, 2.5.3, 2.5.4, 2.5.5, and Proposition 2.3.2. □

Next, we bound the optimality gap between the approximate optimal decision rule $\mathtt{a}^*_{\mathtt{RS}}$ and the true optimal decision. The bound, in particular, depends on the curvature of $R(a, \theta_0)$ around the true optimal decision, defined using the growth condition in Assumption 2.2.6.

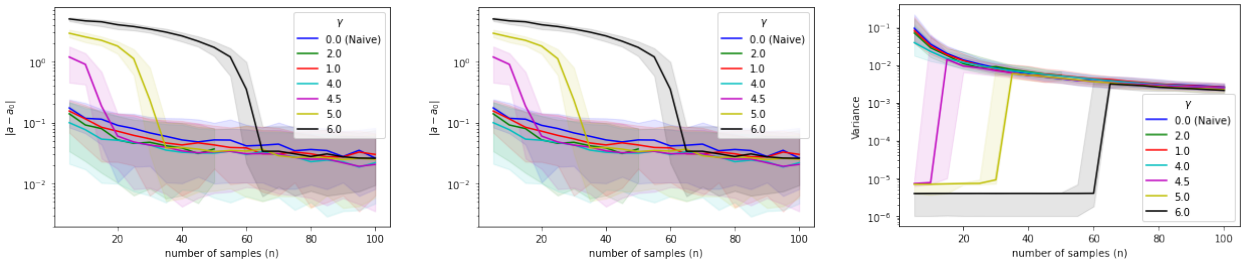
Theorem 2.5.2. *Fix $\gamma > 0$. Suppose that the set $\mathcal{A}$ is compact and $R(a, \theta_0)$ satisfies the growth condition in Assumption 2.2.6, with $\Psi(d)$ such that $\Psi(d)/d^\delta = \kappa$, for any $\delta > 0$. Then, for the newsvendor model with exponentially distributed demand with rate $\theta \in \Theta = [T, \infty)$, prior distribution $\Pi(\cdot) = Inv - \Gamma_\Theta(\cdot; \alpha, \beta) = Inv - \Gamma(A \cap \Theta; \alpha, \beta)/Inv - \Gamma(\Theta; \alpha, \beta)$, and the variational family fixed to shifted (by $T > 0$) gamma distributions, and for any $\tau > 0$, the P_0^n- probability of the following event*

$$\left\{ \tilde{X}_n : H\left(\mathtt{a}^*_{\mathtt{RS}}(\tilde{X}_n), \arg\min_{z \in \mathcal{A}} R(z, \theta_0) \right) \leq \left[\frac{2\tau}{\kappa} M(\gamma) \left(\frac{\log n}{n} \right)^{1/2} \right]^{\frac{1}{\delta}} \right\}$$

is at least $1 - \tau^{-1}$ for sufficiently large n and for some mapping $M : \mathbb{R}^+ \to \mathbb{R}^+$, where $R(\cdot, \theta)$ is the newsvendor model risk.

Proof. The proof is a direct consequence of Theorem 2.3.3, Lemmas 2.5.1, 2.5.2, 2.5.3, 2.5.4, 2.5.5, and Proposition 2.3.2. □

Next, we demonstrate the effect of varying the risk-sensitivity parameter γ. We fix $\theta_0 = 0.1$, $b = 1$, $h = 5$, $\alpha = 1$, and $\beta = 4.1$. We run RSVB algorithm with $\gamma \in \{0(\text{ naive }), 1, 2, 4.5, 5, 6\}$ and repeat the experiment over 100 sample paths. We plot the



(a) OG in Values (b) OG in Decisions (c) Variance

Figure 2.1. Optimality gap in values and decisions, and the variance of the RSVB posterior (mean over 100 sample paths) against the number of samples (n) for various values of γ.

results in Figure 2.1. In Figure 2.1(a) and (b), we plot the optimality gap in values and decisions, that is $R(\mathtt{a}^*_{\mathtt{RS}}(\gamma), \theta_0) - R(a^*_0, \theta_0)$ and $|\mathtt{a}^*_{\mathtt{RS}}(\gamma) - a^*_0|$ respectively, for various values of γ. We observe that the gap decreases when n increases. This observation supports our results in Propositions 2.3.1 and 2.3.2 that establishes the properties of $\eta^R_n(\gamma)$ as n increases. Lastly, in Figure 2.1(c), we plot the variance of the RSVB posterior as n increases for various values of γ; as anticipated the variance reduces as n increases. To observe the effect of γ, first recall that as γ increases the decision maker become more risk averse and so is our algorithmic framework RSVB. Indeed, from the rightmost variance plot in Figure 2.1 it is evident that for larger value of γ (> 4) the RSVB posterior is more concentrated on the subset of Θ, where risk is more and consequently we observe large optimality gaps in values and decision (see first two plots in Figure 2.1). Moreover, as n increase the effect of larger γ reduces, since as n increases the incentive to deviate from the posterior reduces (due to increased KL divergence dominance for larger n in RSVB).

2.5.2 Multi-product newsvendor problem

Analogous to the one-dimensional newsvendor loss function, the loss function in its multi-product version is defined as

$$\ell(a, \xi) := h^T(a - \xi)^+ + b^T(\xi - a)^+$$

where h and b are given vectors of underage and overage costs respectively for each product and mapping $(\cdot)^+$ is defined component-wise. We assume that there are d items or products and $\xi \in \mathbb{R}^d$ denotes the random vector of demands. Let $a \in \mathcal{A} \subset \mathbb{R}^d_+$ be the inventory or decision variable, typically assumed to take values in a compact decision space $\mathcal{A}$ with $\underline{a} := \{\{\min\{a_{\mathrm{i}} : a_{\mathrm{i}} \in \mathcal{A}_{\mathrm{i}}\}\}_{\mathrm{i}=1}^d$ and $\bar{a} := \{\{\max\{a_{\mathrm{i}} : a_{\mathrm{i}} \in \mathcal{A}_{\mathrm{i}}\}\}_{\mathrm{i}=1}^d$, and $\underline{a} > 0$, where $\mathcal{A}_{\mathrm{i}}$ is the marginal set of i^{th} component of $\mathcal{A}$. The random demand is assumed to be multivariate Gaussian, with unknown mean parameter $\theta \in \mathbb{R}^d$ but with known covariance matrix Σ. We also assume that Σ is a symmetric positive definite matrix and can be decomposed as $Q^T \Lambda Q$, where Q is an orthogonal matrix and Λ is a diagonal matrix consisting of respective eigenvalues of Σ. We also define $\overline{\Lambda} = \max_{\mathrm{i}\in\{1,2,\ldots d\}} \Lambda_{\mathrm{ii}}$ and $\underline{\Lambda} = \min_{\mathrm{i}\in\{1,2,\ldots d\}} \Lambda_{\mathrm{ii}}$. The model risk

$$
\begin{aligned}
R(a,\theta) = \mathbb{E}_{P_\theta}[\ell(a,\xi)] &= \sum_{\mathrm{i}=1}^{d} \mathbb{E}_{P_{\theta_{\mathrm{i}}}}[h_{\mathrm{i}}(a_{\mathrm{i}} - \xi_{\mathrm{i}})^+ + b_{\mathrm{i}}(\xi_{\mathrm{i}} - a_{\mathrm{i}})^+] \\
&= \sum_{\mathrm{i}=1}^{d} \left[(h_{\mathrm{i}} + b_{\mathrm{i}}) a_{\mathrm{i}} \Phi\left(\frac{(a_{\mathrm{i}} - \theta_{\mathrm{i}})}{\sigma_{\mathrm{ii}}}\right) - b_{\mathrm{i}} a_{\mathrm{i}} + \theta_{\mathrm{i}}(b_{\mathrm{i}} - h_{\mathrm{i}}) + \sigma_{\mathrm{ii}} \left[h \frac{\phi\left(\frac{(a_{\mathrm{i}}-\theta_{\mathrm{i}})}{\sigma_{\mathrm{ii}}}\right)}{\Phi\left(\frac{(a_{\mathrm{i}}-\theta_{\mathrm{i}})}{\sigma_{\mathrm{ii}}}\right)} + b \frac{\phi\left(\frac{(a_{\mathrm{i}}-\theta_{\mathrm{i}})}{\sigma_{\mathrm{ii}}}\right)}{1 - \Phi\left(\frac{(a_{\mathrm{i}}-\theta_{\mathrm{i}})}{\sigma_{\mathrm{ii}}}\right)} \right] \right],
\end{aligned}
$$

which is convex in a. Here $P_{\theta_{\mathrm{i}}}$ is the marginal distribution of ξ for i^{th} product, $\phi(\cdot)$ and $\Phi(\cdot)$ are probability and cumulative distribution function of the standard Normal distribution. We also assume that the true mean parameter θ_0 lies in a compact subspace $\Theta \subset \mathbb{R}^d$. We fix the prior to be uniformly distributed on Θ with no correlation across its components, that is $\boldsymbol{\pi}(A) = \frac{}{m(\Theta)} = \prod_{\mathrm{i}=1}^d \frac{}{m(\Theta_{\mathrm{i}})}$, where $m(B)$ is the Lebesgue measure (or volume) of $B \subset \mathbb{R}^d$ As in the previous example, we fix the sieve set $\Theta_n(\epsilon) = \Theta$, which clearly implies that $\Pi(\theta)$ places no mass on the complement of this set and therefore satisfies Assumption 2.2.2.

Then under the condition that the true demand distribution has a multivariate Gaussian distribution (with known Σ) and mean θ_0 ($P_0 \equiv P_{\theta_0}$), we demonstrate the existence of test functions satisfying Assumption 2.2.1 by constructing a test function unlike the single-product newsvendor problem with exponential demand..

Lemma 2.5.6. *Fix $n \geq 1$. Then, for any $\epsilon > \epsilon_n := \frac{1}{\sqrt{n}}$ with $\epsilon_n \to 0$, and $n\epsilon_n^2 \geq 1$ and test function $\phi_{n,\epsilon} := \mathbb{1}_{\left\{\tilde{X}_n : \left\|\hat{\theta}_n - \theta_0\right\| > \sqrt{\tilde{C}\epsilon^2}\right\}}$, $L_n^{MNV}(\theta,\theta_0) = n\left(\sup_{a\in\mathcal{A}} |R(a,\theta) - R(a,\theta_0)|\right)^2$ satisfies*

$$\mathbb{E}_{P_0^n}[\phi_n] \leq C_0 \exp(-Cn\epsilon^2), \tag{2.22}$$

$$\sup_{\{\theta\in\Theta: L_n^{MNV}(\theta,\theta_0)\geq C_1 n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_n] \leq \exp(-Cn\epsilon^2), \tag{2.23}$$

with $C_0 = 1$, $C_1 = 4K^2C$ *and* $C = 1/8\left(\frac{\tilde{C}}{d\overline{\Lambda}} - 1\right)$ *for sufficiently large* $\tilde{C}$ *such that* $C > 1$ *and* $\overline{\Lambda} = \max_{i\in\{1,2,\ldots d\}} \Lambda_{ii}$, *where* $K = \sup_{\mathcal{A},\Theta} \|\partial_\theta R(a,\theta)\|$.

In the following result, we show that there exist appropriate constants such that prior distribution satisfies Assumption 2.2.3 when the demand distribution is a multivariate Gaussian with unknown mean.

Lemma 2.5.7. *Fix* $n_2 \geq 2$ *and any* $\lambda > 1$. *Let* $A_n := \{\theta \in \Theta : D_{1+\lambda}(P_0^n \| P_\theta^n) \leq C_3 n\epsilon_n^2\}$, *where* $D_{1+\lambda}(P_0^n \| P_\theta^n)$ *is the Rényi Divergence between* P_0^n *and* P_θ^n. *Then for* $\epsilon_n^2 = \frac{\log n}{n}$ *and any* $C_3 > 0$ *such that* $C_2 = \frac{4d}{\overline{\Lambda}(\lambda+1)\left(\prod_{i=1}^d m(\Theta_i)\right)^{2/d}} C_3 \geq 2$ *and for large enough* n, *the uncorrelated uniform prior restricted to* Θ *satisfies*

$$\Pi(A_n) \geq \exp(-nC_2\epsilon_n^2).$$

Next, it is straightforward to see that the multi-product newsvendor model risk $R(a,\theta)$ is bounded below for a given $a \in \mathcal{A}$ on a compact set Θ and thus it satisfies Assumption 2.2.5. Finally, we also show that the newsvendor model risk satisfies Assumption 2.2.4.

Lemma 2.5.8. *Fix* $n \geq 1$ *and* $\gamma > 0$. *For any* $\epsilon > \epsilon_n$ *and* $a \in \mathcal{A}$, $R(a,\theta)$ *satisfies*

$$\mathbb{E}_\Pi[\mathbb{1}_{\{G(a,\theta)\gamma > C_4(\gamma) n\epsilon^2\}} e^{\gamma G(a,\theta)}] \leq \exp(-C_5(\gamma) n\epsilon_n^2),$$

for any $C_4(\gamma) > 2\gamma \sup_{\{a,\theta\}\in\mathcal{A}\otimes\Theta} G(a,\theta)$ *and* $C_5(\gamma) = C_4(\gamma) - 2\gamma \sup_{\{a,\theta\}\in\mathcal{A}\otimes\Theta} G(a,\theta)$.

Similar to single product example, in our next result, we show that in the multi-product newsvendor setting, we can construct a sequence $\{Q_n(\theta)\} \in \mathcal{Q}$ that satisfies Assumption 2.3.1, and thus identify ϵ_n and constant C_9. We fix $\mathcal{Q}$ to be the family of uncorrelated Gaussian distributions restricted to Θ.

Lemma 2.5.9. *Let $\{Q_n(\theta)\}$ be a sequence of product of d univariate Gaussian distribution defined as $q_n^{\mathrm{i}}(\theta) \propto \frac{1}{\sqrt{2\pi\sigma_{\mathrm{i},n}^2}} \mathrm{e}^{-\frac{1}{2\sigma_{\mathrm{i},n}^2}(\theta-\mu_{\mathrm{i},n})^2} \mathbb{1}_{\Theta_{\mathrm{i}}} = \frac{\mathcal{N}(\theta_{\mathrm{i}}|\mu_{\mathrm{i},n},\sigma_{\mathrm{i},n})\mathbb{1}_{\Theta_{\mathrm{i}}}}{\mathcal{N}(\Theta_{\mathrm{i}}|\mu_{\mathrm{i},n},\sigma_{\mathrm{i},n})}$ and fix $\sigma_{\mathrm{i},n} = 1/\sqrt{n}$ and $\theta_{\mathrm{i}} = \theta_0^{\mathrm{i}}$ for all $\mathrm{i} \in \{1, 2, \ldots, d\}$. Then for uncorrelated uniform distribution restricted to Θ and multivariate normal likelihood model*

$$\frac{1}{n}\left[\mathrm{KL}\left(Q_n(\theta)\|\Pi(\theta)\right) + \mathbb{E}_{Q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n(\tilde{X}_n))\|dP_\theta^n(\tilde{X}_n)\right)\right]\right] \leq C_9\epsilon_n^2,$$

where $\epsilon_n^2 = \frac{\log n}{n}$ and $C_9 := \frac{d}{2} + \max\left(0, -\sum_{\mathrm{i}=1}^d \left[\log(\sqrt{2\pi\mathrm{e}}) - \log(m(\Theta_{\mathrm{i}}))\right] + \frac{d}{2}\underline{\Lambda}^{-1}\right)$.

Now, using the result established in Lemmas above, we bound the optimality gap in values for the multi-product newsvendor model risk.

Theorem 2.5.3. *Fix $\gamma > 0$. Suppose that the set $\mathcal{A}$ is compact. Then, for the multi-product newsvendor model with multivariate Gaussian distributed demand with known covariance matrix Σ and unknown mean vector θ lying in a compact subset $\Theta \subset \mathbb{R}^d$, prior $\Pi(\cdot) = \prod_{\mathrm{i}=1}^d \frac{m(\{\cdot\}\cap\Theta_{\mathrm{i}})}{m(\Theta_{\mathrm{i}})}$, and the variational family fixed to uncorrelated Gaussian distribution restricted to Θ, and for any $\tau > 0$, the P_0^n- probability of the following event*

$$\left\{\tilde{X}_n : R(\mathtt{a}_{\mathtt{RS}}^*, \theta_0) - \inf_{z\in\mathcal{A}} R(z, \theta_0) \leq 2\tau M(\gamma)\left(\frac{\log n}{n}\right)^{1/2}\right\} \tag{2.24}$$

is at least $1-\tau^{-1}$ for sufficiently large n and for some mapping $M : \mathbb{R}^+ \to \mathbb{R}^+$, where $R(\cdot, \theta)$ is the multi-product newsvendor model risk.

Proof. The proof is a direct consequence of Theorem 2.3.2, Lemmas 2.5.6, 2.5.7, 2.5.8, 2.5.9, and Proposition 2.3.2. □

Next, we bound the optimality gap between the approximate optimal decision rule $\mathtt{a}_{\mathtt{RS}}^*$ and the true optimal decision.

Theorem 2.5.4. *Fix $\gamma > 0$. Suppose that the set $\mathcal{A}$ is compact and $R(a, \theta_0)$ satisfies the growth condition in Assumption 2.2.6, with $\Psi(d)$ such that $\Psi(d)/d^\delta = \kappa$, for any $\delta > 0$. Then, for the multi-product newsvendor model with multivariate Gaussian distributed demand with known covariance matrix Σ and unknown mean vector θ lying in a compact subset*

$\Theta \subset \mathbb{R}^d$, *prior* $\Pi(\cdot) = \prod_{\mathrm{i}=1}^{d} \frac{m(\{\cdot\} \cap \Theta_{\mathrm{i}})}{m(\Theta_{\mathrm{i}})}$, *and the variational family fixed to uncorrelated Gaussian distribution restricted to* Θ, *and for any* $\tau > 0$, *the* P_0^n- *probability of the following event*

$$\left\{ \tilde{X}_n : H\left(\mathtt{a}^*_{\mathtt{RS}}(\tilde{X}_n), \arg\min_{z \in \mathcal{A}} R(z, \theta_0) \right) \leq \left[\frac{2\tau}{\kappa} M(\gamma) \left(\frac{\log n}{n} \right)^{1/2} \right]^{\frac{1}{\delta}} \right\}$$

is at least $1 - \tau^{-1}$ *for sufficiently large* n *and for a known function* $M(\gamma)$, *where* $R(\cdot, \theta)$ *is the multi-product newsvendor model risk.*

Proof. The proof is a direct consequence of Theorem 2.3.3, Lemmas 2.5.6, 2.5.7, 2.5.8, 2.5.9, and Proposition 2.3.2. □

2.5.3 Gaussian process classification

Consider a problem of classifying an input pattern or features Y lying in measure space $([0,1]^d, \mathcal{Y}, \nu)$ into one of two classes $\{-1, 1\}$, where $\xi \in \{-1, 1\}$ denote the class of Y. For a given Y, we model the classifier using a Bernoulli distribution $p(\xi|Y, \theta) = \Psi_\xi(\theta(Y))$, where $\theta : [0,1]^d \to \mathbb{R}$ is a non-parametric model parameter in a separable Banach space $(\Theta, \|\cdot\|)$ and measurable functions $\Psi_1(x) = (1 + \mathrm{e}^{-x})^{-1}$ and $\Psi_{-1}(x) = 1 - \Psi_1(x)$. Note that $\Psi_1(\cdot)$ is a logistic function. We denote $\psi(\cdot)$ as the derivative of $\Psi_1(\cdot)$. We assume that $\nu(\cdot)$ is independent of ξ. Thus the sequence of independent observations $\{\tilde{Y}_n, \tilde{X}_n\} = \{(Y_1, \xi_1), (Y_2, \xi_2), \ldots, (Y_n, \xi_n)\}$ are assumed to be generated from model

$$P_\theta(\xi, Y) = p(\xi|Y, \theta)\nu(Y).$$

In the above binary classification problem, the objective is to estimate $\theta(\cdot)$ using the observation vector $\{\tilde{Y}_n, \tilde{X}_n\}$. We posit a Gaussian process (GP) prior $\Pi(\cdot)$ on $\theta(\cdot) \in \Theta$ (to be defined later). We also assume that $\nu(\cdot)$ is known and we do not place any prior on it. Consequently, the posterior distribution over $\theta(\cdot)$ given observations $\{\tilde{Y}_n, \tilde{X}_n\}$ can be defined as

$$d\Pi(\theta|\{\tilde{Y}_n, \tilde{X}_n\}) = \frac{d\Pi(\theta) \prod_{\mathrm{i}=1}^n \Psi_{\xi_{\mathrm{i}}}(\theta(Y_{\mathrm{i}}))\nu(Y_{\mathrm{i}})}{\int \prod_{\mathrm{i}=1}^n \Psi_{\xi_{\mathrm{i}}}(\theta(Y_{\mathrm{i}}))\nu(Y_{\mathrm{i}}) d\Pi(\theta)} = \frac{d\Pi(\theta) \prod_{\mathrm{i}=1}^n \Psi_{\xi_{\mathrm{i}}}(\theta(Y_{\mathrm{i}}))}{\int \prod_{\mathrm{i}=1}^n \Psi_{\xi_{\mathrm{i}}}(\theta(Y_{\mathrm{i}})) d\Pi(\theta)}.$$

Consider the loss function $\ell(a, \xi)$ defined as

$$\ell(a,\xi) := \begin{cases} 0, & \text{if } a = \xi, \\ c_+, & \text{if } a = +1, \xi = -1, \\ c_-, & \text{if } a = -1, \xi = +1, \end{cases} \tag{2.25}$$

where c_+ and c_- are known positive constants. The model risk is given by

$$R(a,\theta) = \mathbb{E}_{P_\theta}[\ell(a,\xi)] = \begin{cases} c_+ \mathbb{E}_\nu[\Psi_{-1}(\theta(y))], & a = +1, \\ c_- \mathbb{E}_\nu[\Psi_1(\theta(y))], & a = -1. \end{cases} \tag{2.26}$$

We define the distance function as $L_n^{GP}(\theta, \theta_0) = n \left(\sup_{a \in \mathcal{A}} |R(a,\theta) - R(a,\theta_0)|\right)^2$. In anticipation of demonstrating that the binary classification model with GP prior and distance function L_n^{GP} satisfy the desired set of assumptions, we recall the following result, from [70], which will be central in establishing Assumptions 2.2.1, 2.2.2, and 2.2.3.

Lemma 2.5.10. *[Theorem 2.1 [70]] Let $\theta(\cdot)$ be a Borel measurable, zero-mean Gaussian random element in a separable Banach space $(\Theta, \|\cdot\|)$ with reproducing kernel Hilbert space (RKHS) $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and let θ_0 be contained in the closure of $\mathbb{H}$ in Θ. For any $\epsilon > \epsilon_n$ satisfying $\varphi_{\theta_0}(\epsilon) \leq n\epsilon^2$, where*

$$\varphi_{\theta_0}(\epsilon) = \inf_{h \in \mathbb{H}: \|h - \theta_0\| < \epsilon} \|h\|_{\mathbb{H}}^2 - \log \Pi(\|\theta\| < \epsilon) \tag{2.27}$$

and any $C_{10} > 1$ with $\mathrm{e}^{-C_{10} n \epsilon_n^2} < 1/2$, there exists a measurable set $\Theta_n(\epsilon) \subset \Theta$ such that

$$\log N(3\epsilon, \Theta_n(\epsilon), \|\cdot\|) \leq 6 C_{10} n \epsilon^2, \tag{2.28}$$

$$\Pi(\theta \notin \Theta_n(\epsilon)) \leq \mathrm{e}^{-C_{10} n \epsilon^2}, \tag{2.29}$$

$$\Pi(\|\theta - \theta_0\| < 4\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2}. \tag{2.30}$$

The proof of their result can be easily adapted from the proof of Vaart and Zanten [70, Theorem 2.1], which is specifically for $\epsilon = \epsilon_n$. Notice that the result above is true for any

norm $\|\cdot\|$ on the Banach space if that satisfies $\varphi_{\theta_0}(\epsilon) \leq n\epsilon^2$. Moreover, if $\varphi_{\theta_0}(\epsilon_n) \leq n\epsilon_n^2$ is true, then it also holds for any $\epsilon > \epsilon_n$, since by definition $\varphi_{\theta_0}(\epsilon)$ is a decreasing function of ϵ.

All the results in the previous lemma depend on $\varphi_{\theta_0}(\epsilon)$ being less than $n\epsilon^2$. In particular, observe that the second term in the definition of $\varphi_{\theta_0}(\epsilon)$ depends on the prior distribution on Θ. Therefore, Vaart and Zanten [70, Theorem 4.5] show that $\varphi_{\theta_0}(\epsilon_n) \leq n\epsilon_n^2$ (with $\|\cdot\|$ as supremum norm) is satisfied by the Gaussian prior of type

$$W(\cdot) = \sum_{\mathrm{j}=1}^{\bar{J}_\alpha} \sum_{k=1}^{2^{\mathrm{j}d}} \mu_{\mathrm{j}} Z_{\mathrm{j},k} \vartheta_{\mathrm{j},k}(\cdot), \tag{2.31}$$

where $\{\mu_{\mathrm{j}}\}$ is a sequence that decreases with j, $\{Z_{\mathrm{i,j}}\}$ are i.i.d. standard Gaussian random variables and $\{\vartheta_{\mathrm{j},k}\}$ form a double-indexed orthonormal basis (with respect to measure ν), that is $\mathbb{E}_\nu[\vartheta_{\mathrm{j},k}\vartheta_{l,m}] = \mathbb{1}_{\{\mathrm{j}=l,k=m\}}$). $\bar{J}_\alpha$ is the smallest integer satisfying $2^{\bar{J}_\alpha d} = n^{d/(2\alpha+d)}$ for a given $\alpha > 0$. In particular, the GP above is constructed using the function class that is supported on $[0,1]^d$ and has a wavelet expansion,

$$w(\cdot) = \sum_{\mathrm{j}=1}^{\infty} \sum_{k=1}^{2^{\mathrm{j}d}} w_{\mathrm{j},k} \vartheta_{\mathrm{j},k}(\cdot).$$

The wavelet function space is equipped with the L_2-norm: $\|w\|_2 = \sum_{\mathrm{j}=1}^{\infty} \left(\sum_{k=1}^{2^{\mathrm{j}d}} |w_{\mathrm{j},k}|^2\right)^{1/2}$; the supremum norm: $\|w\|_\infty = \sum_{\mathrm{j}=1}^{\infty} 2^{\mathrm{j}d} \max_{1\leq k\leq 2^{\mathrm{j}d}} |w_{\mathrm{j},k}|$; and the Besov $(\beta,\infty,\infty)-$norm: $\|w\|_{\beta;\infty,\infty} = \sup_{1\leq \mathrm{j}<\infty} 2^{\mathrm{j}\beta} 2^{\mathrm{j}d} \max_{1\leq k\leq 2^{\mathrm{j}d}} |w_{\mathrm{j},k}|$. Note that W induces a measure over the RKHS $\mathbb{H}$, defined as a collection of truncated wavelet functions

$$w(\cdot) = \sum_{\mathrm{j}=1}^{\bar{J}_\alpha} \sum_{k=1}^{2^{\mathrm{j}d}} w_{\mathrm{j},k} \vartheta_{\mathrm{j},k}(\cdot),$$

with norm induced by inner-product on $\mathbb{H}$ as $\|w\|_{\mathbb{H}}^2 = \sum_{\mathrm{j}=1}^{\bar{J}_\alpha} \sum_{k=1}^{2^{\mathrm{j}d}} \frac{w_{\mathrm{j},k}^2}{\mu_{\mathrm{j}}^2}$. The RKHS kernel $K : [0,1]^d \times [0,1]^d \mapsto \mathbb{R}$ can be easily derived as

$$K(x,y) = \mathbb{E}[W(x)W(y)] = \mathbb{E}\left[\left(\sum_{\mathrm{j}=1}^{\bar{J}_\alpha} \sum_{k=1}^{2^{\mathrm{j}d}} \mu_{\mathrm{j}} Z_{\mathrm{j},k} \vartheta_{\mathrm{j},k}(y)\right)\left(\sum_{\mathrm{j}=1}^{\bar{J}_\alpha} \sum_{k=1}^{2^{\mathrm{j}d}} \mu_{\mathrm{j}} Z_{\mathrm{j},k} \vartheta_{\mathrm{j},k}(x)\right)\right] = \sum_{\mathrm{j}=1}^{\bar{J}_\alpha} \sum_{k=1}^{2^{\mathrm{j}d}} \mu_{\mathrm{j}}^2 \vartheta_{\mathrm{j},k}(y)\vartheta_{\mathrm{j},k}(x).$$

Indeed, by the definition of this kernel and inner product, observe that

$$\langle K(x,\cdot), w(\cdot)\rangle = \sum_{\mathrm{j}=1}^{\bar{J}_\alpha}\sum_{k=1}^{2^{\mathrm{j}d}} w_{\mathrm{j},k}\mu_{\mathrm{j}}^2\vartheta_{\mathrm{j},k}(x)\frac{1}{\mu_{\mathrm{j}}^2} = w(x).$$

Moreover, $\langle K(x,\cdot), K(y,\cdot)\rangle = \sum_{\mathrm{j}=1}^{\bar{J}_\alpha}\sum_{k=1}^{2^{\mathrm{j}d}}\mu_{\mathrm{j}}^2\vartheta_{\mathrm{j},k}(x)\mu_{\mathrm{j}}^2\vartheta_{\mathrm{j},k}(y)\frac{1}{\mu_{\mathrm{j}}^2} = K(x,y)$. It is clear from its definition that W is a centered Gaussian random field on the RKHS.

Next, using the definition of the kernel, we derive the covariance operator of the Gaussian random field W. Recall that $Y \sim \nu$, which enables us to define the covariance operator, following [20, (6.19)] as

$$(\mathcal{C}h_\nu)(x) = \int_{[0,1]^d} K(x,y)h_\nu(y)d\nu(y).$$

Also, observe that $\{\mu_{\mathrm{j}}^2, \varphi_{\mathrm{j},k}\}$ is the eigenvalue and eigen function pair of the covariance operator $\mathcal{C}$. Consequently, using Karhunen Loéve expansion [20, Theorem 6.19] the prior induced by W on $\mathbb{H}$ is a Gaussian distribution denoted as $\mathcal{N}(0,\mathcal{C})$. We also recall the Cameron-Martin space denoted as $\mathrm{Im}(\mathcal{C}^{1/2})$ associated with a Gaussian measure $\mathcal{N}(0,\mathcal{C})$ on $\mathbb{H}$ to be the intersection of all linear spaces of full measure under $\mathcal{N}(0,\mathcal{C})$ [20, (page 530)]. In particular, $\mathrm{Im}(\mathcal{C}^{1/2})$ is the Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{C}} = \langle\mathcal{C}^{-1/2}\cdot, \mathcal{C}^{-1/2}\cdot\rangle$.

Next, we show the existence of test functions in the following result.

Lemma 2.5.11. *For any $\epsilon > \epsilon_n$ with $\epsilon_n \to 0$, $n\epsilon_n^2 \geq 2\log 2$, and $\varphi_{\theta_0}(\epsilon) \leq n\epsilon^2$, there exists a test function ϕ_n (depending on ϵ) such that $L_n^{GP}(\theta,\theta_0) = n\left(\sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|\right)^2$ satisfies*

$$\mathbb{E}_{P_0^n}[\phi_n] \leq C_0\exp(-Cn\epsilon^2), \tag{2.32}$$

$$\sup_{\{\theta\in\Theta: L_n^{GP}(\theta,\theta_0)\geq C_1 n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_n] \leq \exp(-Cn\epsilon^2), \tag{2.33}$$

where $C = 1/6$, $C_0 = 2$ and $C_1 = (\max(c_+, c_-))^2$.

Assumption 2.2.2 is a direct consequence of (2.29) in Lemma 2.5.10. Next, we prove that prior distribution and the likelihood model satisfy Assumption 2.2.3 using (2.30) of Lemma 2.5.10.

Lemma 2.5.12. *For any $\lambda > 1$, let $A_n := \{\theta \in \Theta : D_{1+\lambda}(P_0^n \| P_\theta^n) \leq C_3 n\epsilon_n^2\}$, where $D_{1+\lambda}(P_0^n \| P_\theta^n)$ is the Rényi Divergence between P_0^n and P_θ^n. Then for any $\epsilon > \epsilon_n$ satisfying $\varphi_{\theta_0}(\epsilon) \leq n\epsilon^2$ and $C_3 = 16(\lambda + 1)$ and $C_2 = 1$, the GP prior satisfies*

$$\Pi(A_n) \geq \exp(-nC_2\epsilon_n^2).$$

Assumption 2.2.4 and 2.2.5 are straightforward to satisfy since the model risk function $R(a, \theta)$ is bounded from above and below.

Now, suppose the variational family $\mathcal{Q}_{GP}$ is a class of Gaussian distributions on Θ, defined as $\mathcal{N}(m_q, \mathcal{C}_q)$, m_q belongs to Θ and $\mathcal{C}_q$ is the covariance operator defined as $\mathcal{C}_q = \mathcal{C}^{1/2}(I - S)\mathcal{C}^{1/2}$, for any S which is a symmetric and Hilbert-Schmidt (HS) operator on Θ (eigenvalues of HS operator are square summable). Note that S and m_q span the distributions in $\mathcal{Q}_{GP}$.

The following lemma verifies Assumption 2.3.1, for a specific sequence of distributions in $\mathcal{Q}$.

Lemma 2.5.13. *For a given $J \in \mathbb{N}$, let $\{Q_n\}$ be a sequence variational distribution such that Q_n is the measure induced by a GP, $W_Q(\cdot) = \theta_0^J(y) + \sum_{\mathrm{j}=1}^{J} \sum_{k=1}^{2^{\mathrm{j}d}} \zeta_{\mathrm{j}}^2 Z_{\mathrm{j},k} \vartheta_{\mathrm{j},k}(\cdot)$, where $\theta_0^J(\cdot) = \sum_{\mathrm{j}=1}^{J} \sum_{k=1}^{2^{\mathrm{j}d}} \theta_{0;\mathrm{j},k} \vartheta_{\mathrm{j},k}(\cdot)$ and $\zeta_{\mathrm{j}}^2 = \frac{\mu_{\mathrm{j}}^2}{1+n\epsilon_n^2\tau_{\mathrm{j}}^2}$. Then for GP prior induced by $W = \sum_{\mathrm{j}=1}^{J} \sum_{k=1}^{2^{\mathrm{j}d}} \mu_{\mathrm{j}} Z_{\mathrm{j},k} \vartheta_{\mathrm{j},k}$ and $\mu_{\mathrm{j}} = 2^{-\mathrm{j}d/2 - \mathrm{j}a}$ for some $a > 0$, $\|\theta_0\|_{\beta;\infty,\infty} < \infty$, and $\theta_0^J(y)$ lie in the Cameron-Martin space $Im(\mathcal{C}^{1/2})$, we have*

$$\frac{1}{n}\mathrm{KL}(\mathcal{N}(\bar{\theta}_0^J, \mathcal{C}_q) \| \mathcal{N}(0, \mathcal{C})) + \frac{1}{n}\mathbb{E}_{Q_n}\mathrm{KL}(P_0^n \| P_\theta^n) \leq C_9\epsilon_n^2,$$

where

$$\epsilon_n = \begin{cases} n^{-\beta/(2\alpha+d)} \log n & \mathrm{if} a \leq \beta \leq \alpha \\ n^{-\alpha/(2\alpha+d)} \log n & \mathrm{if} a \leq \alpha \leq \beta \\ n^{-a/(2a+d)} (\log n)^{d/(2a+d)} & \mathrm{if} \alpha \leq a \leq \beta \\ n^{-\beta/(2a+d)} (\log n)^{d/(2a+d)} & \mathrm{if} \alpha \leq \beta \leq a. \end{cases} \tag{2.34}$$

and $C_9 := \max\left(\|\theta_0\|^2_{\beta,\infty,\infty}, \frac{2^{-2a}-2^{-2Ja-2a}}{1-2^{-2a}}, 2^d/(2^d-1), C\right)$*, where* C *is a positive constant satisfying* $\|\theta_0(y) - \theta_0^J(y))\|^2_\infty \leq C2^{-2J\beta}$.

Using the result above together with Proposition 2.3.2 implies that the RSVB posterior converges at the same rate as the true posterior, where the convergence rate of the true posterior is derived in [70, Theorem 4.5] for the binary GP classification problem with truncated wavelet GP prior.

Finally, we use the Lemmas above to obtain bound on the optimality gap in values of the binary GP classification problem.

Theorem 2.5.5. *Fix* $\gamma > 0$ *and for a given* $J \in \mathbb{N}$*. For the binary GP classification problem with GP prior induced by* $W = \sum_{\mathrm{j}=1}^{J}\sum_{k=1}^{2^{\mathrm{j}d}} \mu_{\mathrm{j}} Z_{\mathrm{j},k}\vartheta_{\mathrm{j},k}$ *and* $\mu_{\mathrm{j}} = 2^{-\mathrm{j}d/2-\mathrm{j}a}$ *for some* $a > 0$*,* $\|\theta_0\|_{\beta;\infty,\infty} < \infty$*, and* $\theta_0^J(y)$ *lie in the Cameron-Martin space* $Im(\mathcal{C}^{1/2})$*, the variational family* $\mathcal{Q}_{GP}$*, and for any* $\tau > 0$*, the* P_0^n- *probability of the following event*

$$\left\{\tilde{X}_n : R(\mathsf{a}^*_{\mathsf{RS}}, \theta_0) - \inf_{z\in\mathcal{A}} R(z, \theta_0) \leq 2\tau M(\gamma)\epsilon_n\right\} \tag{2.35}$$

is at least $1-\tau^{-1}$ *for sufficiently large* n *and for some mapping* $M : \mathbb{R}^+ \to \mathbb{R}^+$*, where* $R(\cdot,\theta)$ *is defined in* (2.26) *and* ϵ_n *as derived in* (2.34)*.*

Proof. The proof is a direct consequence of Theorem 2.3.2, Lemmas 2.5.11, 2.5.12, 2.5.13, and Proposition 2.3.2. □

2.6 Conclusion

Data-driven decision-making has received significant research interest in the recent literature, in particular since the nature of the interplay between data and optimal decision-making can be quite different from the standard machine learning setting. While much of the literature focuses on empirical methods, Bayesian methods afford advantages particularly when making decisions in context of stochastic models. However, Bayesian methods are also hampered by integration requirements that can be hard to satisfy in practice.

In this paper we presented the risk-sensitive variational Bayesian computational framework for Bayes-predictive data-driven decision-making, and analyzed the statistical per-

formance of any computational algorithm derived from this framework by providing non-asymptotic bounds on the optimality gap. We also analyzed two specific algorithms, and for both the naive VB (NVB) and loss-calibrated VB (LCVB) algorithms we provide statistical analyses of the 'goodness' of the optimal decisions in terms of the true data generating model. We also compared the methods against the Bayes optimal solution on a newsvendor problem.

Our current methodology essentially relies on optimizing lower bounds to the 'true' problem at hand. One of our future objectives is to obtain sharp upper bounds on the true objective that can then provide a means of 'squeezing' the true optimal solution between these bounds. A second objective is to fully understand the interplay between robustness and our variational approximations. In some sense, robust methods aim to find the 'worst' distribution out of a set of distributions centered (in an appropriate sense) around a nominal distribution. On the other hand, VB methods find the closest distribution from a family that does not include the nominal distribution (if it did, then we could compute the posterior). There is almost a sense of duality between these perspectives that is worthy of further investigation. Third, from a methodological viewpoint, we are investigating the role of variational autoencoders ([96]) in the context of data-driven decision-making. Currently, our decision-making model requires us to fully specify the likelihood and prior models, while in practice it would be beneficial to make this fully data-driven – precisely where autoencoder technology would be useful. To the best of our knowledge very little is known about the statistical properties of these models, or their role in decision-making contexts.

2.7 Proofs

2.7.1 Alternative derivation of LCVB

We present the alternative derivation of LCVB. Consider the logarithm of the Bayes posterior risk,

$$\begin{aligned}\log \mathbb{E}_{\Pi(\theta|\tilde{X}_n)}[\exp(R(a,\theta))] &= \log \int_\Theta \exp(R(a,\theta)) d\Pi(\theta|\tilde{X}_n) \\ &= \log \int_\Theta \frac{dQ(\theta)}{dQ(\theta)} \exp(R(a,\theta)) d\Pi(\theta|\tilde{X}_n) \\ &\geq -\int_\Theta dQ(\theta) \log \frac{dQ(\theta)}{\exp(R(a,\theta)) d\Pi(\theta|\tilde{X}_n)} =: \mathcal{F}(a; Q(\cdot), \tilde{X}_n) \quad (2.36)\end{aligned}$$

where the inequality follows from an application of Jensen's inequality (since, without loss of generality, $\exp(R(a,\theta)) > 0$ for all $a \in \mathcal{A}$ and $\theta \in \Theta$), and $Q \in \mathcal{Q}$. Then, it follows that

$$\begin{aligned}\min_{a\in\mathcal{A}} \log \mathbb{E}_{\Pi(\theta|\tilde{X}_n)}[\exp(R(a,\theta))] &\geq \min_{a\in\mathcal{A}} \max_{q\in\mathcal{Q}} \mathcal{F}(a; Q(\theta), \tilde{X}_n) \\ &= \min_{a\in\mathcal{A}} \max_{q\in\mathcal{Q}} -\mathrm{KL}\left(Q(\theta)||\Pi(\theta|\tilde{X}_n)\right) + \int_\Theta R(a,\theta) dQ(\theta).\end{aligned} \quad (2.37)$$

2.7.2 Proof of Theorem 2.3.1:

We prove our main result after series of important lemmas. For brevity we denote $\mathcal{LR}_n(\theta,\theta_0) = \frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}$.

Lemma 2.7.1. *For any $a \in \mathcal{A}$, $\gamma > 0$, and $\zeta > 0$,*

$$\begin{aligned}&\mathbb{E}_{P_0^n}\left[\zeta \int_\Theta L_n(\theta,\theta_0)\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \\ &\leq \log \mathbb{E}_{P_0^n}\left[\int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)} \frac{\mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}\right] + \inf_{Q\in\mathcal{Q}} \mathbb{E}_{P_0^n}\Bigg[\mathrm{KL}(Q(\theta)\|\Pi(\theta|\tilde{X}_n)) \\ &\quad - \gamma \inf_{a\in\mathcal{A}} \mathbb{E}_Q[R(a,\theta)]\Bigg] + \log \mathbb{E}_{P_0^n}\left[\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \frac{\mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}\right].\end{aligned} \quad (2.38)$$

Proof. For any fixed $a \in \mathcal{A}, \gamma > 0$, and $\zeta > 0$, and using the fact that KL is non-negative, observe that the integral in the LHS of equation (2.38) satisfies,

$$\begin{aligned}
\zeta \int_\Theta L_n(\theta, \theta_0)\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n) &\leq \int_\Theta \log \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n) \\
&\qquad + \mathrm{KL}\left(dQ^*_{a,\gamma}(\theta|\tilde{X}_n) \middle\| \frac{\mathrm{e}^{\zeta L_n(\theta,\theta_0)}\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n)}{\int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n)}\right) \\
&= \int_\Theta \log \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\ + \log \int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n) \\
&\qquad + \int_\Theta dQ^*_{a,\gamma}(\theta|\tilde{X}_n) \log \frac{dQ^*_{a,\gamma}(\theta|\tilde{X}_n)}{\mathrm{e}^{\zeta L_n(\theta,\theta_0)}\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n)} \\
&= \log \int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n) + \int_\Theta dQ^*_{a,\gamma}(\theta|\tilde{X}_n) \log \frac{dQ^*_{a,\gamma}(\theta|\tilde{X}_n)}{\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n)}.
\end{aligned}$$

Next, using the definition of $Q^*_{a,\gamma}(\theta|\tilde{X}_n)$ in the second term of last equality, for any other $Q(\cdot) \in \mathcal{Q}$

$$\zeta \int_\Theta L_n(\theta, \theta_0)\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n) \leq \log \int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n) + \int_\Theta dQ(\theta) \log \frac{dQ(\theta)}{\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n)}.$$

Finally, it follows from the definition of the posterior distribution that

$$\begin{aligned}
&\zeta \int_\Theta L_n(\theta, \theta_0)\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n) \\
&\quad \leq \log \int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\mathrm{e}^{\gamma R(a,\theta)}\ \frac{\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)} + \int_\Theta dQ(\theta) \log \frac{dQ(\theta)}{\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n)}, \\
&\quad = \log \int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\ \frac{\mathrm{e}^{\gamma R(a,\theta)}\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)} + \int_\Theta dQ(\theta) \log \frac{dQ(\theta)}{\mathrm{e}^{\gamma R(a,\theta)}\ d\Pi(\theta|\tilde{X}_n)} \\
&\quad + \log \int_\Theta\ \mathrm{e}^{\gamma R(a,\theta)} \frac{\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)},
\end{aligned} \tag{2.39}$$

where the last equality follows from adding and subtracting $\log \int_\Theta e^{\gamma R(a,\theta)} \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)$. Now taking expectation on either side of equation (2.39) and using Jensen's inequality on the first and the last term in the RHS yields

$$\begin{aligned}
&\mathbb{E}_{P_0^n}\left[\zeta \int_\Theta L_n(\theta,\theta_0)\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \\
&\leq \log \mathbb{E}_{P_0^n}\left[\int_\Theta e^{\zeta L_n(\theta,\theta_0)} \frac{e^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}{\int_\Theta e^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}\right] + \inf_{Q\in\mathcal{Q}} \mathbb{E}_{P_0^n}\left[\int_\Theta dQ(\theta) \log \frac{dQ(\theta)}{d\Pi(\theta|\tilde{X}_n)}\right. \\
&\quad \left. - \gamma \inf_{a\in\mathcal{A}} \int_\Theta Q(\theta) R(a,\theta) d\theta\right] + \log \mathbb{E}_{P_0^n}\left[\int_\Theta e^{\gamma R(a,\theta)}\ \frac{\mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}\right], \qquad (2.40)
\end{aligned}$$

where in the second term in RHS of (2.39), we first take infimum over all $a \in \mathcal{A}$ which upper bounds the second term in (2.39) and then take infimum over all $Q \in \mathcal{Q}$, since the LHS does not depend on Q. □

Next, we state a technical result that is important in proving our next lemma.

Lemma 2.7.2 (Lemma 6.4 of [28]). *Suppose random variable X satisfies*

$$\mathbb{P}(X \geq t) \leq c_1 \exp(-c_2 t),$$

for all $t \geq t_0 > 0$. Then for any $0 < \beta \leq c_2/2$,

$$\mathbb{E}[\exp(\beta X)] \leq \exp(\beta t_0) + c_1.$$

Proof. Refer Lemma 6.4 of [28].

□

In the following result, we bound the first term on the RHS of equation (2.38).

Lemma 2.7.3. *Under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, and 2.2.5 and for* $\min(C, C_4(\gamma) + C_5(\gamma)) > C_2 + C_3 + C_4(\gamma) + 2$ *and any $\epsilon \geq \epsilon_n$,*

$$\mathbb{E}_{P_0^n}\left[\int_\Theta e^{\zeta L_n(\theta,\theta_0)} \frac{e^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}{\int_\Theta e^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0) d\Pi(\theta)}\right] \leq e^{\zeta C_1 n\epsilon^2} + (1 + C_0 + 3W^{-\gamma}), \qquad (2.41)$$

for $0 < \zeta \leq C_{10}/2$, where $C_{10} = \min\{\lambda, C, 1\}/C_1$ for any $\lambda > 0$.

Proof. First define the set

$$B_n := \left\{ \tilde{X}_n : \int_\Theta \mathcal{LR}_n(\theta, \theta_0) d\Pi(\theta) \geq \mathrm{e}^{-(1+C_3)n\epsilon^2} \Pi(A_n) \right\}, \tag{2.42}$$

where set A_n is defined in Assumption 2.2.3. We demonstrate that, under Assumption 2.2.3, $P_0^n\left(B_n^c\right)$ is bounded above by an exponentially decreasing(in n) term. Note that for A_n as defined in Assumption 2.2.3:

$$\begin{aligned} &\mathbb{P}_0^n \left(\frac{1}{\Pi(A_n)} \int_\Theta \mathcal{LR}_n(\theta, \theta_0) d\Pi(\theta) \leq \mathrm{e}^{-(1+C_3)n\epsilon^2} \right) \\ &\quad \leq \mathbb{P}_0^n \left(\frac{1}{\Pi(A_n)} \int_{\Theta \cap A_n} \mathcal{LR}_n(\theta, \theta_0) d\Pi(\theta) \leq \mathrm{e}^{-(1+C_3)n\epsilon^2} \right). \end{aligned} \tag{2.43}$$

Let $d\tilde{\Pi}(\theta) := \frac{\mathbb{1}_{\{\Theta \cap A_n\}}(\theta)}{\Pi(A_n)} d\Pi(\theta)$, and use this in (2.43) for any $\lambda > 0$ to obtain,

$$\begin{aligned} \mathbb{P}_0^n \left(\frac{1}{\Pi(A_n)} \int_\Theta \mathcal{LR}_n(\theta, \theta_0) d\Pi(\theta) \leq \mathrm{e}^{-(1+C_3)n\epsilon^2} \right) &\leq \mathbb{P}_0^n \left(\int_\Theta \mathcal{LR}_n(\theta, \theta_0) d\tilde{\Pi}(\theta) \leq \mathrm{e}^{-(1+C_3)n\epsilon^2} \right) \\ &= \mathbb{P}_0^n \left(\left[\int_\Theta \mathcal{LR}_n(\theta, \theta_0) d\tilde{\Pi}(\theta) \right]^{-\lambda} \geq \mathrm{e}^{(1+C_3)\lambda n\epsilon^2} \right). \end{aligned}$$

Then, using the Chernoff's inequality in the last equality above, we have

$$\begin{aligned} \mathbb{P}_0^n \left(\frac{1}{\Pi(A_n)} \int_\Theta \mathcal{LR}_n(\theta, \theta_0) d\Pi(\theta) \leq \mathrm{e}^{-(1+C_3)n\epsilon^2} \right) &\leq \mathrm{e}^{-(1+C_3)\lambda n\epsilon^2} \mathbb{E}_{P_0^n} \left(\left[\int_\Theta \mathcal{LR}_n(\theta, \theta_0) d\tilde{\Pi}(\theta) \right]^{-\lambda} \right) \\ &\leq \mathrm{e}^{-(1+C_3)\lambda n\epsilon^2} \left[\int_\Theta \mathbb{E}_{P_0^n} \left(\left[\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)} \right]^{-\lambda} \right) d\tilde{\Pi}(\theta) \right] \\ &= \mathrm{e}^{-(1+C_3)\lambda n\epsilon^2} \left[\int_\Theta \exp(\lambda D_{\lambda+1}\left(P_0^n \| P_\theta^n\right)) d\tilde{\Pi}(\theta) \right] \\ &\leq \mathrm{e}^{-(1+C_3)\lambda n\epsilon^2} \mathrm{e}^{\lambda C_3 n\epsilon_n^2} \leq \epsilon^{-\lambda n\epsilon^2}, \end{aligned} \tag{2.44}$$

where the second inequality follows from first applying Jensen's inequality (on the term inside $[\,\cdot\,]$) and then using Fubini's theorem, and the penultimate inequality follows from Assumption 2.2.3 and the definition of $\tilde{\Pi}(\theta)$.

Next, define the set $K_n := \{\theta \in \Theta : L_n(\theta, \theta_0) > C_1 n\epsilon^2\}$. Notice that set K_n is the set of alternate hypothesis as defined in Assumption 2.2.1. We bound the calibrated posterior

probability of this set K_n to get a bound on the first term in the RHS of equation (2.38). Recall the sequence of test function $\{\phi_{n,\epsilon}\}$ from Assumption 2.2.1. Observe that

$$\begin{aligned}
&\mathbb{E}_{P_0^n}\left[\frac{\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \\
&\quad= \mathbb{E}_{P_0^n}\left[(\phi_{n,\epsilon})\frac{\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] + \mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\frac{\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \\
&\quad\le \mathbb{E}_{P_0^n}\phi_{n,\epsilon} + \mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\mathbb{1}_{B_n^C}\right] + \mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\mathbb{1}_{B_n}\frac{\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \\
&\quad\le \mathbb{E}_{P_0^n}\phi_{n,\epsilon} + \mathbb{E}_{P_0^n}\left[\mathbb{1}_{B_n^C}\right] + \mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\mathbb{1}_{B_n}\frac{\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right], \qquad (2.45)
\end{aligned}$$

where in the second inequality, we first divide the second term over set B_n and its complement, and then use the fact that $\frac{\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)} \le 1$. The third inequality is due the fact that $\phi_{n,\epsilon} \in [0,1]$. Next, using Assumption 2.2.3 and 2.2.5 observe that on set B_n

$$\begin{aligned}
\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta) &\ge W^\gamma \int_\Theta \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta) \\
&\ge W^\gamma \mathrm{e}^{-(1+C_2+C_3)n\epsilon_n^2} \ge W^\gamma \mathrm{e}^{-(1+C_2+C_3)n\epsilon^2}.
\end{aligned}$$

Substituting the equation above in the third term of equation (2.45), we obtain

$$\begin{aligned}
&\mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\mathbb{1}_{B_n}\frac{\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \\
&\le W^{-\gamma}\mathrm{e}^{(1+C_2+C_3)n\epsilon^2}\mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\mathbb{1}_{B_n}\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)\right] \\
&\le W^{-\gamma}\mathrm{e}^{(1+C_2+C_3)n\epsilon^2}\mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)\right]. \qquad (\star)
\end{aligned}$$

Now using Fubini's theorem observe that,

$$\begin{aligned}
(\star) &= W^{-\gamma}\mathrm{e}^{(1+C_2+C_3)n\epsilon^2}\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\mathbb{E}_{P_\theta^n}\left[(1-\phi_{n,\epsilon})\right]d\Pi(\theta) \\
&\le W^{-\gamma}\mathrm{e}^{(1+C_2+C_3+C_4(\gamma))n\epsilon^2}\left[\int_{K_n\cap\{\mathrm{e}^{\gamma R(a,\theta)}\le \mathrm{e}^{C_4(\gamma)n\epsilon^2}\}} \mathbb{E}_{P_\theta^n}\left[(1-\phi_{n,\epsilon})\right]d\Pi(\theta)\right. \\
&\qquad\left. + \mathrm{e}^{-C_4(\gamma)n\epsilon^2}\int_{K_n\cap\{\mathrm{e}^{\gamma R(a,\theta)}> \mathrm{e}^{C_4(\gamma)n\epsilon^2}\}} \mathrm{e}^{\gamma R(a,\theta)}d\Pi(\theta)\right],
\end{aligned}$$

where in the last inequality, we first divide the integral over set $\{\theta \in \Theta : \mathrm{e}^{\gamma R(a,\theta)} \leq \mathrm{e}^{C_4(\gamma)n\epsilon^2}\}$ and its complement and then use the upper bound on $\mathrm{e}^{\gamma R(a,\theta)}$ in the first integral. Now, it follows that

$$
\begin{aligned}
(\star) &\leq W^{-\gamma}\mathrm{e}^{(1+C_2+C_3+C_4(\gamma))n\epsilon^2}\left[\int_{K_n}\mathbb{E}_{P_\theta^n}\left[(1-\phi_{n,\epsilon})\right]d\Pi(\theta) + \mathrm{e}^{-C_4(\gamma)n\epsilon^2}\int_{\{\mathrm{e}^{\gamma R(a,\theta)}>\mathrm{e}^{C_4(\gamma)n\epsilon^2}\}}\mathrm{e}^{\gamma R(a,\theta)}d\Pi(\theta)\right] \\
&= W^{-\gamma}\mathrm{e}^{(1+C_2+C_3+C_4(\gamma))n\epsilon^2}\left[\int_{K_n\cap\Theta_n(\epsilon)}\mathbb{E}_{P_\theta^n}\left[(1-\phi_{n,\epsilon})\right]d\Pi(\theta) + \int_{K_n\cap\Theta_n(\epsilon)^c}\mathbb{E}_{P_\theta^n}\left[(1-\phi_{n,\epsilon})\right]d\Pi(\theta)\right. \\
&\qquad\qquad \left. + \mathrm{e}^{-C_4(\gamma)n\epsilon^2}\int_{\{\mathrm{e}^{\gamma R(a,\theta)}>\mathrm{e}^{C_4(\gamma)n\epsilon^2}\}}\mathrm{e}^{\gamma R(a,\theta)}d\Pi(\theta)\right] \\
&\leq W^{-\gamma}\mathrm{e}^{(1+C_2+C_3+C_4(\gamma))n\epsilon^2}\left[\int_{K_n\cap\Theta_n(\epsilon)}\mathbb{E}_{P_\theta^n}\left[(1-\phi_{n,\epsilon})\right]d\Pi(\theta) + \Pi(\Theta_n(\epsilon)^c)\right. \\
&\qquad\qquad \left. + \mathrm{e}^{-C_4(\gamma)n\epsilon^2}\int_{\{\mathrm{e}^{\gamma R(a,\theta)}>\mathrm{e}^{C_4(\gamma)n\epsilon^2}\}}\mathrm{e}^{\gamma R(a,\theta)}d\Pi(\theta)\right],
\end{aligned}
$$

where the second equality is obtained by dividing the first integral on set $\Theta_n(\epsilon)$ and its complement, and the third inequality is due the fact that $\phi_{n,\epsilon} \in [0,1]$. Now, using the equation above and Assumption 2.2.1, 2.2.2, and 2.2.4 observe that

$$
\mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\mathbb{1}_{B_n}\frac{\int_{K_n}\mathrm{e}^{\gamma R(a,\theta)}\,\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\,\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \leq W^{-\gamma}\mathrm{e}^{(1+C_2+C_3+C_4(\gamma))n\epsilon^2}\left[2\mathrm{e}^{-Cn\epsilon^2} + \mathrm{e}^{-(C_5(\gamma)+C_4(\gamma))n\epsilon^2}\right].
$$

Hence, choosing $C, C_2, C_3, C_4(\gamma)$ and $C_5(\gamma)$ such that $-1 > 1+C_2+C_3+C_4(\gamma)-\min(C,(C_4(\gamma)+C_5(\gamma)))$ implies

$$
\mathbb{E}_{P_0^n}\left[(1-\phi_{n,\epsilon})\mathbb{1}_{B_n}\frac{\int_{K_n}\mathrm{e}^{\gamma R(a,\theta)}\,\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\,\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \leq 3W^{-\gamma}\mathrm{e}^{-n\epsilon^2}. \tag{2.46}
$$

By Assumption 2.2.1, we have

$$
\mathbb{E}_{P_0^n}\phi_{n,\epsilon} \leq C_0\mathrm{e}^{-Cn\epsilon^2}. \tag{2.47}
$$

Therefore, substituting equation (2.44), equation (2.46), and (2.47) into (2.45), we obtain

$$\mathbb{E}_{P_0^n}\left[\frac{\int_{K_n} \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \leq (1+C_0+3W^{-\gamma})\mathrm{e}^{-C_{10}C_1 n\epsilon^2}, \tag{2.48}$$

where $C_{10} = \min\{\lambda, C, 1\}/C_1$. Using Fubini's theorem, observe that the LHS in the equation (2.48) can be expressed as $\mu(K_n)$, where

$$d\mu(\theta) = \mathbb{E}_{P_0^n}\left[\frac{\mathcal{LR}_n(\theta,\theta_0)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right]\Pi(\theta)\mathrm{e}^{\gamma R(a,\theta)}d\theta.$$

Next, recall that the set $K_n = \{\theta \in \Theta : L_n(\theta,\theta_0) > C_1 n\epsilon^2\}$. Applying Lemma 2.7.2 above with $X = L_n(\theta,\theta_0)$, $c_1 = (1+C_0+3W^{-\gamma})$, $c_2 = C_{10}$, $t_0 = C_1 n\epsilon_n^2$, and for $0 < \zeta \leq C_{10}/2$, we obtain

$$\mathbb{E}_{P_0^n}\left[\int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\frac{\mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}d\theta\right] \leq \mathrm{e}^{\zeta C_1 n\epsilon_n^2} + (1+C_0+3W^{-\gamma}). \tag{2.49}$$

□

Further, we have another technical lemma, that will be crucial in proving the subsequent lemma that upper bounds the last term in the equation (2.38).

Lemma 2.7.4. *Suppose a positive random variable X satisfies*

$$\mathbb{P}(X \geq \mathrm{e}^t) \leq c_1 \exp(-(c_2+1)t),$$

for all $t \geq t_0 > 0$, $c_1 > 0$, and $c_2 > 0$. Then,

$$\mathbb{E}[X] \leq \exp(t_0) + \frac{c_1}{c_2}.$$

Proof. For any $Z_0 > 1$,

$$\mathbb{E}[X] \leq Z_0 + \int_{Z_0}^\infty \mathbb{P}(X \geq x)dx = Z_0 + \int_{\ln Z_0}^\infty \mathbb{P}(X \geq \mathrm{e}^y)\mathrm{e}^y dy \leq Z_0 + c_1\int_{\ln Z_0}^\infty \exp(-c_2 y)dy.$$

Therefore, choosing $Z_0 = \exp(t_0)$,

$$\mathbb{E}[X] \leq \exp(t_0) + \frac{c_1}{c_2}\exp(-c_2 t_0) \leq \exp(t_0) + \frac{c_1}{c_2}.$$

□

Next, we establish the following bound on the last term in equation (2.38).

Lemma 2.7.5. *Under Assumptions 2.2.1, 2.2.2, 2.2.3, 2.2.4, 2.2.5, and for* $C_4(\gamma)+C_5(\gamma) > C_2 + C_3 + 2$,

$$\mathbb{E}_{P_0^n}\left[\int_\Theta \frac{\mathrm{e}^{\gamma R(a,\theta)}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \leq \mathrm{e}^{C_4(\gamma)n\epsilon_n^2} + 2C_4(\gamma). \tag{2.50}$$

Proof. Define the set

$$M_n := \{\theta \in \Theta : \mathrm{e}^{\gamma R(a,\theta)} > \mathrm{e}^{C_4(\gamma)n\epsilon^2}\}. \tag{2.51}$$

Using the set B_n in equation (2.42), observe that the measure of the set M_n, under the posterior distribution satisfies,

$$\mathbb{E}_{P_0^n}\left[\frac{\int_{M_n}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \leq \mathbb{E}_{P_0^n}\left[\mathbb{1}_{B_n^c}\right] + \mathbb{E}_{P_0^n}\left[\mathbb{1}_{B_n}\frac{\int_{M_n}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right]. \tag{2.52}$$

Now, the second term of equation (2.52) can be bounded as follows: recall Assumption 2.2.3 and the definition of set B_n, both together imply that,

$$\begin{aligned}\mathbb{E}_{P_0^n}\left[\mathbb{1}_{B_n}\frac{\int_{M_n}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] &\leq \mathrm{e}^{(1+C_2+C_3)n\epsilon^2}\mathbb{E}_{P_0^n}\left[\mathbb{1}_{B_n}\int_{M_n}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)\right] \\ &\leq \mathrm{e}^{(1+C_2+C_3)n\epsilon^2}\mathbb{E}_{P_0^n}\left[\int_{M_n}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)\right]. \qquad (\star\star)\end{aligned}$$

Then, using Fubini's Theorem $(\star\star) = \mathrm{e}^{(1+C_2+C_3)n\epsilon^2}\Pi(M_n)$. Next, using the definition of set M_n and then Assumption 2.2.4, we obtain

$$\begin{aligned}\mathbb{E}_{P_0^n}\left[\mathbb{1}_{B_n}\frac{\int_{M_n}\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta\ \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] &\leq \mathrm{e}^{(1+C_2+C_3)n\epsilon^2}\mathrm{e}^{-C_4(\gamma)n\epsilon^2}\int_{M_n}\mathrm{e}^{\gamma R(a,\theta)}d\Pi(\theta) \\ &\leq \mathrm{e}^{(1+C_2+C_3)n\epsilon^2}\mathrm{e}^{-C_4(\gamma)n\epsilon^2}\mathrm{e}^{-C_5(\gamma)n\epsilon^2},\end{aligned}$$

Hence, choosing the constants $C_2, C_3, C_4(\gamma)$ and $C_5(\gamma)$ such that $-1 > 1 + C_2 + C_3 - C_5(\gamma)$ implies

$$\mathbb{E}_{P_0^n}\left[\mathbb{1}_{B_n}\frac{\int_{M_n}\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \leq \mathrm{e}^{-(1+C_4(\gamma))n\epsilon^2} \tag{2.53}$$

Therefore, substituting (2.44) and (2.53) into (2.52)

$$\mathbb{E}_{P_0^n}\left[\frac{\int_{M_n}\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \leq 2\mathrm{e}^{-C_4(\gamma)(C_{11}(\gamma)+1)n\epsilon^2}, \tag{2.54}$$

where $C_{11} = \min\{\lambda, 1 + C_4(\gamma)\}/C_4(\gamma) - 1$. Using Fubini's theorem, observe that the RHS in (2.54) can be expressed as $\nu(M_n)$, where the measure

$$d\nu(\theta) = \mathbb{E}_{P_0^n}\left[\frac{\mathcal{LR}_n(\theta,\theta_0)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right]d\Pi(\theta).$$

Applying Lemma 2.7.4 for $X = \mathrm{e}^{\gamma R(a,\theta)}$,$c_1 = 2$, $c_2 = C_{11}(\gamma)$, $t_0 = C_4(\gamma)n\epsilon_n^2$ and $\lambda \geq 1 + C_4(\gamma)$, we obtain

$$\mathbb{E}_{P_0^n}\left[\int_\Theta \frac{\mathrm{e}^{\gamma R(a,\theta)}\,\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] \leq \mathrm{e}^{C_4 n\epsilon_n^2} + \frac{2}{C_{11}(\gamma)} \leq \mathrm{e}^{C_4 n\epsilon_n^2} + 2C_4(\gamma). \tag{2.55}$$

□

Proof. Proof of Theorem 2.3.1: Finally, recall (2.38),

$$\begin{aligned}
&\zeta\mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0)\; dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \\
&\quad\leq \log \mathbb{E}_{P_0^n}\left[\int_\Theta \mathrm{e}^{\zeta L_n(\theta,\theta_0)}\frac{\mathrm{e}^{\gamma R(a,\theta)}\,\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\,\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right] + \inf_{Q\in\mathcal{Q}}\mathbb{E}_{P_0^n}\Big[\mathrm{KL}(Q(\theta)\|\Pi(\theta|\tilde{X}_n)) \\
&\qquad - \gamma\inf_{a\in\mathcal{A}}\mathbb{E}_Q[R(a,\theta)]\Big] + \log\mathbb{E}_{P_0^n}\left[\int_\Theta \mathrm{e}^{\gamma R(a,\theta)}\;\frac{\mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}{\int_\Theta \mathcal{LR}_n(\theta,\theta_0)d\Pi(\theta)}\right].
\end{aligned} \tag{2.56}$$

Substituting (2.50) and (2.41) into the above equation and then using the definition of $\eta_n^R(\gamma)$, we get

$$\begin{aligned}
&\mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0)\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \\
&\quad\le \frac{1}{\zeta}\left\{\log(\mathrm{e}^{\zeta C_1 n\epsilon_n^2}+(1+C_0+3W^{-\gamma}))+\log\left(\mathrm{e}^{C_4(\gamma)n\epsilon_n^2}+2C_4(\gamma)\right)+n\eta_n^R(\gamma)\right\} \\
&\quad\le \left(C_1+\frac{1}{\zeta}C_4(\gamma)\right)n\epsilon_n^2+\frac{1}{\zeta}n\eta_n^R(\gamma)+\frac{(1+C_0+3W^{-\gamma})\mathrm{e}^{(-\zeta C_1 n\epsilon_n^2)}}{\zeta}+\frac{2C_4(\gamma)\mathrm{e}^{-C_4(\gamma)n\epsilon_n^2}}{\zeta},
\end{aligned}$$

where the last inequality uses the fact that $\log x \le x-1$. Choosing $\zeta = C_{10}/2 = \frac{\min(C,\lambda,1)}{2C_1}$,

$$\begin{aligned}
&\mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0)\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \\
&\quad\le M(\gamma)n(\epsilon_n^2)+Mn\eta_n^R(\gamma)+\frac{2(1+C_0+3W^{-\gamma})\mathrm{e}^{(-\frac{C_{10}}{2}n\epsilon_n^2)}}{C_{10}}+\frac{4C_4(\gamma)\mathrm{e}^{-C_4(\gamma)n\epsilon_n^2}}{C_{10}} \qquad (2.57)
\end{aligned}$$

where $M(\gamma) = C_1+\frac{1}{\zeta}C_4(\gamma)$ and $M=\frac{1}{\zeta}$ depend on $C, C_1, C_4(\gamma), W$ and λ. Since the last two terms in (2.57) decrease and the first term increases as n increases, we can choose M large enough, such that for all $n \ge 1$

$$Mn\eta_n^R(\gamma) > \frac{2(1+C_0+3W^{-\gamma})}{C_{10}}+\frac{4C_4(\gamma)}{C_{10}},$$

and therefore for $M = 2M$,

$$\mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0)\ dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \le M(\gamma)n(\epsilon_n^2)+Mn\eta_n^R(\gamma). \qquad (2.58)$$

Also, observe that the LHS in the above equation is always positive, therefore $M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma) \ge 0\ \forall n \ge 1$ and $\gamma > 0$.

□

2.7.3 Proof of Theorem 2.3.2 and 2.3.3

Lemma 2.7.6. *Given $a \in \mathcal{A}$ and for a constant M, as defined in Theorem 2.3.1*

$$\mathbb{E}_{P_0^n}\left[\sup_{a\in\mathcal{A}}\left|\mathbb{E}_{Q^*_{a,\gamma}(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)\right|\right] \leq \left[M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma)\right]^{\frac{1}{2}}. \tag{2.59}$$

Proof. First, observe that

$$\begin{aligned}\left(\sup_{a\in\mathcal{A}}\left|\mathbb{E}_{Q^*_{a,\gamma}(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta)\right|\right)^2 &\leq \left(\int \sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right)^2 \\ &\leq \int\left(\sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|\right)^2 dQ^*_{a,\gamma}(\theta|\tilde{X}_n),\end{aligned}$$

where the last inequality follows from Jensen's inequality. Now, using the Jensen's inequality again

$$\begin{aligned}&\left(\mathbb{E}_{P_0^n}\left[\sup_{a\in\mathcal{A}}\left|\mathbb{E}_{Q^*_{a,\gamma}(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)\right|\right]\right)^2 \\ &\qquad\leq \mathbb{E}_{P_0^n}\left[\left(\sup_{a\in\mathcal{A}}\left|\mathbb{E}_{Q^*_{a,\gamma}(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)\right|\right)^2\right].\end{aligned}$$

Now, using Theorem 2.3.1 the result follows immediately.

□

Proof of Theorem 2.3.2. Observe that

$$\begin{aligned}&R(\mathtt{a}^*_{\mathtt{RS}},\theta_0) - \inf_{z\in\mathcal{A}} R(z,\theta_0) \\ &= |R(\mathtt{a}^*_{\mathtt{RS}},\theta_0) - \inf_{z\in\mathcal{A}} R(z,\theta_0)| \\ &= R(\mathtt{a}^*_{\mathtt{RS}},\theta_0) - \int R(\mathtt{a}^*_{\mathtt{RS}},\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) + \int R(\mathtt{a}^*_{\mathtt{RS}},\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - \inf_{z\in\mathcal{A}} R(z,\theta_0) \\ &\leq \left|R(\mathtt{a}^*_{\mathtt{RS}},\theta_0) - \int R(\mathtt{a}^*_{\mathtt{RS}},\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n)\right| + \left|\int R(\mathtt{a}^*_{\mathtt{RS}},\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - \inf_{a\in\mathcal{A}} R(a,\theta_0)\right| \\ &\leq 2\sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - R(a,\theta_0)\right|.\end{aligned}$$

Given $\mathtt{a}^*_{\mathtt{RS}} \in \mathcal{A}$ and for a constant M (defined in Theorem 2.3.1), we have from Lemma 2.7.6 for $a = \mathtt{a}^*_{\mathtt{RS}}$

$$\mathbb{E}_{P_0^n}\left[\sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - R(a,\theta_0)\right|\right] \leq \left[M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma)\right]^{\frac{1}{2}}. \tag{2.60}$$

It follows from above that the P_0^n- probability of the following event is at least $1-\tau^{-1}$:

$$\left\{\tilde{X}_n : R(\mathtt{a}^*_{\mathtt{RS}},\theta_0) - \inf_{z\in\mathcal{A}} R(z,\theta_0) \leq 2\tau\left[M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma)\right]^{\frac{1}{2}}\right\}. \tag{2.61}$$

□

Proof of Theorem 2.3.3: Since, the above result holds for any $a \in \mathcal{A}$, fix $a = \mathtt{a}^*_{\mathtt{RS}}$ and observe that for any $\gamma > 0$ and $\tau > 0$, the result in Lemma 2.7.6 implies that P_0^n- probability of

$$\left\{\left[M(\epsilon_n^2 + \eta_n^R(\gamma))\right]^{-\frac{1}{2}} \sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - R(a,\theta_0)\right| > \tau\right\} \tag{2.62}$$

is at most τ^{-1}. For $\mathtt{a}^*_{\mathtt{RS}}$, it follows from the definition of $\Psi(\cdot)$ that

$$\begin{aligned}
&\Psi\left(H(\mathtt{a}^*_{\mathtt{RS}}, \arg\min_{a\in\mathcal{A}} R(a,\theta_0))\right)\\
&\leq R(\mathtt{a}^*_{\mathtt{RS}},\theta_0) - \inf_{z\in\mathcal{A}} R(a,\theta_0)\\
&= R(\mathtt{a}^*_{\mathtt{RS}},\theta_0) - \int R(\mathtt{a}^*_{\mathtt{RS}},\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) + \int R(\mathtt{a}^*_{\mathtt{RS}},\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - \inf_{z\in\mathcal{A}} R(z,\theta_0)\\
&\leq \left|R(\mathtt{a}^*_{\mathtt{RS}},\theta_0) - \int R(\mathtt{a}^*_{\mathtt{RS}},\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n)\right| + \left|\int R(\mathtt{a}^*_{\mathtt{RS}},\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - \inf_{a\in\mathcal{A}} R(a,\theta_0)\right|\\
&\leq 2\sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - R(a,\theta_0)\right|.
\end{aligned} \tag{2.63}$$

It follows from the above inequality that

$$\begin{aligned}
&\left\{\left[M(\epsilon_n^2 + \eta_n^R(\gamma))\right]^{-\frac{1}{2}} \Psi\left(H(\mathtt{a}^*_{\mathtt{RS}}, \arg\min_{z\in\mathcal{A}} R(a,\theta_0))\right) > 2\tau\right\}\\
&\qquad\subseteq \left\{\left[M(\epsilon_n^2 + \eta_n^R(\gamma))\right]^{-\frac{1}{2}} \sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*_{\mathtt{a}^*_{\mathtt{RS}},\gamma}(\theta|\tilde{X}_n) - R(a,\theta_0)\right| > \tau\right\}.
\end{aligned} \tag{2.64}$$

Therefore, using the condition on the growth function in the statement of the theorem that, $\frac{\Psi\left(H\left(\mathtt{a}^*_{\mathtt{RS}}, \arg\min_{a\in\mathcal{A}} R(a,\theta_0)\right)\right)}{H\left(\mathtt{a}^*_{\mathtt{RS}}, \arg\min_{a\in\mathcal{A}} R(a,\theta_0)\right)^{\delta}} = \kappa$, the P_0^n- probability of the following event is at least $1-\tau^{-1}$:

$$\left\{H(\mathtt{a}^*_{\mathtt{RS}}, \arg\min_{a\in\mathcal{A}} R(a,\theta_0)) \leq \tau^{\frac{1}{\delta}} \left[\frac{2\left[M(\gamma)\epsilon_n^2 + M\eta_n^R(\gamma)\right]^{\frac{1}{2}}}{\kappa}\right]^{\frac{1}{\delta}}\right\}. \tag{2.65}$$

This concludes the proof.

□

2.7.4 Proofs in Section 2.3.1

Proof of Proposition 2.3.1. Using the definition of $\eta_n^R(\gamma)$ and the posterior distribution $\Pi(\theta|\tilde{X}_n)$, observe that

$$\begin{aligned}
n\eta_n^R(\gamma) &= \inf_{Q\in\mathcal{Q}} \mathbb{E}_{P_0^n}\left[\mathrm{KL}(Q(\theta)\|\Pi(\theta|\tilde{X}_n)) - \gamma \inf_{a\in\mathcal{A}} \mathbb{E}_Q[R(a,\theta)]\right] \\
&= \inf_{Q\in\mathcal{Q}} \mathbb{E}_{P_0^n}\left[\mathrm{KL}(Q(\theta)\|\Pi(\theta)) + \int_\Theta dQ(\theta)\log\left(\frac{\int d\Pi(\theta)p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta)}\right) - \gamma \inf_{a\in\mathcal{A}} \mathbb{E}_Q[R(a,\theta)]\right] \\
&= \inf_{Q\in\mathcal{Q}}\left[\mathrm{KL}(Q(\theta)\|\Pi(\theta)) - \gamma \inf_{a\in\mathcal{A}} \mathbb{E}_Q[R(a,\theta)] + \mathbb{E}_{P_0^n}\left[\int_\Theta dQ(\theta)\log\left(\frac{\int d\Pi(\theta)p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta)}\right)\right]\right].
\end{aligned}$$

Now, using Fubini's in the last term of the equation above, we obtain

$$\begin{aligned}
n\eta_n^R(\gamma) = \inf_{Q\in\mathcal{Q}}\Bigg[&\mathrm{KL}(Q(\theta)\|\Pi(\theta)) - \gamma \inf_{a\in\mathcal{A}} \mathbb{E}_Q[R(a,\theta)] \\
&+ \mathbb{E}_Q\left[\mathrm{KL}\left(dP_0^n\|p(\tilde{X}_n|\theta)\right) - \mathrm{KL}\left(dP_0^n\middle\|\int d\Pi(\theta)p(\tilde{X}_n|\theta)\right)\right]\Bigg].
\end{aligned} \tag{2.66}$$

Observe that, $\int_{\mathcal{X}^n} \int d\Pi(\theta) p(\tilde{X}_n|\theta) d\tilde{X}_n = 1$. Since, KL is always non-negative, it follows from the equation above that

$$\begin{aligned}
&\eta_n^R(\gamma) \\
&\leq \frac{1}{n} \inf_{Q \in \mathcal{Q}} \left[\mathrm{KL}\left(Q(\theta) \| \Pi(\theta)\right) - \gamma \inf_{a \in \mathcal{A}} \mathbb{E}_Q[R(a,\theta)] + \mathbb{E}_Q\left[\mathrm{KL}\left(dP_0^n \| p(\tilde{X}_n|\theta)\right)\right]\right] \\
&\leq \frac{1}{n} \inf_{Q \in \mathcal{Q}} \left[\mathrm{KL}\left(Q(\theta) \| \Pi(\theta)\right) + \mathbb{E}_Q\left[\mathrm{KL}\left(dP_0^n \| p(\tilde{X}_n|\theta)\right)\right]\right] - \frac{1}{n}\gamma \inf_{Q \in \mathcal{Q}} \inf_{a \in \mathcal{A}} \mathbb{E}_Q[R(a,\theta)], \quad (2.67)
\end{aligned}$$

where the last inequality follows from the following fact, for any functions $f(\cdot)$ and $g(\cdot)$,

$$\inf(f - g) \leq \inf f - \inf g.$$

Recall $\epsilon_n \geq \frac{1}{\sqrt{n}}$. Now, using Assumption 2.3.1, it is straightforward to observe that the first term in (2.67),

$$\frac{1}{n} \inf_{Q \in \mathcal{Q}} \left[\mathrm{KL}\left(Q(\theta) \| \Pi(\theta)\right) + \mathbb{E}_Q\left[\mathrm{KL}\left(dP_0^n \| p(\tilde{X}_n|\theta)\right)\right]\right] \leq C_9 \epsilon_n^2. \quad (2.68)$$

Now consider the last term in (2.67). Notice that the coefficient of $\frac{1}{n}$ is independent of n and is bounded from below. Therefore, there exist a constant $C_8 = -\inf_{Q \in \mathcal{Q}} \inf_{a \in \mathcal{A}} \mathbb{E}_Q[R(a,\theta)]$, such that with equation (2.68) it follows that $\eta_n^R(\gamma) \leq \gamma n^{-1} C_8 + C_9 \epsilon_n^2$ and the result follows.

□

Proof of Proposition 2.3.2. First recall that

$$\begin{aligned}
n\eta_n^R(\gamma) &= \inf_{Q \in \mathcal{Q}} \mathbb{E}_{P_0^n}\left[\mathrm{KL}(Q(\theta) \| \Pi(\theta|\tilde{X}_n)) - \gamma \inf_{a \in \mathcal{A}} \mathbb{E}_Q[R(a,\theta)]\right] \\
&= \inf_{Q \in \mathcal{Q}} \mathbb{E}_{P_0^n}\left[\mathrm{KL}(Q(\theta) \| \Pi(\theta|\tilde{X}_n))\right] - \gamma \inf_{a \in \mathcal{A}} \mathbb{E}_Q[R(a,\theta)]. \quad (2.69)
\end{aligned}$$

Observe that the optimization problem is equivalent to solving :

$$\min_{Q \in \mathcal{Q}} \mathbb{E}_{P_0^n}\left[\mathrm{KL}(Q(\theta) \| \Pi(\theta|\tilde{X}_n))\right] \text{ s.t. } -\inf_{a \in \mathcal{A}} \mathbb{E}_Q[R(a,\theta)] \leq 0. \quad (2.70)$$

Now for any $\gamma > 0$, $Q^*_\gamma(\theta) \in \mathcal{Q}$ that minimizes the objective in (2.69) is primal feasible if

$$-\inf_{a\in\mathcal{A}} \int_\Theta dQ^*_\gamma(\theta) R(a,\theta) \leq 0.$$

Therefore, it is straightforward to observe that as γ increases $n\eta^R_n(\gamma)$ decreases that is

$$\mathbb{E}_{P^n_0}\left[\int_\Theta dQ^*_\gamma(\theta) \log \frac{dQ^*_\gamma(\theta)}{d\Pi(\theta|\tilde{X}_n)} - \gamma \inf_{a\in\mathcal{A}} \int_\Theta dQ^*_\gamma(\theta) R(a,\theta)\right].$$

□

Proof. Proof of Corollary 2.3.1: For any $\delta > 0$, using Markov inequality

$$\begin{aligned} P^n_0\left(Q^*_{a,\gamma}\left[\left\{\frac{1}{n}L_n(\theta,\theta_0) > M_n(\epsilon^2_n + \eta^R_n(\gamma))\right\}\middle|\tilde{X}_n\right] > \delta\right) &\leq \frac{1}{\delta}\mathbb{E}_{P^n_0} Q^*_{a,\gamma}\left[\left\{\frac{1}{n}L_n(\theta,\theta_0) > M_n(\epsilon^2_n + \eta^R_n(\gamma))\right\}\middle|\tilde{X}_n\right] \\ &\leq \frac{1}{n\delta M_n(\epsilon^2_n + \eta^R_n(\gamma))}\mathbb{E}_{P^n_0}\left[\int_\Theta L_n(\theta,\theta_0) dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \\ &\leq \frac{nM(\epsilon^2_n + \eta^R_n(\gamma))}{n\delta M_n(\epsilon^2_n + \eta^R_n(\gamma))} = \frac{M}{\delta M_n}, \end{aligned}$$

where the last inequality follows from Theorem 2.3.2. Since M_n is a diverging sequence, convergence in P^n_0-probability follows.

□

2.7.5 Proofs in Section 2.3.2

Proof of Lemma 2.3.1: Refer Theorem 7.1 of [91]. □

Proof of Lemma 2.3.2: For any positive k and ϵ, let $\theta \in [\theta_0 - k\epsilon, \theta_0 + k\epsilon]^d \subset \Theta \subset \mathbb{R}^d$. Now consider a set $H_{\mathrm{i}} = \{\theta^0_{\mathrm{i}}, \theta^1_{\mathrm{i}}, \ldots \theta^J_{\mathrm{i}}, \theta^{J+1}_{\mathrm{i}}\}$ and $H = \bigotimes_d H_{\mathrm{i}}$ with $J = \lfloor \frac{2k\epsilon}{\delta} \rfloor$, where $\theta^{\mathrm{j}}_{\mathrm{i}} = \theta_0 - k\epsilon + \mathrm{i}\delta$ for $\mathrm{j} = \{0, 1, \ldots, J\}$ and $\theta^{J+1}_{\mathrm{i}} = \theta_0 + k\epsilon$. Observe that for any $\theta \in [\theta_0 - k\epsilon, \theta_0 + k\epsilon]^d$, there exists a $\theta^{\mathrm{j}} \in H$ such that $\|\theta - \theta^{\mathrm{j}}\| < \delta$. Hence, union of the $\delta-$balls for each element in set H covers $[\theta_0 - k\epsilon, \theta_0 + k\epsilon]^d$, therefore $N(\delta, [\theta_0 - k\epsilon, \theta_0 + k\epsilon]^d, \|\cdot\|) = (J+2)^d$.

Now, due to Assumption 2.3.3, for any $\theta \in [\theta_0 - k\epsilon, \theta_0 + k\epsilon]^d$

$$d_L(\theta,\theta_0) \leq K_2\|\theta - \theta^{\mathrm{j}}\| \leq K_2\delta,$$

For brevity, we denote $n^{-1}L_n(\theta, \theta_0)$ by $d_L(\theta, \theta_0)$, that is

$$d_L(\theta_1, \theta_2) := \sup_{a \in \mathcal{A}} |R(a, \theta_1) - R(a, \theta_2)|, \ \forall \{\theta_1, \theta_2\} \in \Theta, \tag{2.71}$$

and the covering number of the set $T(\epsilon) := \{P_\theta : d_L(\theta, \theta_0) < \epsilon\}$ as $N(\delta, T(\epsilon), d_L)$, where $\delta > 0$ is the radius of each ball in the cover.

Hence, δ-cover of set $[\theta_0 - k\epsilon, \theta_0 + k\epsilon]^d$ is $K_1\delta$ cover of set $T(\epsilon)$ with $k = 1/K_2$. Finally,

$$N(K_2\delta, T(\epsilon), d_L) \leq (J+2)^d \leq \left(\frac{2k\epsilon}{\delta} + 2\right)^d = \left(\frac{2\epsilon}{K_2\delta} + 2\right)^d$$

which implies for $\delta = K_2\delta$,

$$N(\delta, T(\epsilon), d_L) \leq \left(\frac{2\epsilon}{\delta} + 2\right)^s.$$

□

Proof of Lemma 2.3.3: Recall $d_L(\theta, \theta_0) = (\sup_{a \in \mathcal{A}} |R(a, \theta) - R(a, \theta_0)|)$ and $T(\epsilon) = \{P_\theta : d_L(\theta, \theta_0) < \epsilon\}$. Using Lemma 2.3.2, observe that for every $\epsilon > \epsilon_n > 0$,

$$N\left(\frac{\epsilon}{2}, \{\theta : \epsilon \leq d_L(\theta, \theta_0) \leq 2\epsilon\}, d_L\right) \leq N\left(\frac{\epsilon}{2}, \{\theta : d_L(\theta, \theta_0) \leq 2\epsilon\}, d_L\right) < 10^d.$$

Next, using Assumption 2.3.2 we have

$$d_L(\theta, \theta_0) \leq K_1 d_H(\theta, \theta_0).$$

It follows from the above two observations and Lemma 2 that, for every $\epsilon > \epsilon_n > 0$, there exist tests $\{\phi_{n,\epsilon}\}$ such that

$$\mathbb{E}_{P_0^n}[\phi_{n,\epsilon}] \leq 10^d \frac{\exp(-Cn\epsilon^2)}{1 - \exp(-Cn\epsilon^2)}, \tag{2.72}$$

$$\sup_{\{\theta \in \Theta : d_L(\theta, \theta_0) \geq \epsilon\}} \mathbb{E}_{P_\theta^n}[1 - \phi_{n,\epsilon}] \leq \exp(-Cn\epsilon^2), \tag{2.73}$$

where $C = \frac{1}{2K_1^2}$. Since the above two conditions hold for every $\epsilon > \epsilon_n$, we can choose a constant $K > 0$ such that for every $\epsilon > \epsilon_n$

$$\mathbb{E}_{P_0^n}[\phi_{n,\epsilon}] \leq 10^d \frac{\exp(-CK^2 n\epsilon^2)}{1-\exp(-CK^2 n\epsilon^2)} \leq 2(10^d)\exp(-CK^2 n\epsilon^2), \quad (2.74)$$

$$\sup_{\{\theta\in\Theta: L_n(\theta,\theta_0)\geq K^2 n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_{n,\epsilon}] = \sup_{\{\theta\in\Theta: d_L(\theta,\theta_0)\geq K\epsilon\}} \mathbb{E}_{P_\theta^n}[1-\phi_{n,\epsilon}] \leq \exp(-CK^2 n\epsilon^2), \quad (2.75)$$

where the second inequality in (2.74) holds $\forall n \geq n_0$, where $n_0 := \min\{n \geq 1 : CK^2 n\epsilon^2 \geq \log(2)\}$ Hence, the result follows for $C_1 = K^2$ and $C = CK^2$. □

Proof of Corollary 2.3.2: Using Lemma 2.3.3 observe that for any $\Theta_n(\epsilon) \subseteq \Theta$, $L_n(\theta,\theta_0)$ satisfies Assumption 2.2.1 with $C_0 = 2 * 10^s$, $C = \frac{C_1}{2K_1^2}$ and for any $C_1 > 0$, since

$$\sup_{\{\theta\in\Theta_n(\epsilon): L_n(\theta,\theta_0)\geq C_1 n\epsilon_n^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_{n,\epsilon}] \leq \sup_{\{\theta\in\Theta: L_n(\theta,\theta_0)\geq C_1 n\epsilon_n^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_{n,\epsilon}] \leq \exp(-Cn\epsilon_n^2).$$

Hence, applying Theorem 2.3.1 the proof follows. □

2.7.6 Proof of Theorem 2.4.1, 2.4.2, and 2.4.3

Proof of Theorem 2.4.1: The proof follows immediately from Theorem 2.3.1 by taking limit $\gamma \to 0^+$ on either side of its main result, that is

$$\mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0)\, dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \leq Mn(\epsilon_n^2 + \eta_n^R(\gamma)). \quad (2.76)$$

Fix $n \geq 1$. Now first consider the LHS, use the fact that for any $a \in \mathcal{A}$, $\lim_{\gamma\to 0^+} Q^*_{a,\gamma}(\theta|\tilde{X}_n) = Q^*(\theta|\tilde{X}_n)$ (2.13), the integrand is also non-negative , and $n\eta_n^R(\gamma) < \infty$ due to Proposition 2.3.2 (since a decreasing sequence is bounded given $\eta_n^R(0) < \infty$), therefore, using Fatou's Lemma we have

$$\mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0)\, dQ^*(\theta|\tilde{X}_n)\right] \leq \liminf_{\gamma\to 0^+} \mathbb{E}_{P_0^n}\left[\int_\Theta L_n(\theta,\theta_0)\, dQ^*_{a,\gamma}(\theta|\tilde{X}_n)\right] \quad (2.77)$$

On the other hand, using similar argument as used in (2.13) to show that $Q^*_{a,\gamma}(\theta|\tilde{X}_n) \to Q^*(\theta|\tilde{X}_n)$ as $\gamma \to 0^+$, it follows that

$$\liminf_{\gamma \to 0^+} \eta_n^R(\gamma) = \eta_n^R(0).$$

Thus the result follows.

□

Next, we obtain a finite sample bound on the regret, defined as the uniform difference between the Naive VB approximate posterior risk and the expected loss under the true data generating measure P_0.

Lemma 2.7.7. *For a constant M as defined in Theorem 2.4.1*

$$\mathbb{E}_{P_0^n}\left[\sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*(\theta|\tilde{X}_n) - R(a,\theta_0)\right|\right] \leq \left[M(\epsilon_n^2 + \eta_n(0))\right]^{\frac{1}{2}}. \tag{2.78}$$

Proof. The result follows immediately from the following inequalities

$$\begin{aligned}\left(\sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*(\theta|\tilde{X}_n) - R(a,\theta_0)\right|\right)^2 &\leq \left(\int \sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|dQ^*(\theta|\tilde{X}_n)\right)^2 \\ &\leq \int\left(\sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|\right)^2 dQ^*(\theta|\tilde{X}_n),\end{aligned}$$

where the last inequality is a consequence of Jensens' inequality. Now, using Jensen's inequality again

$$\begin{aligned}&\left(\mathbb{E}_{P_0^n}\left[\sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*(\theta|\tilde{X}_n) - R(a,\theta_0)\right|\right]\right)^2 \\ &\quad\leq \mathbb{E}_{P_0^n}\left[\left(\sup_{a\in\mathcal{A}}\left|\int R(a,\theta)dQ^*(\theta|\tilde{X}_n) - R(a,\theta_0)\right|\right)^2\right].\end{aligned}$$

Now the result follows immediately using Theorem 2.4.1. □

Proof of Theorem 2.4.2 and 2.4.3 . The proof is similar to Theorem 2.3.2 and 2.3.3 and hence omitted.

□

2.7.7 Proof of Theorem 2.4.4, 2.4.5 and 2.4.6

Proof of Theorem 2.4.4: The proof follows immediately from Theorem 2.4.4 by substituting $\gamma = 1$.

□

Proof. Proof of Theorem 2.4.5 and 2.4.6: The proof is similar to Theorem 2.3.2 and 2.3.3 and hence omitted.

□

2.7.8 Newsvendor Problem

We fix $n^{-1/2}\sqrt{L_n^{NV}(\theta,\theta_0)} = (\sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|)$. Next, we aim to show that the exponentially distributed model P_θ satisfies Assumption 2.2.1, for distance function $L_n^{NV}(\theta,\theta_0)$. To show this, in the next result we first prove that $d_L^{NV}(\theta,\theta_0) = n^{-1/2}\sqrt{L_n^{NV}(\theta,\theta_0)}$ satisfy Assumption 2.3.2. Also, recall that the square of Hellinger distance between two exponential distributions with rate parameter θ and θ_0 is $d_H^2(\theta,\theta_0) = 1 - 2\frac{\sqrt{\theta\theta_0}}{\theta+\theta_0} = 1 - 2\frac{\sqrt{\theta_0/\theta}}{1+\theta_0/\theta}$.

Lemma 2.7.8. *For any $\theta \in \Theta = [T,\infty)$, and $a \in \mathcal{A}$,*

$$d_L^{NV}(\theta,\theta_0) \le \left[\frac{\left(\frac{h}{\theta_0} - \frac{h}{T}\right)^2 + (b+h)^2\left(\frac{e^{-\underline{a}T}}{T} - \frac{e^{-\underline{a}\theta_0}}{\theta_0}\right)^2}{d_H^2(T,\theta_0)}\right]^{1/2} d_H(\theta,\theta_0)$$

where $\underline{a} := \min\{a \in \mathcal{A}\}$ and $\underline{a} > 0$ and θ_0 lies in the interior of Θ.

Proof. Observe that for any $a \in \mathcal{A}$,

$$\begin{aligned}
|R(a,\theta) - R(a,\theta_0)|^2 &= \left|\frac{h}{\theta_0} - \frac{h}{\theta} + (b+h)\left(\frac{e^{-a\theta}}{\theta} - \frac{e^{-a\theta_0}}{\theta_0}\right)\right|^2 \\
&= \left(\frac{h}{\theta_0} - \frac{h}{\theta}\right)^2 + (b+h)^2\left(\frac{e^{-a\theta}}{\theta} - \frac{e^{-a\theta_0}}{\theta_0}\right)^2 + 2\left(\frac{h}{\theta_0} - \frac{h}{\theta}\right)(b+h)\left(\frac{e^{-a\theta}}{\theta} - \frac{e^{-a\theta_0}}{\theta_0}\right) \\
&\le \left(\frac{h}{\theta_0} - \frac{h}{\theta}\right)^2 + (b+h)^2\left(\frac{e^{-a\theta}}{\theta} - \frac{e^{-a\theta_0}}{\theta_0}\right)^2, \qquad (2.79)
\end{aligned}$$

where the last inequality follows since for $\theta \geq \theta_0$, $\left(\frac{h}{\theta_0} - \frac{h}{\theta}\right) \geq 0$ and $\left(\frac{\mathrm{e}^{-a\theta}}{\theta} - \frac{\mathrm{e}^{-a\theta_0}}{\theta_0}\right) < 0$ and vice versa if $\theta < \theta_0$ that together makes the last term in the penultimate equality negative for all $\theta \in \Theta$. Moreover, the first derivative of the upperbound with respect to θ is

$$2\left(\frac{h}{\theta_0} - \frac{h}{\theta}\right)\frac{h}{\theta^2} - 2(b+h)^2\left(\frac{\mathrm{e}^{-a\theta}}{\theta} - \frac{\mathrm{e}^{-a\theta_0}}{\theta_0}\right)\mathrm{e}^{-a\theta}\left[\frac{1}{\theta^2} + \frac{a}{\theta}\right],$$

and it is negative when $\theta \leq \theta_0$ and positive when $\theta > \theta_0$ for all $b > 0, h > 0$, and $a \in \mathcal{A}$. Therefore, the upperbound in (2.79) above is decreasing function of θ for all $\theta \leq \theta_0$ and increasing function of θ for all $\theta > \theta_0$. The upperbound is tight at $\theta = \theta_0$.

Now recall that the squared Hellinger distance between two exponential distributions with rate parameter θ and θ_0 is

$$d_H^2(\theta, \theta_0) = 1 - 2\frac{\sqrt{\theta\theta_0}}{\theta + \theta_0} = 1 - 2\frac{\sqrt{\theta_0/\theta}}{1 + \theta_0/\theta} = \frac{(1 - \sqrt{\theta_0/\theta})^2}{1 + (\sqrt{\theta_0/\theta})^2}.$$

Note that for $\theta \leq \theta_0$, $d_H^2(\theta, \theta_0)$ is a decreasing function of θ and for all $\theta > \theta_0$ it is an increasing function of θ. Also, note that as $\theta \to \infty$, the squared Hellinger distance as well as the upperbound computed in (2.79) converges to a constant for a given h, b, θ_0 and a. However, as $\theta \to 0$, the $d_H^2(\theta, \theta_0) \to 1$ but the upperbound computed in (2.79) diverges.

Since, $\Theta = [T, \infty)$ for some $T > 0$ and $T \leq \theta_0$, observe that if we scale $d_H^2(\theta, \theta_0)$ by factor by which the upperbound computed in (2.79) is greater than d_H at $\theta = T$, then

$$\begin{aligned}\left(\frac{h}{\theta_0} - \frac{h}{\theta}\right)^2 + (b+h)^2\left(\frac{\mathrm{e}^{-a\theta}}{\theta} - \frac{\mathrm{e}^{-a\theta_0}}{\theta_0}\right)^2 &\leq \frac{\left(\frac{h}{\theta_0} - \frac{h}{T}\right)^2 + (b+h)^2\left(\frac{\mathrm{e}^{-aT}}{T} - \frac{\mathrm{e}^{-a\theta_0}}{\theta_0}\right)^2}{d_H^2(T, \theta_0)} d_H^2(\theta, \theta_0) \\ &\leq \frac{\left(\frac{h}{\theta_0} - \frac{h}{T}\right)^2 + (b+h)^2\left(\frac{\mathrm{e}^{-\underline{a}T}}{T} - \frac{\mathrm{e}^{-\underline{a}\theta_0}}{\theta_0}\right)^2}{d_H^2(T, \theta_0)} d_H^2(\theta, \theta_0),\end{aligned} \tag{2.80}$$

where $\underline{a} = \inf\{a : a \in \mathcal{A}\}$ and in the last inequality we used the fact that $\left(\frac{\mathrm{e}^{-aT}}{T} - \frac{\mathrm{e}^{-a\theta_0}}{\theta_0}\right)^2$ is a decreasing function of a for any b, h, T, and θ_0 . Since, the RHS in the equation above

does not depend on a, it follows from the result in (2.79) and the definition of $L_n^{NV}(\theta,\theta_0)$ that $d_L^{NV}(\theta,\theta_0) \leq \left[\frac{\left(\frac{h}{\theta_0}-\frac{h}{T}\right)^2+(b+h)^2\left(\frac{e^{-aT}}{T}-\frac{e^{-a\theta_0}}{\theta_0}\right)^2}{d_H^2(T,\theta_0)}\right]^{1/2} d_H(\theta,\theta_0)$. □

Lemma 2.7.9. *For any $\theta \in \Theta = [T,\infty)$, for sufficiently small $T > 0$, and θ_0 lying in the interior of Θ, we have*

$$d_H^2(\theta,\theta_0) = 1 - 2\frac{\sqrt{\theta\theta_0}}{\theta+\theta_0} \leq \left(\frac{\theta_0}{(T+\theta_0)^2}\left(\sqrt{\frac{\theta_0}{T}}-\sqrt{\frac{T}{\theta_0}}\right)\right)|\theta-\theta_0|.$$

Proof. Observe that

$$\frac{\partial d_H^2(\theta,\theta_0)}{\partial\theta} = -2\frac{(\theta+\theta_0)\frac{\sqrt{\theta_0}}{2\sqrt{\theta}}-\sqrt{\theta\theta_0}}{(\theta+\theta_0)^2} = -\frac{(\theta+\theta_0)\sqrt{\theta_0}-2\theta\sqrt{\theta_0}}{\sqrt{\theta}(\theta+\theta_0)^2} = \frac{\theta\sqrt{\theta_0}-\theta_0\sqrt{\theta_0}}{\sqrt{\theta}(\theta+\theta_0)^2} = \frac{\theta_0}{(\theta+\theta_0)^2}\left(\sqrt{\frac{\theta}{\theta_0}}-\sqrt{\frac{\theta_0}{\theta}}\right.$$

Observe that $\theta \to 0$, $\frac{\partial d_H^2(\theta,\theta_0)}{\partial\theta} \to \infty$. Since,$\theta \in \Theta = [T,\infty)$, therefore the $\sup_{\theta\in\Theta}\left|\frac{\partial d_H^2(\theta,\theta_0)}{\partial\theta}\right| < \infty$. In fact, for sufficiently small $T > 0$, $\sup_{\theta\in\Theta}\left|\frac{\partial d_H^2(\theta,\theta_0)}{\partial\theta}\right| = \left|\frac{\theta_0}{(T+\theta_0)^2}\left(\sqrt{\frac{T}{\theta_0}}-\sqrt{\frac{\theta_0}{T}}\right)\right| = \left(\frac{\theta_0}{(T+\theta_0)^2}\left(\sqrt{\frac{\theta_0}{T}}-\sqrt{\frac{T}{\theta_0}}\right)\right)$. Now the result follows immediately since the derivative of $d_H^2(\theta,\theta_0)$ is bounded on Θ, which implies that $d_H^2(\theta,\theta_0)$ is Lipschitz on Θ. □

Lemma 2.7.10. *For any $\theta \in \Theta = [T,\infty)$, and $a \in \mathcal{A}$,*

$$d_L^{NV}(\theta,\theta_0) \leq \frac{h}{T^2}|\theta-\theta_0|.$$

Proof. Recall,

$$R(a,\theta) = ha - \frac{h}{\theta} + (b+h)\frac{e^{-a\theta}}{\theta}.$$

First, observe that for any $a \in \mathcal{A}$,

$$\frac{\partial R(a,\theta)}{\partial\theta} = \frac{h}{\theta^2} - a(b+h)\frac{e^{-a\theta}}{\theta} - (b+h)\frac{e^{-a\theta}}{\theta^2} = \frac{1}{\theta^2}\left(h-(b+h)e^{-a\theta}(1+a\theta)\right) \leq \frac{h}{\theta^2}. \quad (2.81)$$

The result follows immediately, since $\sup_{\theta\in\Theta}\frac{\partial R(a,\theta)}{\partial\theta} \leq \frac{h}{T^2}$. □

Proof. Proof of Lemma 2.5.1

It follows from Lemma 2.7.8 that $d_L^{NV}(\theta,\theta_0)$ for any $\theta \in \Theta = [T,\infty)$ and θ_0 lying the interior of Θ, satisfies Assumption 2.3.2 with $K_1 = \left[\frac{\left(\frac{h}{\theta_0}-\frac{h}{T}\right)^2+(b+h)^2\left(\frac{e^{-\underline{a}T}}{T}-\frac{e^{-\underline{a}\theta_0}}{\theta_0}\right)^2}{d_H^2(T,\theta_0)}\right]^{1/2} := K_1^{NV}$. Similarly, it follows from Lemma and 2.7.10 that for sufficiently small $T>0$, $d_L^{NV}(\theta,\theta_0)$ satisfies Assumption 2.3.3 with $K_2 = h/T^2 := K_2^{NV}$. Now using similar arguments as used in Lemma 2.3.2 and Lemma 2.2.1, for a given $\epsilon_n > 0$ and every $\epsilon > \epsilon_n$, such that $n\epsilon_n^2 \geq 1$, it can be shown that , $L_n^{NV}(\theta,\theta_0) = n\left(\sup_{a\in\mathcal{A}}|R(a,\theta)-R(a,\theta_0)|\right)^2$ satisfies

$$\mathbb{E}_{P_0^n}[\phi_n] \leq C_0 \exp(-Cn\epsilon^2), \tag{2.82}$$

$$\sup_{\{\theta\in\Theta: L_n^{NV}(\theta,\theta_0)\geq C_1 n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_n] \leq \exp(-Cn\epsilon^2), \tag{2.83}$$

where $C_0 = 20$ and $C = \frac{C_1}{2(K_1^{NV})^2}$ for a constant $C_1 > 0$. □

Proof. Proof of Proposition 2.5.2:

First, we write the Rényi divergence between P_0^n and P_θ^n,

$$D_{1+\lambda}\left(P_0^n\|P_\theta^n\right) = \frac{1}{\lambda}\log\int\left(\frac{dP_0^n}{dP_\theta^n}\right)^\lambda dP_0^n = n\frac{1}{\lambda}\log\int\left(\frac{dP_0}{dP_\theta}\right)^\lambda dP_0 = n\left(\log\frac{\theta_0}{\theta} + \frac{1}{\lambda}\log\frac{\theta_0}{(\lambda+1)\theta_0-\lambda\theta}\right),$$

when $((\lambda+1)\theta_0-\lambda\theta) > 0$ and $D_{1+\lambda}\left(P_0^n\|P_\theta^n\right) = \infty$ otherwise. Also, observe that, $D_{1+\lambda}\left(P_0^n\|P_\theta^n\right)$ is non-decreasing in λ (this also follows from non-decreasing property of the Rényi divergence with respect to λ). Therefore, observe that

$$\Pi(D_{1+\lambda}\left(P_0^n\|P_\theta^n\right) \leq C_3 n\epsilon_n^2) \geq \Pi(D_\infty\left(P_0^n\|P_\theta^n\right) \leq C_3 n\epsilon_n^2) = \Pi\left(0 \leq \log\frac{\theta_0}{\theta} \leq C_3\epsilon_n^2\right) = \Pi\left(\theta_0 e^{-C_3\epsilon_n^2} \leq \theta \leq \theta_0\right)$$

Now, recall that for a set $A \subseteq \Theta = [T,\infty)$, we define $\Pi(A) = \text{Inv}-\Gamma(A\cap\Theta)/\text{Inv}-\Gamma(\Theta)$. Now, observe that for sufficiently small T and large enough n, we have

$$\Pi\left(\theta_0 e^{-C_3\epsilon_n^2} \leq \theta \leq \theta_0\right) \geq \text{Inv}-\Gamma\left(\theta_0 e^{-C_3\epsilon_n^2} \leq \theta \leq \theta_0\right)$$

The cumulative distribution function of inverse-gamma distribution is $\text{Inv}-\Gamma(\{\theta\in\Theta:\theta < t\}) := \frac{\Gamma\left(\alpha,\frac{\beta}{t}\right)}{\Gamma(\alpha)}$, where $\alpha(>0)$ is the shape parameter, $\beta(>0)$ is the scale parameter, $\Gamma(\cdot)$ is

the Gamma function, and $\Gamma(\cdot,\cdot)$ is the incomplete Gamma function. Therefore, it follows for $\alpha > 1$ that

$$\begin{aligned}\text{Inv}-\Gamma\left(\theta_0 \mathrm{e}^{-C_3\epsilon_n^2} \le \theta \le \theta_0\right) &= \frac{\Gamma\left(\alpha, \beta/\theta_0\right) - \Gamma\left(\alpha, \beta/\theta_0 \mathrm{e}^{C_3\epsilon_n^2}\right)}{\Gamma(\alpha)} = \frac{\int_{\beta/\theta_0}^{\beta/\theta_0 \mathrm{e}^{C_3\epsilon_n^2}} \mathrm{e}^{-x} x^{\alpha-1} dx}{\Gamma(\alpha)} \\ &\ge \frac{\mathrm{e}^{-\beta/\theta_0 \mathrm{e}^{C_3\epsilon_n^2} + \alpha C_3\epsilon_n^2}}{\alpha\Gamma(\alpha)} \left(\frac{\beta}{\theta_0}\right)^{\alpha} \left[1 - \mathrm{e}^{-\alpha C_3\epsilon_n^2}\right] \\ &\ge \frac{\mathrm{e}^{-\beta/\theta_0 \mathrm{e}^{C_3}}}{\alpha\Gamma(\alpha)} \left(\frac{\beta}{\theta_0}\right)^{\alpha} \left[\mathrm{e}^{-\alpha C_3 n\epsilon_n^2}\right]\end{aligned}$$

where the penultimate inequality folows since $0 < \epsilon_n^2 < 1$ and the last inequality follows from the fact that, $1-\mathrm{e}^{-\alpha C_3\epsilon_n^2} \ge \mathrm{e}^{-\alpha C_3 n\epsilon_n^2}$, for large enough n. Also note that, $1-\mathrm{e}^{-\alpha C_3\epsilon_n^2} \ge \mathrm{e}^{-\alpha C_3 n\epsilon_n^2}$ can't hold true for $\epsilon_n^2 = 1/n$. However, for $\epsilon_n^2 = \frac{\log n}{n}$ it holds for any $n \ge 2$ when $\alpha C_3 > 2$. Therefore, for inverse-Gamma prior restricted to Θ, $C_2 = \alpha C_3$ and any $\lambda > 1$ the result follows for sufficiently large n.

□

Proof. Proof of Proposition 2.5.3: Recall,

$$R(a,\theta) = ha - \frac{h}{\theta} + (b+h)\frac{\mathrm{e}^{-a\theta}}{\theta}.$$

First, observe that for any $a \in \mathcal{A}$,

$$\frac{\partial R(a,\theta)}{\partial\theta} = \frac{h}{\theta^2} - a(b+h)\frac{\mathrm{e}^{-a\theta}}{\theta} - (b+h)\frac{\mathrm{e}^{-a\theta}}{\theta^2} = \frac{1}{\theta^2}\left(h - (b+h)\mathrm{e}^{-a\theta}(1+a\theta)\right). \tag{2.84}$$

Using the above equation the (finite) critical point θ^* must satisfy, $h - (b+h)\mathrm{e}^{-a\theta^*}(1 + a\theta^*) = 0$. Therefore,

$$R(a,\theta) \ge R(a,\theta^*) = h\left(a - \frac{1}{\theta^*} + \frac{1}{\theta^*(1+a\theta^*)}\right) = \frac{ha^2\theta^*}{(1+a\theta^*)}.$$

Since $h, b > 0$ and $a\theta^* > 0$, hence

$$R(a,\theta) \ge \frac{h\underline{a}^2\theta^*}{(1+a\theta^*)},$$

where $\underline{a} := \min\{a \in \mathcal{A}\}$ and $\underline{a} > 0$.

□

Proof. Proof of Proposition 2.5.4:

First, observe that $R(a,\theta)$ is bounded above in θ for a given $a \in \mathcal{A}$

$$\begin{aligned} R(a,\theta) &= ha - \frac{h}{\theta} + (b+h)\frac{\mathrm{e}^{-a\theta}}{\theta} \\ &\leq ha + \frac{b}{\theta}. \end{aligned}$$

Using the above fact and the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \int_{\left\{\mathrm{e}^{\gamma R(a,\theta)} > \mathrm{e}^{C_4(\gamma) n\epsilon_n^2}\right\}} \mathrm{e}^{\gamma R(a,\theta)} \boldsymbol{\pi}(\theta) d\theta &\leq \left(\int \mathrm{e}^{2\gamma R(a,\theta)} \boldsymbol{\pi}(\theta) d\theta\right)^{1/2} \left(\int \mathbb{1}_{\mathrm{e}^{\gamma R(a,\theta)} > \mathrm{e}^{C_4(\gamma) n\epsilon_n^2}} \boldsymbol{\pi}(\theta) d\theta\right)^{1/2} \\ &\leq \left(\int \mathrm{e}^{2\gamma\left(ha + \frac{b}{\theta}\right)} \boldsymbol{\pi}(\theta) d\theta\right)^{1/2} \left(\int \mathbb{1}_{\left\{\mathrm{e}^{\gamma\left(ha+\frac{b}{\theta}\right)} > \mathrm{e}^{C_4(\gamma) n\epsilon_n^2}\right\}} \boldsymbol{\pi}(\theta) d\theta\right)^{1/2} \\ &\leq \mathrm{e}^{-C_4(\gamma) n\epsilon_n^2} \left(\int \mathrm{e}^{2\gamma\left(ha + \frac{b}{\theta}\right)} \boldsymbol{\pi}(\theta) d\theta\right), \end{aligned} \tag{2.85}$$

where the last inequality follows from using the Chebyshev's inequality.

Now using the definition of the prior distribution, which is an inverse gamma prior restricted to $\Theta = [T, \infty)$, we have

$$\begin{aligned} \int_{\left\{\mathrm{e}^{\gamma R(a,\theta)} > \mathrm{e}^{C_4(\gamma) n\epsilon_n^2}\right\}} \mathrm{e}^{\gamma R(a,\theta)} \boldsymbol{\pi}(\theta) d\theta &\leq \mathrm{e}^{-C_4(\gamma) n\epsilon_n^2} \left(\int \mathrm{e}^{2\gamma\left(ha + \frac{b}{\theta}\right)} \boldsymbol{\pi}(\theta) d\theta\right) \\ &\leq \mathrm{e}^{-C_4(\gamma) n\epsilon_n^2} \mathrm{e}^{2\gamma\left(h\overline{a} + \frac{b}{T}\right)}, \end{aligned}$$

where $\overline{a} := \max\{a \in \mathcal{A}\}$ and $\overline{a} > 0$. Since $n\epsilon_n^2 \geq 1$, we must fix $C_4(\gamma)$ such that $\mathrm{e}^{C_4(\gamma)} > \mathrm{e}^{2\gamma\left(h\overline{a} + \frac{b}{T}\right)}$, that is $C_4(\gamma) > 2\gamma\left(h\overline{a} + \frac{b}{T}\right)$ and $C_5(\gamma) = C_4(\gamma) - 2\gamma\left(h\overline{a} + \frac{b}{T}\right)$.

□

Proof. Proof of Proposition 2.5.5: Since family $\mathcal{Q}$ contains all shifted-gamma distributions, observe that $\{q_n(\cdot) \in \mathcal{Q}\} \forall n \geq 1$. By definition, $q_n(\theta) = \frac{n^n}{\theta_0^n \Gamma(n)}(\theta - T)^{n-1} \mathrm{e}^{-n\frac{(\theta-T)}{\theta_0}}$. Now consider the first term; using the definition of the KL divergence it follows that

$$\mathrm{KL}(q_n(\theta) \| \boldsymbol{\pi}(\theta)) = \int_T^\infty q_n(\theta) \log(q_n(\theta)) d\theta - \int_T^\infty q_n(\theta) \log(\boldsymbol{\pi}(\theta)) d\theta. \tag{2.86}$$

Substituting $q_n(\theta)$ in the first term of the equation above and expanding the logarithm term, we obtain

$$\begin{aligned}\int_T^\infty q_n(\theta) \log(q_n(\theta)) d\theta &= (n-1) \int_T^\infty \log(\theta - T) \frac{n^n}{\theta_0^n \Gamma(n)} (\theta - T)^{n-1} \mathrm{e}^{-n\frac{\theta-T}{\theta_0}} d\theta - n + \log\left(\frac{n^n}{\theta_0^n \Gamma(n)}\right) \\ &= -\log \theta_0 + (n-1) \int_T^\infty \log \frac{\theta - T}{\theta_0} \frac{n^n}{\theta_0^n \Gamma(n)} (\theta - T)^{n-1} \mathrm{e}^{-n\frac{\theta-T}{\theta_0}} d\theta - n + \log\left(\frac{n^n}{\Gamma(n)}\right)\end{aligned} \tag{2.87}$$

Now consider the second term in the equation above. Substitute $\theta = \frac{t\theta_0}{n} + T$ into the integral, we have

$$\begin{aligned}\int_T^\infty \log \frac{\theta - T}{\theta_0} \frac{n^n}{\theta_0^n \Gamma(n)} (\theta - T)^{n-1} \mathrm{e}^{-n\frac{\theta-T}{\theta_0}} d\theta &= \int_0^\infty \log \frac{t}{n} \frac{1}{\Gamma(n)} t^{n-1} \mathrm{e}^{-t} dt \\ &\leq \int \left(\frac{t}{n} - 1\right) \frac{1}{\Gamma(n)} t^{n-1} \mathrm{e}^{-t} dt = 0.\end{aligned} \tag{2.88}$$

Substituting the above result into (2.87), we get

$$\begin{aligned}\int_T^\infty q_n(\theta) \log(q_n(\theta)) d\theta &\leq -\log \theta_0 - n + \log\left(\frac{n^n}{\Gamma(n)}\right) \\ &\leq -\log \theta_0 - n + \log\left(\frac{n^n}{\sqrt{2\pi n} n^{n-1} \mathrm{e}^{-n}}\right) \\ &= -\log \sqrt{2\pi}\theta_0 + \frac{1}{2} \log n,\end{aligned} \tag{2.89}$$

where the second inequality uses the fact that $\sqrt{2\pi n}n^n\mathrm{e}^{-n} \leq n\Gamma(n)$. Recall $\boldsymbol{\pi}(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-\alpha-1}\mathrm{e}^{-\frac{\beta}{\theta}}$. Now consider the second term in (2.86). Using the definition of inverse-gamma prior and expanding the logarithm function, we have

$$\begin{aligned}
-\int_T^\infty q_n(\theta)\log(\boldsymbol{\pi}(\theta))d\theta &= -\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + (\alpha+1)\int_T^\infty \log\theta \frac{n^n}{\theta_0^n\Gamma(n)}(\theta-T)^{n-1}\mathrm{e}^{-n\frac{\theta-T}{\theta_0}}d\theta + \beta\frac{n}{(n-1)\theta_0} \\
&= -\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \int_T^\infty \log\frac{\theta}{\theta_0}\frac{n^n}{\theta_0^n\Gamma(n)}(\theta-T)^{n-1}\mathrm{e}^{-n\frac{\theta-T}{\theta_0}}d\theta \\
&\quad + \beta\frac{n}{(n-1)\theta_0} + (\alpha+1)\log\theta_0 \\
&\leq -\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \int_T^\infty \frac{\theta-T}{\theta_0}\frac{n^n}{\theta_0^n\Gamma(n)}(\theta-T)^{n-1}\mathrm{e}^{-n\frac{\theta-T}{\theta_0}}d\theta \\
&\quad + \beta\frac{n}{(n-1)\theta_0} + (\alpha+1)\log\theta_0 \\
&= -\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \beta\frac{n}{(n-1)\theta_0} + (\alpha+1)\log\theta_0, \qquad (2.90)
\end{aligned}$$

where the first inequlity is due to fact that $\mathbb{E}_{q_n}[\beta/\theta] \leq \mathbb{E}_{q_n}[\beta/(\theta-T)]$ for any $\theta > T$ and the penultimate inequality follows from the observation in (2.88) and the fact that $\log\frac{\theta}{\theta_0} \leq \frac{\theta}{\theta_0} - 1 \leq \frac{\theta}{\theta_0} - \frac{T}{\theta_0}$ for any $\theta_0 > T$. Substituting (2.90) and (2.89) into (2.86) and dividing either sides by n, we obtain

$$\begin{aligned}
\frac{1}{n}\mathrm{KL}(q_n(\theta)\|\boldsymbol{\pi}(\theta)) &\leq \frac{1}{n}\left(-\log\sqrt{2\pi}\theta_0 + \frac{1}{2}\log n - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \beta\frac{n}{(n-1)\theta_0} + (\alpha+1)\log\theta_0\right) \\
&= \frac{1}{2}\frac{\log n}{n} + \beta\frac{1}{(n-1)\theta_0} + \frac{1}{n}\left(-\log\sqrt{2\pi} - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + (\alpha)\log\theta_0\right). \qquad (2.91)
\end{aligned}$$

Now, consider the second term in the assertion of the lemma. Since $\xi_{\mathrm{i}}, \mathrm{i} \in \{1, 2\ldots n\}$ are independent and identically distributed, we obtain

$$\frac{1}{n}\mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n\|p(\tilde{X}_n|\theta)\right)\right] = \mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0\|p(\xi|\theta)\right)\right]$$

Now using the expression for KL divergence between the two exponential distributions, we have

$$\frac{1}{n}\mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n \| p(\tilde{X}_n|\theta)\right)\right] = \int_T^\infty \left(\log\frac{\theta_0}{\theta} + \frac{\theta}{\theta_0} - 1\right)\frac{n^n}{\theta_0^n\Gamma(n)}(\theta - T)^{n-1}\mathrm{e}^{-n\frac{\theta-T}{\theta_0}}d\theta \leq \frac{n}{n-1} + 1 - 2 = \frac{1}{n-} \tag{2.92}$$

where second inequality uses the fact that $\log x \leq x - 1 \leq x - \frac{T}{\theta_0}$ for $\theta_0 > T$. Combined together (2.92) and (2.91) for $n \geq 2$ implies that

$$\begin{aligned}&\frac{1}{n}\left[\mathrm{KL}\left(q_n(\theta)\|\boldsymbol{\pi}(\theta)\right) + \mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n)\|p(\tilde{X}_n|\theta)\right)\right]\right] \\ &\quad\leq \frac{1}{2}\frac{\log n}{n} + \frac{1}{n}\left(2 + \frac{2\beta}{\theta_0} - \log\sqrt{2\boldsymbol{\pi}} - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \alpha\log\theta_0\right) \leq C_9\frac{\log n}{n}.\end{aligned} \tag{2.93}$$

where $C_9 := \frac{1}{2} + \max\left(0, 2 + \frac{2\beta}{\theta_0} - \log\sqrt{2\boldsymbol{\pi}} - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \alpha\log\theta_0\right)$ and the result follows. □

Proof. Proof of Proposition 2.5.5: Since family $\mathcal{Q}$ contains all gamma distributions, observe that $\{q_n(\cdot) \in \mathcal{Q}\}\forall n \geq 1$. By definition, $q_n(\theta) = \frac{n^n}{\theta_0^n\Gamma(n)}\theta^{n-1}\mathrm{e}^{-n\frac{\theta}{\theta_0}}$. Now consider the first term; using the definition of the KL divergence it follows that

$$\mathrm{KL}(q_n(\theta)\|\boldsymbol{\pi}(\theta)) = \int q_n(\theta)\log(q_n(\theta))d\theta - \int q_n(\theta)\log(\boldsymbol{\pi}(\theta))d\theta. \tag{2.94}$$

Substituting $q_n(\theta)$ in the first term of the equation above and expanding the logarithm term, we obtain

$$\begin{aligned}\int q_n(\theta)\log(q_n(\theta))d\theta &= (n-1)\int \log\theta\frac{n^n}{\theta_0^n\Gamma(n)}\theta^{n-1}\mathrm{e}^{-n\frac{\theta}{\theta_0}}d\theta - n + \log\left(\frac{n^n}{\theta_0^n\Gamma(n)}\right) \\ &= -\log\theta_0 + (n-1)\int \log\frac{\theta}{\theta_0}\frac{n^n}{\theta_0^n\Gamma(n)}\theta^{n-1}\mathrm{e}^{-n\frac{\theta}{\theta_0}}d\theta - n + \log\left(\frac{n^n}{\Gamma(n)}\right)\end{aligned} \tag{2.95}$$

Now consider the second term in the equation above. Substitute $\theta = \frac{t\theta_0}{n}$ into the integral, we have

$$\begin{aligned}\int \log \frac{\theta}{\theta_0} \frac{n^n}{\theta_0^n \Gamma(n)} \theta^{n-1} \mathrm{e}^{-n\frac{\theta}{\theta_0}} d\theta &= \int \log \frac{t}{n} \frac{1}{\Gamma(n)} t^{n-1} \mathrm{e}^{-t} dt \\ &\leq \int \left(\frac{t}{n} - 1\right) \frac{1}{\Gamma(n)} t^{n-1} \mathrm{e}^{-t} dt = 0. \end{aligned} \tag{2.96}$$

Substituting the above result into (2.95), we get

$$\begin{aligned}\int q_n(\theta) \log(q_n(\theta)) d\theta &\leq -\log \theta_0 - n + \log\left(\frac{n^n}{\Gamma(n)}\right) \\ &\leq -\log \theta_0 - n + \log\left(\frac{n^n}{\sqrt{2\pi n} n^{n-1} \mathrm{e}^{-n}}\right) \\ &= -\log \sqrt{2\pi}\theta_0 + \frac{1}{2}\log n, \end{aligned} \tag{2.97}$$

where the second inequality uses the fact that $\sqrt{2\pi n} n^n \mathrm{e}^{-n} \leq n\Gamma(n)$. Recall $\boldsymbol{\pi}(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-\alpha-1} \mathrm{e}^{-\frac{\beta}{\theta}}$. Now consider the second term in (2.94). Using the definition of inverse-gamma prior and expanding the logarithm function, we have

$$\begin{aligned}-\int q_n(\theta) \log(\boldsymbol{\pi}(\theta)) d\theta &= -\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + (\alpha+1) \int \log\theta \frac{n^n}{\theta_0^n \Gamma(n)} \theta^{n-1} \mathrm{e}^{-n\frac{\theta}{\theta_0}} d\theta + \beta \frac{n}{(n-1)\theta_0} \\ &= -\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + (\alpha+1) \int \log\frac{\theta}{\theta_0} \frac{n^n}{\theta_0^n \Gamma(n)} \theta^{n-1} \mathrm{e}^{-n\frac{\theta}{\theta_0}} d\theta \\ &\quad + \beta \frac{n}{(n-1)\theta_0} + (\alpha+1)\log\theta_0 \\ &\leq -\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \beta \frac{n}{(n-1)\theta_0} + (\alpha+1)\log\theta_0, \end{aligned} \tag{2.98}$$

where the last inequality follows from the observation in (2.96). Substituting (2.98) and (2.97) into (2.94) and dividing either sides by n, we obtain

$$\begin{aligned}\frac{1}{n}\mathrm{KL}(q_n(\theta) \| \boldsymbol{\pi}(\theta)) &\leq \frac{1}{n}\left(-\log\sqrt{2\pi}\theta_0 + \frac{1}{2}\log n - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \beta\frac{n}{(n-1)\theta_0} + (\alpha+1)\log\theta_0\right) \\ &= \frac{1}{2}\frac{\log n}{n} + \beta\frac{1}{(n-1)\theta_0} + \frac{1}{n}\left(-\log\sqrt{2\pi} - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + (\alpha)\log\theta_0\right). \end{aligned} \tag{2.99}$$

Now, consider the second term in the assertion of the lemma. Since, $\xi_{\mathrm{i}}, \mathrm{i} \in \{1, 2 \dots n\}$ are independent and identically distributed, we obtain

$$\frac{1}{n}\mathbb{E}_{q(\theta)}\left[\mathrm{KL}\left(dP_0^n \| p(\tilde{X}_n|\theta)\right)\right] = \mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0 \| p(\xi|\theta)\right)\right]$$

Now using the expression for KL divergence between the two exponential distributions, we have

$$\frac{1}{n}\mathbb{E}_{q(\theta)}\left[\mathrm{KL}\left(dP_0^n \| p(\tilde{X}_n|\theta)\right)\right] = \int \left(\log\frac{\theta_0}{\theta} + \frac{\theta}{\theta_0} - 1\right)\frac{n^n}{\theta_0^n \Gamma(n)}\theta^{n-1}\mathrm{e}^{-n\frac{\theta}{\theta_0}}d\theta \leq \frac{n}{n-1} + 1 - 2 = \frac{1}{n-1}, \tag{2.100}$$

where second inequality uses the fact that $\log x \leq x - 1$. Combined together (2.100) and (2.99) for $n \geq 2$ implies that

$$\begin{aligned}&\frac{1}{n}\left[\mathrm{KL}\left(q(\theta)\|\boldsymbol{\pi}(\theta)\right) + \mathbb{E}_{q(\theta)}\left[\mathrm{KL}\left(dP_0^n)\|p(\tilde{X}_n|\theta)\right)\right]\right] \\ &\quad \leq \frac{1}{2}\frac{\log n}{n} + \frac{1}{n}\left(2 + \frac{2\beta}{\theta_0} - \log\sqrt{2\boldsymbol{\pi}} - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \alpha\log\theta_0\right) \leq C_9\frac{\log n}{n}.\end{aligned} \tag{2.101}$$

where $C_9 := \frac{1}{2} + \max\left(0, 2 + \frac{2\beta}{\theta_0} - \log\sqrt{2\boldsymbol{\pi}} - \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right) + \alpha\log\theta_0\right)$ and the result follows. □

2.7.9 Multi-product Newsvendor problem

In the multi-dimensional newsvendor problem, we fix $n^{-1/2}\sqrt{L_n^{MNV}(\theta, \theta_0)} = (\sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|)$, where

$$R(a,\theta) = \sum_{\mathrm{i}=1}^{d}\left[(h_{\mathrm{i}} + b_{\mathrm{i}})a_{\mathrm{i}}\Phi(a_{\mathrm{i}}) - b_{\mathrm{i}}a_{\mathrm{i}} + \theta_{\mathrm{i}}(b_{\mathrm{i}} - h_{\mathrm{i}}) + \sigma_{\mathrm{ii}}\left[h\frac{\phi((a_{\mathrm{i}} - \theta_{\mathrm{i}})/\sigma_{\mathrm{ii}})}{\Phi((a_{\mathrm{i}} - \theta_{\mathrm{i}})/\sigma_{\mathrm{ii}})} + b\frac{\phi((a_{\mathrm{i}} - \theta_{\mathrm{i}})/\sigma_{\mathrm{ii}})}{1 - \Phi((a_{\mathrm{i}} - \theta_{\mathrm{i}})/\sigma_{\mathrm{ii}})}\right]\right].$$

For brevity, we denote $d_L^{MNV}(\theta, \theta_0) = n^{-1/2}\sqrt{L_n^{MNV}(\theta, \theta_0)}$. First, we show that

Lemma 2.7.11. *For any compact decision space $\mathcal{A}$ and compact model space Θ,*

$$d_L^{MNV}(\theta, \theta_0) \leq K\|\theta - \theta_0\|,$$

for a constant K depending on compact sets $\mathcal{A}$ and Θ and given b, h and Σ.

Proof. Observe that

$$\begin{aligned}\partial_{\theta_i} R(a,\theta) &= (b_i - h_i) + (a_i-\theta_i)/\sigma_{ii}\phi((a_i-\theta_i)/\sigma_{ii})\left[\frac{h}{\Phi((a_i-\theta_i)/\sigma_{ii})} + \frac{b}{1-\Phi((a_i-\theta_i)/\sigma_{ii})}\right] \\ &+ \sigma_{ii}\phi\left(\frac{(a_i-\theta_i)}{\sigma_{ii}}\right)\left[\frac{h\phi((a_i-\theta_i)/\sigma_{ii})}{\sigma_{ii}\Phi((a_i-\theta_i)/\sigma_{ii})^2} - \frac{b\phi((a_i-\theta_i)/\sigma_{ii})}{\sigma_{ii}(1-\Phi((a_i-\theta_i)/\sigma_{ii}))^2}\right] \\ &= (b_i - h_i) + (a_i-\theta_i)/\sigma_{ii}\phi((a_i-\theta_i)/\sigma_{ii})\left[\frac{h}{\Phi((a_i-\theta_i)/\sigma_{ii})} + \frac{b}{1-\Phi((a_i-\theta_i)/\sigma_{ii})}\right] \\ &+ \phi\left(\frac{(a_i-\theta_i)}{\sigma_{ii}}\right)\left[\frac{h\phi((a_i-\theta_i)/\sigma_{ii})}{\Phi((a_i-\theta_i)/\sigma_{ii})^2} - \frac{b\phi((a_i-\theta_i)/\sigma_{ii})}{(1-\Phi((a_i-\theta_i)/\sigma_{ii}))^2}\right]. \end{aligned} \tag{2.102}$$

Since, $\mathcal{A}$ and Θ are compact sets, therefore $\{(a_i-\theta_i)/\sigma_{ii}\}_{i=1}^d$ lie in a compact set. Consequently, $\phi((a_i-\theta_i)/\sigma_{ii})$ and $\Phi((a_i-\theta_i)/\sigma_{ii})$ also lie in bounded subset of $\mathbb{R}$ and thus $\sup_{\mathcal{A},\Theta}\|\partial_{\theta_i} R(a,\theta)\| \leq K$ for a given b, h and Σ. Since , the norm of the derivative of $R(a,\theta)$ is bounded on Θ for any $a \in \mathcal{A}$, therefore, $d_L^{MNV}(\theta,\theta_0)$ is uniformly Lipschitz in $\mathcal{A}$ with Lipschitz constant K, that is

$$d_L^{MNV}(\theta,\theta_0) \leq K\|\theta-\theta_0\|.$$

□

Next, we show that the P_θ satisfies Assumption 2.2.1, for distance function $L_n^{MNV}(\theta,\theta_0)$.

Proof. Proof of Lemma 2.5.6:

First consider the following test function, constructed using $\tilde{X}_n = \{\xi_1, \xi_2, \ldots, \xi_n\}$.

$$\phi_{n,\epsilon} := \mathbb{1}_{\left\{\tilde{X}_n : \left\|\hat{\theta}_n - \theta_0\right\| > \sqrt{C\epsilon^2}\right\}},$$

where $\hat{\theta}_n = \frac{\sum_{i=1}^n \xi_i}{n}$. Note that $\hat{\theta}_n - \theta_0 \sim \mathcal{N}(\cdot|0, \frac{1}{n}\Sigma)$, where $\frac{1}{n}\Sigma$ is a symmetric positive definite matrix. Therefore it can be decomposed as $\Sigma = Q^T\Lambda Q$, where Q is an orthogonal matrix and Λ is a daigonal matrix consisting of respective eigen values and consequently $\hat{\theta}_n - \theta_0 \sim Q\mathcal{N}(\cdot|0, \frac{1}{n}\Lambda)$. So, we have $\|\hat{\theta}_n - \theta_0\|^2 \sim \|\mathcal{N}(\cdot|0, \frac{1}{n}\Lambda)\|^2$. Notice that $\|\mathcal{N}(\cdot|0, \frac{1}{n}\Lambda)\|^2$ is a linear combination of d $\chi^2_{(1)}$ random variable weighted by elements of the diagonal matrix

$\frac{1}{n}\Lambda$. Using this observation, we first verify that $\phi_{n,\epsilon}$ satisfies condition (i) of the Lemma. Observe that

$$\mathbb{E}_{P_0^n}[\phi_n] = P_0^n\left(\tilde{X}_n : \left\|\hat{\theta}_n - \theta_0\right\|^2 > C\epsilon^2\right) = P_0^n\left(\tilde{X}_n : \|\mathcal{N}(\cdot|0,\Lambda)\|^2 > Cn\epsilon^2\right).$$

Note that $\chi^2_{(1)}$ is Γ distributed with shape $1/2$ and scale 2, which implies $\chi^2_{(1)} - 1$ is a sub-gamma random variable with scale factor 2 and variance factor 2. Now observe that for $\hat{\Lambda} = \max_{\mathrm{i}\in\{1,2,\ldots d\}} \Lambda_{\mathrm{ii}}$,

$$\begin{aligned}
P_0^n\left(\tilde{X}_n : \|\mathcal{N}(\cdot|0,\Lambda)\|^2 > Cn\epsilon^2\right) \leq P_0^n\left(\tilde{X}_n : \chi^2_{(1)} > \frac{1}{d\hat{\Lambda}}Cn\epsilon^2\right) &\leq P_0^n\left(\tilde{X}_n : \chi^2_{(1)} > \frac{1}{d\hat{\Lambda}}Cn\epsilon^2\right) \\
&= P_0^n\left(\tilde{X}_n : \chi^2_{(1)} - 1 > \frac{1}{d\hat{\Lambda}}Cn\epsilon^2 - 1\right) \\
&\leq \mathrm{e}^{-\frac{\left(\frac{1}{d\hat{\Lambda}}Cn\epsilon^2 - 1\right)^2}{2\left(2+2\left(\frac{1}{d\hat{\Lambda}}Cn\epsilon^2-1\right)\right)}} \\
&\leq \mathrm{e}^{-1/8\frac{1}{d\hat{\Lambda}}Cn\epsilon^2 + 1/8} \leq \mathrm{e}^{-1/8\left(\frac{C}{d\hat{\Lambda}}-1\right)n\epsilon^2},
\end{aligned} \tag{2.103}$$

where in the third inequality we used the well known tail bound for sub-gamma random variable (Lemma 3.12 [97]) assuming that C is sufficiently large such that $\left(\frac{1}{d\hat{\Lambda}}Cn\epsilon^2 - 1\right) > 1$ and in the last inequality follows from the assumption that $n\epsilon^2 > n\epsilon_n^2 \geq 1$.

Now, we fix the alternate set to be $\{\theta \in \mathbb{R}^d : \|\theta - \theta_0\| \geq 2\sqrt{C\epsilon^2}\}$. Next, we verify that $\phi_{n,\epsilon}$ satisfies condition (ii) of the lemma. First, observe that

$$\mathbb{E}_{P_\theta^n}[1-\phi_n] = P_\theta^n\left(\tilde{X}_n : \left\|\hat{\theta}_n - \theta_0\right\|^2 \leq C\epsilon^2\right) \leq P_\theta^n\left(\tilde{X}_n : \|\hat{\theta}_n - \theta\| \geq \|\theta - \theta_0\| - \sqrt{C\epsilon^2}\right), \tag{2.104}$$

where in the last inequality, we used the fact that $\|\theta-\theta_0\| \leq \|\hat{\theta}_n-\theta\| + \left\|\hat{\theta}_n-\theta_0\right\|$. Now on alternate set $\{\theta \in \mathbb{R}^d : \|\theta-\theta_0\| \geq 2\sqrt{C\epsilon^2}\}$,

$$\begin{aligned}\mathbb{E}_{P_\theta^n}[1-\phi_n] &\leq P_\theta^n\left(\tilde{X}_n : \|\hat{\theta}_n-\theta\| \geq \|\theta-\theta_0\| - \sqrt{C\epsilon^2}\right)\\ &\leq P_\theta^n\left(\tilde{X}_n : \|\hat{\theta}_n-\theta\| \geq \|\theta-\theta_0\| - \sqrt{C\epsilon^2}\right)\\ &\leq P_\theta^n\left(\tilde{X}_n : \|\hat{\theta}_n-\theta\| \geq \sqrt{C\epsilon^2}\right). \end{aligned} \tag{2.105}$$

Now, it follows from (2.103) and $\Theta \subset \mathbb{R}^d$ that

$$\mathbb{E}_{P_0^n}[\phi_n] \leq \mathrm{e}^{-1/8\left(\frac{C}{d\hat{\Lambda}}-1\right)n\epsilon^2}, \tag{2.106}$$

$$\sup_{\{\theta\in\Theta:\|\theta-\theta_0\|\geq 2\sqrt{C\epsilon^2}\}} \mathbb{E}_{P_\theta^n}[1-\phi_n] \leq \sup_{\{\theta\in\mathbb{R}^d:\|\theta-\theta_0\|\geq 2\sqrt{C\epsilon^2}\}} \mathbb{E}_{P_\theta^n}[1-\phi_n] \leq \mathrm{e}^{-1/8\left(\frac{C}{d\hat{\Lambda}}-1\right)n\epsilon^2}. \tag{2.107}$$

Using Lemma 2.7.11,

$$\{\theta\in\Theta : n^{-1/2}\sqrt{L_n^{MNV}(\theta,\theta_0)} \geq 2K\sqrt{C\epsilon^2}\} = \{\theta\in\Theta : d_L^{MNV}(\theta,\theta_0) \geq 2K\sqrt{C\epsilon^2}\} \subseteq \{\theta\in\Theta : \|\theta-\theta_0\| \geq 2\sqrt{C}$$

which implies that

$$\sup_{\{\theta\in\Theta:L_n^{MNV}(\theta,\theta_0)\geq 4K^2Cn\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_n] \leq \sup_{\{\theta\in\Theta:\|\theta-\theta_0\|\geq 2\sqrt{C\epsilon^2}\}} \mathbb{E}_{P_\theta^n}[1-\phi_n].$$

Therefore, P_θ for $\theta \in \Theta$, satisifes Assumptions 2.2.1 for $L_n(\theta,\theta_0) = L_n^{MNV}(\theta,\theta_0)$ for $C_0 = 1$, $C_1 = 4K^2C$ and $C = 1/8\left(\frac{C}{d\hat{\Lambda}}-1\right)$. □

Proof. Proof of Proposition 2.5.7:

First, we write the Rényi divergence between two multivariate Gaussian distribution with known Σ as

$$D_{1+\lambda}(\mathcal{N}(\cdot|\theta_0)\|\mathcal{N}(\cdot|\theta)) = \frac{\lambda+1}{2}(\theta-\theta_0)^T\Sigma(\theta-\theta_0), \tag{2.108}$$

and $D_{1+\lambda}(\mathcal{N}(\cdot|\theta)\|\mathcal{N}(\cdot|\theta_0)) < \infty$ if and only if Σ^{-1} is positive definite [98].

Since, we assumed that the sequence of models are iid, therefore,

$$D_{1+\lambda}\left(P_0^n \| P_\theta^n\right) = \frac{1}{\lambda}\log\int\left(\frac{dP_0^n}{dP_\theta^n}\right)^\lambda dP_0^n = n\frac{1}{\lambda}\log\int\left(\frac{dP_0}{dP_\theta}\right)^\lambda dP_0 = n\left(\frac{\lambda+1}{2}(\theta-\theta_0)^T\Sigma(\theta-\theta_0)\right),$$

when Σ^{-1} is positive definite and $D_{1+\lambda}\left(P_0^n \| P_\theta^n\right) = \infty$ otherwise. Now observe that

$$\begin{aligned}
\Pi(D_{1+\lambda}\left(P_0^n \| P_\theta^n\right) \leq nC_3\epsilon_n^2) &= \Pi\left(\left((\theta-\theta_0)^T\Sigma(\theta-\theta_0)\right) \leq \frac{2}{\lambda+1}C_3\epsilon_n^2\right) \\
&= \Pi\left(\left([(\theta-\theta_0)Q]^T\Lambda[Q(\theta-\theta_0)]\right) \leq \frac{2}{\lambda+1}C_3\epsilon_n^2\right) \\
&\geq \Pi\left(\left([(\theta-\theta_0)Q]^T[Q(\theta-\theta_0)]\right) \leq \frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2\right), \\
&= \Pi\left(\left([(\theta-\theta_0)]^T[(\theta-\theta_0)]\right) \leq \frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2\right), \qquad (2.109)
\end{aligned}$$

where $\hat{\Lambda} = \max_{\mathrm{i}\in\{1,2,\ldots d\}}\Lambda_{\mathrm{ii}}$ and in the second equality we used eigen value decomposition of $\Sigma = Q^T\Lambda Q$. Next, observe that,

$$\begin{aligned}
\Pi(D_{1+\lambda}\left(P_0^n \| P_\theta^n\right) \leq nC_3\epsilon_n^2) &= \Pi\left(\left([(\theta-\theta_0)]^T[(\theta-\theta_0)]\right) \leq \frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2\right) \\
&= \Pi\left(\|(\theta-\theta_0)\| \leq \sqrt{\frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2}\right) \\
&\geq \Pi\left(\|(\theta-\theta_0)\|_\infty \leq \sqrt{\frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2}\right) \\
&= \prod_{\mathrm{i}=1}^{d}\Pi_{\mathrm{i}}\left(|(\theta_{\mathrm{i}}-\theta_0^{\mathrm{i}})| \leq \sqrt{\frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2}\right),
\end{aligned}$$

where in the last equality we used the fact that the prior distribution is uncorrelated. Now, the result follows immediately for sufficiently large n, if the prior distribution is uncorrelated

and uniformly distributed on the compact set Θ_{i}, for each $\mathrm{i} \in \{1, 2, \ldots, d\}$. In particular observe that for large enough n, we have

$$\Pi(D_{1+\lambda}\left(P_0^n \| P_\theta^n\right) \leq nC_3\epsilon_n^2) \geq \prod_{\mathrm{i}=1}^{d} \frac{\theta_0^{\mathrm{i}} + \sqrt{\frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2} - \theta_0^{\mathrm{i}} + \sqrt{\frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2}}{m(\Theta_{\mathrm{i}})}$$
$$= \frac{2^d\left(\frac{2}{\hat{\Lambda}(\lambda+1)}C_3\epsilon_n^2\right)^{d/2}}{\prod_{\mathrm{i}=1}^{d} m(\Theta_{\mathrm{i}})} = \left(\frac{8}{\hat{\Lambda}(\lambda+1)\left(\prod_{\mathrm{i}=1}^{d} m(\Theta_{\mathrm{i}})\right)^{2/d}}C_3\epsilon_n^2\right)^{d/2}, \tag{2.110}$$

where $m(A)$ is the Lebesgue measure (volume) of any set $A \subset \mathbb{R}$. Now if $\epsilon_n^2 = \frac{\log n}{n}$, then for $\frac{8}{\hat{\Lambda}(\lambda+1)\left(\prod_{\mathrm{i}=1}^{d} m(\Theta_{\mathrm{i}})\right)^{2/d}}C_3 > 2$, $\frac{8}{\hat{\Lambda}(\lambda+1)\left(\prod_{\mathrm{i}=1}^{d} m(\Theta_{\mathrm{i}})\right)^{2/d}}C_3\epsilon_n^2 \geq \mathrm{e}^{-\frac{8}{\hat{\Lambda}(\lambda+1)\left(\prod_{\mathrm{i}=1}^{d} m(\Theta_{\mathrm{i}})\right)^{2/d}}C_3 n\epsilon_n^2}$ for all $n \geq 2$, therefore,

$$\Pi(D_{1+\lambda}\left(P_0^n \| P_\theta^n\right) \leq nC_3\epsilon_n^2) \geq \mathrm{e}^{-\frac{4d}{\hat{\Lambda}(\lambda+1)\left(\prod_{\mathrm{i}=1}^{d} m(\Theta_{\mathrm{i}})\right)^{2/d}}C_3 n\epsilon_n^2}.$$

□

Proof. Proof of Proposition 2.5.9: Since family $\mathcal{Q}$ contains all uncorrelated Gaussian distributions restricted to Θ, observe that $\{q_n(\cdot) \in \mathcal{Q}\}\forall n \geq 1$. By definition, $q_n^{\mathrm{i}}(\theta) \propto \frac{1}{\sqrt{2\pi\sigma_{\mathrm{i},n}^2}}\mathrm{e}^{-\frac{1}{2\sigma_{\mathrm{i},n}^2}(\theta-\mu_{\mathrm{i},n})^2}\mathbb{1}_{\Theta_{\mathrm{i}}} = \frac{\mathcal{N}(\theta_{\mathrm{i}}|\mu_{\mathrm{i},n},\sigma_{\mathrm{i},n})\mathbb{1}_{\Theta_{\mathrm{i}}}}{\mathcal{N}(\Theta_{\mathrm{i}}|\mu_{\mathrm{i},n},\sigma_{\mathrm{i},n})}$ and fix $\sigma_{\mathrm{i},n} = 1/\sqrt{n}$ and $\theta_{\mathrm{i}} = \theta_0^{\mathrm{i}}$ for all $\mathrm{i} \in \{1, 2, \ldots, d\}$. Now consider the first term; using the definition of the KL divergence it follows that

$$\mathrm{KL}(q_n(\theta)\|\boldsymbol{\pi}(\theta)) = \int q_n(\theta)\log(q_n(\theta))d\theta - \int q_n(\theta)\log(\boldsymbol{\pi}(\theta))d\theta. \tag{2.111}$$

Substituting $q_n(\theta)$ in the first term of the equation above and expanding the logarithm term, we obtain

$$\begin{aligned}\int q_n(\theta)\log(q_n(\theta))d\theta &= \sum_{\mathrm{i}=1}^{d}\int q_n^{\mathrm{i}}(\theta_{\mathrm{i}})\log(q_n^{\mathrm{i}}(\theta_{\mathrm{i}}))d\theta_{\mathrm{i}} \\ &\leq \sum_{\mathrm{i}=1}^{d}\int \mathcal{N}(\theta_{\mathrm{i}}|\mu_{\mathrm{i},n},\sigma_{\mathrm{i},n})\log\mathcal{N}(\theta_{\mathrm{i}}|\mu_{\mathrm{i},n},\sigma_{\mathrm{i},n})d\theta_{\mathrm{i}} \\ &= -\sum_{\mathrm{i}=1}^{d}[\log(\sqrt{2\pi\mathrm{e}}) + \log\sigma_{\mathrm{i},n}], \end{aligned} \tag{2.112}$$

where in the last equality, we used the well known expression for the differential entropy of Gaussian distributions. Recall $\boldsymbol{\pi}(\theta) = \prod_{\mathrm{i}=1}^{d}\frac{1}{m(\Theta_{\mathrm{i}})}$. Now consider the second term in (2.111). It is straightforward to observe that,

$$-\int q_n(\theta)\log(\boldsymbol{\pi}(\theta))d\theta = \sum_{\mathrm{i}=1}^{d}\log(m(\Theta_{\mathrm{i}})). \tag{2.113}$$

Substituting (2.113) and (2.112) into (2.111) and dividing either sides by n and substituting $\sigma_{\mathrm{i},n}$, we obtain

$$\begin{aligned}\frac{1}{n}\mathrm{KL}(q_n(\theta)\|\boldsymbol{\pi}(\theta)) &\leq -\frac{1}{n}\sum_{\mathrm{i}=1}^{d}[\log(\sqrt{2\pi\mathrm{e}}) - \log(m(\Theta_{\mathrm{i}})) - \frac{1}{2}\log n] \\ &= \frac{d}{2}\frac{\log n}{n} - \frac{1}{n}\sum_{\mathrm{i}=1}^{d}[\log(\sqrt{2\pi\mathrm{e}}) - \log(m(\Theta_{\mathrm{i}}))]. \end{aligned} \tag{2.114}$$

Now, consider the second term in the assertion of the lemma. Since $\xi_{\mathrm{i}}, \mathrm{i} \in \{1, 2\ldots n\}$ are independent and identically distributed, we obtain

$$\frac{1}{n}\mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n\|p(\tilde{X}_n|\theta)\right)\right] = \mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0\|p(\xi|\theta)\right)\right]$$

Now using the expression for KL divergence between the two multivariate Gaussian distributions, we have

$$\begin{aligned}\frac{1}{n}\mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n \| p(\tilde{X}_n|\theta)\right)\right] &= \frac{1}{2}\mathbb{E}_{q_n(\theta)}\left[(\theta-\theta_0)^T\Sigma^{-1}(\theta-\theta_0)\right] \\ &\leq \frac{\check{\Lambda}^{-1}}{2}\mathbb{E}_{q_n(\theta)}\left[(\theta-\theta_0)^T(\theta-\theta_0)\right] \\ &\leq \frac{d}{n}\frac{\check{\Lambda}^{-1}}{2} \end{aligned} \tag{2.115}$$

where $\check{\Lambda} = \min_{\mathrm{i}\in\{1,2,\ldots d\}}\Lambda_{\mathrm{ii}}$, and $\Sigma^{-1} = Q^T\Lambda^{-1}Q$, where Q is an orthogonal matrix and Λ is a daigonal matrix consisting of the respective eigen values of Σ. Combined together (2.115) and (2.114) implies that

$$\begin{aligned}&\frac{1}{n}\left[\mathrm{KL}\left(q_n(\theta)\|\boldsymbol{\pi}(\theta)\right) + \mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n)\|p(\tilde{X}_n|\theta)\right)\right]\right] \\ &\quad\leq \frac{d}{2}\frac{\log n}{n} - \frac{1}{n}\sum_{\mathrm{i}=1}^{d}\left[\log(\sqrt{2\pi\mathrm{e}}) - \log(m(\Theta_{\mathrm{i}}))\right] + \frac{d}{n}\frac{\check{\Lambda}^{-1}}{2} \leq C_9\frac{\log n}{n}. \end{aligned} \tag{2.116}$$

where $C_9 := \frac{d}{2} + \max\left(0, -\sum_{\mathrm{i}=1}^{d}\left[\log(\sqrt{2\pi\mathrm{e}}) - \log(m(\Theta_{\mathrm{i}}))\right] + \frac{d}{2}\check{\Lambda}^{-1}\right)$ and the result follows. □

2.7.10 Gaussian process classification

Proof of Proposition 2.5.11. In view of Theorem 7.1 in [91], it suffices to show that

$$N\left(\epsilon, \Theta_n(\epsilon), d_{\mathrm{TV}}\right) \leq \mathrm{e}^{\bar{C}n\epsilon^2},$$

for some $\bar{C} > 0$. Now, first observe that

$$\begin{aligned} d_{\mathrm{TV}}(P_{\theta(y)}, P_{\theta_0(y)}) &= \frac{1}{2}\mathbb{E}_\nu\left(|\Psi_1(\theta(y)) - \Psi_1(\theta_0(y))| + |\Psi_{-1}(\theta(y)) - \Psi_{-1}(\theta_0(y))|\right) \\ &= \mathbb{E}_\nu\left(|\Psi_1(\theta(y)) - \Psi_1(\theta_0(y))|\right) \\ &\leq \mathbb{E}_\nu\left(|\theta(y) - \theta_0(y)|\right) \leq \|\theta(y) - \theta_0(y)\|_\infty, \end{aligned} \tag{2.117}$$

where the second equality uses the definition of $\Psi_{-1}(\cdot)$. Since, total-variation distance is bounded above by supremum norm, hence there exists a constant $0 < c < 1/2$, such that

$$N\left(\epsilon, \Theta_n(\epsilon), d_{\mathrm{TV}}\right) \leq N\left(c\epsilon, \Theta_n(\epsilon), \|\cdot\|_\infty\right) \leq \mathrm{e}^{\frac{2}{3}c^2 C_{10} n\epsilon^2}, \tag{2.118}$$

where the last inequality follows from (2.28) in Lemma 2.5.10. Then if follows from Theorem 7.1 in [91] that for every $\epsilon > \epsilon_n$, there exists a test ϕ_n (depending on $\epsilon > 0$) such that, for every $\mathrm{j} \geq 1$,

$$\mathbb{E}_{P_0^n}[\phi_n] \leq \mathrm{e}^{\frac{2}{3}c^2 C_{10} n\epsilon^2} \mathrm{e}^{-\frac{1}{2}n\epsilon^2} \frac{1}{1 - \exp\left(-\frac{1}{2}n\epsilon^2\right)}, \text{and}$$
$$\sup_{\{\theta \in \Theta_n(\epsilon): d_{TV}(P_\theta, P_{\theta_0}) > \mathrm{j}\epsilon\}} \mathbb{E}_{P_\theta^n}[1 - \phi_n] \leq \exp\left(-\frac{1}{2}n\epsilon^2 \mathrm{j}\right).$$

Now for all n such that $n\epsilon^2 > n\epsilon_n^2 > 2\log 2$ and $C_{10} = c^{-2}/4 > 1$ and $\mathrm{j} = 1$, we have

$$\mathbb{E}_{P_0^n}[\phi_n] \leq 2\mathrm{e}^{-\frac{1}{3}n\epsilon^2}, \text{and} \tag{2.119}$$
$$\sup_{\{\theta \in \Theta_n(\epsilon): d_{TV}(P_\theta, P_{\theta_0}) > \epsilon\}} \mathbb{E}_{P_\theta^n}[1 - \phi_n] \leq \mathrm{e}^{-\frac{1}{2}n\epsilon^2} \leq \mathrm{e}^{-\frac{1}{3}n\epsilon^2}. \tag{2.120}$$

Now observe that

$$\begin{aligned}
\sup_{a \in \mathcal{A}} |G(a,\theta) - G(a,\theta_0)| &= \max\left(c_+ |\mathbb{E}_\nu[\Psi_{-1}(\theta(y))] - \mathbb{E}_\nu[\Psi_{-1}(\theta_0(y))]|, c_- |\mathbb{E}_\nu[\Psi_1(\theta(y))] - \mathbb{E}_\nu[\Psi_1(\theta_0(y))]|\right) \\
&= \max\left(c_+ |\mathbb{E}_\nu[\Psi_1(\theta_0(y))] - \mathbb{E}_\nu[\Psi_1(\theta(y))]|, c_- |\mathbb{E}_\nu[\Psi_1(\theta(y))] - \mathbb{E}_\nu[\Psi_1(\theta_0(y))]|\right) \\
&= \max(c_+, c_-) |\mathbb{E}_\nu[\Psi_1(\theta_0(y))] - \mathbb{E}_\nu[\Psi_1(\theta(y))]| \\
&\leq \max(c_+, c_-) \mathbb{E}_\nu[|\Psi_1(\theta_0(y)) - \Psi_1(\theta(y))|] \\
&\leq \max(c_+, c_-) d_{TV}(P_\theta, P_{\theta_0}) \qquad (2.121)
\end{aligned}$$

where the second equality uses the fact that $\Psi_{-1}(\cdot) = 1 - \Psi_1(\cdot)$.Consequently,

$$\{\theta \in \Theta_n(\epsilon) : \sup_{a \in \mathcal{A}} |G(a,\theta) - G(a,\theta_0)| > \max(c_+, c_-)\epsilon\} \subseteq \{\theta \in \Theta_n(\epsilon) : d_{TV}(P_\theta, P_{\theta_0}) > \epsilon\}$$

Therefore, it follows from (2.119) and (2.120) and the definition of $L_n(\theta, \theta_0)$ that

$$\mathbb{E}_{P_0^n}[\phi_n] \leq 2\mathrm{e}^{-\frac{1}{3}n\epsilon^2}, \text{and} \tag{2.122}$$

$$\sup_{\{\theta\in\Theta_n(\epsilon):L_n(\theta,\theta_0)>(\max(c_+,c_-))^2n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_n] \leq \mathrm{e}^{-\frac{1}{2}n\epsilon^2} \leq \mathrm{e}^{-\frac{1}{3}n\epsilon^2}. \tag{2.123}$$

Finally, the result follows for $C = 1/3$, $C_0 = 2$ and $C_1 = (\max(c_+, c_-))^2$.

□

Proof of Proposition 2.5.12. The Rényi divergence

$$\begin{aligned} D_{1+\lambda}(P_0^n \| P_\theta^n) &= n\frac{1}{\lambda}\ln\int \left(\Psi_1(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{-\lambda} + \Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{-\lambda}\right)\nu(dy) \\ &= n\frac{1}{\lambda}\ln\int \mathrm{e}^{\lambda\frac{1}{\lambda}\ln\left(\Psi_1(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{-\lambda}+\Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{-\lambda}\right)}\nu(dy). \end{aligned}$$

Note that the derivative of the exponent in the integrand above with respect to $\theta(y)$ is

$$\begin{aligned} &\frac{\left(-\lambda\Psi_1(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{-\lambda-1}\psi(\theta(y)) + \lambda\Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{-\lambda-1}\psi(\theta(y))\right)}{(\Psi_1(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{-\lambda} + \Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{-\lambda})} \\ &= \lambda\psi(\theta(y))\frac{\left(-\Psi_1(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{-\lambda-1} + \Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{-\lambda-1}\right)}{(\Psi_1(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{-\lambda} + \Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{-\lambda})} \\ &= \lambda\frac{\psi(\theta(y))}{\Psi_1(\theta(y))\Psi_{-1}(\theta(y))}\frac{\left(-\Psi_1(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{\lambda+1} + \Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{\lambda+1}\right)}{(\Psi_1(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{\lambda} + \Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{\lambda})} \\ &= \lambda\frac{\left(-\Psi_1(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{\lambda+1} + \Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{\lambda+1}\right)}{(\Psi_1(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{\lambda} + \Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{\lambda})} \\ &= \lambda\frac{\left(-\mathrm{e}^{-(\lambda+1)\theta(y)} + \mathrm{e}^{-(1+\lambda)\theta_0(y)}\right)}{\left(\mathrm{e}^{-\lambda\theta(y)} + \mathrm{e}^{-(\lambda+1)\theta_0(y)}\right)\left(1+\mathrm{e}^{-\theta(y)}\right)} \\ &= \lambda\frac{\mathrm{e}^{-(1+\lambda)\theta_0(y)}\left(1-\mathrm{e}^{-(\lambda+1)(\theta(y)-\theta_0(y))}\right)}{\left(\mathrm{e}^{-\lambda\theta(y)} + \mathrm{e}^{-(\lambda+1)\theta_0(y)}\right)\left(1+\mathrm{e}^{-\theta(y)}\right)} \\ &\leq \lambda\frac{(\lambda+1)(\theta(y)-\theta_0(y))}{\left(\mathrm{e}^{-\lambda\theta(y)+(\lambda+1)\theta_0(y)}+1\right)\left(1+\mathrm{e}^{-\theta(y)}\right)} \\ &\leq \lambda(\lambda+1)|\theta(y)-\theta_0(y)|, \end{aligned} \tag{2.124}$$

where in the fourth equality we used definition of the logistic function and the penultimate inequality follows from the well known inequality that $1 - \mathrm{e}^{-x} \leq x$. Consequently, using

Taylor's theorem it follows that the exponent in the integrand is bounded above by $\lambda(\lambda+1)|\theta(y)-\theta_0(y)|^2$ and thus by $\lambda(\lambda+1)\|\theta(y)-\theta_0(y)\|_\infty^2$. Therefore,

$$\begin{aligned} D_{1+\lambda}(P_0^n\|P_\theta^n) &= n\frac{1}{\lambda}\ln\int\left(\Psi_1(\theta_0(y))^{1+\lambda}\Psi_1(\theta(y))^{-\lambda}+\Psi_{-1}(\theta_0(y))^{1+\lambda}\Psi_{-1}(\theta(y))^{-\lambda}\right)\nu(dy) \\ &\leq n\frac{1}{\lambda}\ln\int \mathrm{e}^{\lambda(\lambda+1)\|\theta(y)-\theta_0(y)\|_\infty^2}\nu(dy) \\ &= n(\lambda+1)\|\theta(y)-\theta_0(y)\|_\infty^2. \end{aligned}$$

Now using the inequality for $C_3 = 16(\lambda+1)$ above observe that

$$\begin{aligned} \Pi(A_n) &= \Pi(D_{1+\lambda}\left(P_0^n\|P_\theta^n\right) \leq C_3 n\epsilon_n^2) \\ &\geq \Pi(n(\lambda+1)\|\theta(y)-\theta_0(y)\|_\infty^2 \leq C_3 n\epsilon_n^2) \\ &= \Pi(\|\theta(y)-\theta_0(y)\|_\infty \leq 4\epsilon_n) \geq \mathrm{e}^{-n\epsilon_n^2} \end{aligned} \tag{2.125}$$

and the result follows from (2.30) of Lemma 2.5.10.

□

Proof of Proposition 2.5.13. Let us first analyze the KL divergence between the prior distribution and variational family. Recall that two Gaussian measures on infinite dimensional spaces are either equivalent or singular. [20, Theorem 6.13] specify the condition required for the two Gaussian measures to be equivalent. In particular, note that $\theta_0^J(\cdot) \in \mathrm{Im}(\mathcal{C}^{1/2})$. Now observe that the covariance operator of Q_n has eigenvalues $\{\zeta_{\mathrm{j}}^2\}_{\mathrm{j}=1}^{J}{}_{k=1}^{2^{\mathrm{j}d}}$, therefore operator S in the definition of $\mathcal{C}_q$ has eigenvalues $\{1-\zeta_{\mathrm{j}}^2/\mu_{\mathrm{j}}^2\}_{\mathrm{j}=1}^{J}{}_{k=1}^{2^{\mathrm{j}d}}$. For $\tau_{\mathrm{j}}^2 = 2^{-2\mathrm{j}a-\mathrm{j}d}$ for any $a>0$, $\sum_{\mathrm{j}=1}^{J} 2^{\mathrm{j}d}\left(\frac{n\epsilon_n^2 2^{-2\mathrm{j}a-\mathrm{j}d}}{1+n\epsilon_n^2 2^{-2\mathrm{j}a-\mathrm{j}d}}\right)^2 = \sum_{\mathrm{j}=1}^{J} 2^{-\mathrm{j}d}\left(\frac{n\epsilon_n^2 2^{-2\mathrm{j}a}}{1+n\epsilon_n^2 2^{-2\mathrm{j}a-\mathrm{j}d}}\right)^2 < \infty$, therefore S is an HS operator.

For any integer $J \leq J_\alpha$ define $\bar{\theta}_0^J = \int \theta_0^J(y)\nu(dy)$, where $\theta_0^J(\cdot) = \sum_{\mathrm{j}=1}^{J}\sum_{k=1}^{2^{\mathrm{j}d}} \theta_{0;\mathrm{j},k}\vartheta_{\mathrm{j},k}(\cdot)$. Since, $\theta_0^J(\cdot) \in \mathrm{Im}(\mathcal{C}^{1/2})$ and S is a symmetric and HS operator, we invoke Theorem 5 in [99], to write

$$\begin{aligned}
\mathrm{KL}(\mathcal{N}(\bar{\theta}_0^J, \mathcal{C}_q)\|\mathcal{N}(0,\mathcal{C})) &= \frac{1}{2}\|\mathcal{C}^{-1/2}\bar{\theta}_0^J\|^2 - \frac{1}{2}\log\det(I-S) + \frac{1}{2}tr(-S), \\
&= \frac{1}{2}\sum_{\mathrm{j}=1}^{J}\sum_{k=1}^{2^{\mathrm{j}d}}\frac{\theta_{0;k,\mathrm{j}}^2}{\mu_\mathrm{j}^2} - \frac{1}{2}\log\prod_{\mathrm{j}=1}^{J}\prod_{k=1}^{2^{\mathrm{j}d}}(1-\kappa_\mathrm{j}^2) - \frac{1}{2}\sum_{\mathrm{j}=1}^{J}\sum_{k=1}^{2^{\mathrm{j}d}}\kappa_\mathrm{j}^2 \\
&= \frac{1}{2}\sum_{\mathrm{j}=1}^{J}\sum_{k=1}^{2^{\mathrm{j}d}}\frac{\theta_{0;k,\mathrm{j}}^2}{\mu_\mathrm{j}^2} - \frac{1}{2}\log\prod_{\mathrm{j}=1}^{J}(1-\kappa_\mathrm{j}^2)^{2^{\mathrm{j}d}} - \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}d}\kappa_\mathrm{j}^2 \\
&= \frac{1}{2}\sum_{\mathrm{j}=1}^{J}\sum_{k=1}^{2^{\mathrm{j}d}}\frac{\theta_{0;k,\mathrm{j}}^2}{\mu_\mathrm{j}^2} - \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}d}\log(1-\kappa_\mathrm{j}^2) - \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}d}\kappa_\mathrm{j}^2.
\end{aligned}$$

Now for $\mu_\mathrm{j}2^{\mathrm{j}d/2} = 2^{-\mathrm{j}a}$, and using the definition of Besov norm of θ_0 denoted as $\|\theta_0\|_{\beta,\infty,\infty}^2$, and denoting $1-\kappa_\mathrm{j}^2 = \frac{1}{1+n\epsilon_n^2\tau_\mathrm{j}^2}$, we have

$$\begin{aligned}
\mathrm{KL}(\mathcal{N}(\bar{\theta}_0^J, \mathcal{C}_q)\|\mathcal{N}(0,\mathcal{C})) &\leq \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}(2a-2\beta+d)}\|\theta_0\|_{\beta,\infty,\infty}^2 - \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}d}\log(1-\kappa_\mathrm{j}^2) - \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}d}\kappa_\mathrm{j}^2 \\
&= \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}(2a-2\beta+d)}\|\theta_0\|_{\beta,\infty,\infty}^2 - \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}d}\left(\log(1-\kappa_\mathrm{j}^2) + \kappa_\mathrm{j}^2\right) \\
&= \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}(2a-2\beta+d)}\|\theta_0\|_{\beta,\infty,\infty}^2 + \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}d}\left(\log(1+n\epsilon_n^2\tau_\mathrm{j}^2) - \frac{n\epsilon_n^2\tau_\mathrm{j}^2}{1+n\epsilon_n^2\tau_\mathrm{j}^2}\right) \\
&\leq \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}(2a-2\beta+d)}\|\theta_0\|_{\beta,\infty,\infty}^2 + \frac{1}{2}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}d}\left(n\epsilon_n^2\tau_\mathrm{j}^2\right),
\end{aligned}$$

where the last inequality follows from the fact that, $\log(1+x) - \frac{x}{1+x} \leq \frac{x^2}{1+x} \leq x$ for $x > 0$. Substituting $\tau_\mathrm{j}^2 = 2^{-2\mathrm{j}a-\mathrm{j}d}$, we have

$$\begin{aligned}
\frac{1}{n}\mathrm{KL}(\mathcal{N}(\bar{\theta}_0^J, \mathcal{C}_q)\|\mathcal{N}(0,\mathcal{C})) &\leq \frac{1}{2n}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}(2a-2\beta+d)}\|\theta_0\|_{\beta,\infty,\infty}^2 + \frac{\epsilon_n^2}{2}\sum_{\mathrm{j}=1}^{J}2^{-2\mathrm{j}a} \\
&\leq \frac{\|\theta_0\|_{\beta,\infty,\infty}^2}{2n}\sum_{\mathrm{j}=1}^{J}2^{\mathrm{j}(2a-2\beta+d)} + \frac{2^{-2a}}{2}\frac{1-2^{-2Ja}}{1-2^{-2a}}\epsilon_n^2.
\end{aligned}$$

The summation in the first term above is bounded by ϵ_n^2 as derived in [70, Theorem 4.5]. Therefore,

$$\frac{1}{n}\mathrm{KL}(\mathcal{N}(\bar{\theta}_0^J, \mathcal{C}_q)\|\mathcal{N}(0,\mathcal{C})) \leq \max\left(\|\theta_0\|^2_{\beta,\infty,\infty}, \frac{2^{-2a}-2^{-2Ja-2a}}{1-2^{-2a}}\right)\epsilon_n^2. \tag{2.126}$$

Now consider the second term

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_{Q_n}\mathrm{KL}(P_0^n\|P_\theta^n) &= \mathbb{E}_{Q_n}\int\left(\Psi_1(\theta_0(y))\log\frac{\Psi_1(\theta_0(y))}{\Psi_1(\theta(y))} + \Psi_{-1}(\theta_0(y))\log\frac{\Psi_{-1}(\theta_0(y))}{\Psi_{-1}(\theta(y))}\right)\nu(dy)\\
&\leq \mathbb{E}_{Q_n}\int\langle\theta(y)-\theta_0(y),\theta(y)-\theta_0(y)\rangle\nu(dy)\\
&= \mathbb{E}_{Q_n}\int\|\theta(y)-\theta_0^J(y)-(\theta_0(y)-\theta_0^J(y))\|_2^2\nu(dy)\\
&= \mathbb{E}_{Q_n}\int\|\theta(y)-\theta_0^J(y)\|_2^2+\|\theta_0(y)-\theta_0^J(y))\|_2^2-2\langle\theta(y)-\theta_0^J(y),\theta_0(y)-\theta_0^J(y)\rangle\nu(dy)\\
&\leq \mathbb{E}_{Q_n}\int\|\theta(y)-\theta_0^J(y)\|_2^2\nu(dy)+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&= \mathbb{E}_{Q_n}\int|\sum_{\mathrm{j}=1}^J\sum_{k=1}^{2^{\mathrm{j}d}}\zeta_{\mathrm{j}}Z_{\mathrm{j},k}\vartheta_{\mathrm{j},k}(y)|^2\nu(dy)+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&\leq \mathbb{E}_{Q_n}\sum_{\mathrm{j}=1}^J\sum_{k=1}^{2^{\mathrm{j}d}}\zeta_{\mathrm{j}}^2Z_{\mathrm{j},k}^2\int\vartheta_{\mathrm{j},k}(y)^2\nu(dy)+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&= \sum_{\mathrm{j}=1}^J\sum_{k=1}^{2^{\mathrm{j}d}}\zeta_{\mathrm{j}}^2\mathbb{E}_{Q_n}[Z_{\mathrm{j},k}^2]+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&= \sum_{\mathrm{j}=1}^J\sum_{k=1}^{2^{\mathrm{j}d}}\mu_{\mathrm{j}}^2(1-\kappa_{\mathrm{j}}^2)+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&= \sum_{\mathrm{j}=1}^J 2^{\mathrm{j}d}\frac{\mu_{\mathrm{j}}^2}{1+n\epsilon_n^2\tau_{\mathrm{j}}^2}+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&\leq \frac{1}{n\epsilon_n^2}\sum_{\mathrm{j}=1}^J\frac{2^{-2\mathrm{j}a}}{\tau_{\mathrm{j}}^2}+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&= \frac{1}{n\epsilon_n^2}\sum_{\mathrm{j}=1}^J 2^{\mathrm{j}d}+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&= \frac{2^d}{n\epsilon_n^2}\frac{2^{dJ}-1}{2^d-1}+\|\theta_0(y)-\theta_0^J(y))\|_\infty^2\\
&\leq \frac{2^d/(2^d-1)}{(\log n)^2}+C\epsilon_n^2,
\end{aligned}$$

where in the second inequality we used the second assertion of Lemma 3.2 [70] for logistic function, the fifth inequality uses the fact that $\theta(y) - \theta_0^J(y)$ is orthogonal to $\theta_0(y) - \theta_0^J(y)$. For any $a \leq \alpha$ fix $J = J_\alpha$ otherwise $J = J_a$, and then it is straight forward to check from the definition of ϵ_n given in the assertion of the theorem that $(2^{dJ-1}/n\epsilon_n^2) \leq (\log n)^{-2}$. The term $\|\theta_0(y) - \theta_0^J(y))\|_\infty^2$ is also bounded by $C\epsilon_n^2$ as shown in the proof of Theorem 4.5 in [70]. $\square$

3. ASYMPTOTIC CONSISTENCY OF LOSS-CALIBRATED VARIATIONAL BAYES

This chapter establishes the asymptotic consistency of the loss-calibrated variational Bayes (LCVB) method, a special case of the RSVB approach discussed in Chapter 2. This method of loss-aware posterior approximation was first proposed by [10]. Also, the theory in this chapter is applicable only to parametric (finite-dimensional) likelihood models unlike Chapter 2, where the theory was general enough to include non-parametric models. Here, we establish the asymptotic consistency of both the loss-calibrated approximate posterior and the resulting decision rules under relatively easy-to-verify regularity conditions than those required for establishing the rate of convergence for RSVB method. We also establish the asymptotic consistency of decision rules obtained from a "naive" two-stage procedure that first computes a standard variational Bayes approximation and then uses this in the decision-making procedure.

3.1 Introduction

Consider a loss function $R(a, \theta) : \mathcal{A} \times \Theta \mapsto \mathbb{R}$, where $a \in \mathcal{A} \subset \mathbb{R}^s$ is a decision variable and $\theta \in \Theta \subset \mathbb{R}^d$ is the model parameter. Given a set of observations $\tilde{X}_n = \{\xi_1, \ldots, \xi_n\}$ drawn from a distribution with unknown parameter θ_0, $p(\tilde{X}_n|\theta_0)$, our goal is to compute the Bayes optimal decision rule

$$\mathsf{a}^*(\tilde{X}_n) := \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\pi_n}[R(a, \theta)] = \int_\Theta R(a, \theta)\boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta, \tag{3.1}$$

where $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ is the posterior distribution. The latter results when a Bayesian decision-maker places a *prior* distribution $\boldsymbol{\pi}(\theta)$ over the parameter space Θ, capturing *a priori* information about θ such as location or spread. Given $\tilde{X}_n$, the prior and likelihood $p(\tilde{X}_n|\theta)$ together define a *posterior* distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n) \propto p(\tilde{X}_n|\theta)\boldsymbol{\pi}(\theta) =: p(\theta, \tilde{X}_n)$, the conditional distribution over θ given observations. The posterior distribution represents uncertainty over the unknown parameter θ, and contains all information required for further inferences or optimization.

In general, under most realistic modeling assumptions, closed-form analytic expressions are unavailable for $\boldsymbol{\pi}(\theta|\tilde{X}_n)$, making the subsequent integration and optimization problems intractable. In practice, therefore, one uses an approximation to the posterior in the integration in (3.1). It is easy to see that posterior computation can be expressed as a convex optimization problem:

$$\begin{aligned}\min_{q(\cdot)\in\mathcal{M}} \mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)) &= \mathrm{KL}(q(\theta)\|p(\theta,\tilde{X}_n)) + \log p(\tilde{X}_n) \\ &= \mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta)) - \int_\Theta \log p(\tilde{X}_n|\theta)\ q(\theta)d\theta + \log p(\tilde{X}_n)\end{aligned} \tag{3.2}$$

where KL is the Kullback-Leibler divergence and $\mathcal{M}$ is the space of all distributions that are absolutely continuous with respect to the posterior (or, equivalently, the prior). This problem can be immediately recognized as minimizing the 'variational free energy' [100]. Variational Bayesian (VB) procedures [101], in standard form, restrict the optimization in (3.2) to a fixed subset $\mathcal{Q} \subset \mathcal{M}$. Here, we are interested in a generalized version of this procedure where the posterior computation is *calibrated* by the loss function $R(a,\theta)$ for each $a \in \mathcal{A}$:

$$\begin{aligned}\min_{q(\cdot)\in\mathcal{Q}} \mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)) - \int_\Theta \log R(a,\theta)\ q(\theta)d\theta \\ = \mathrm{KL}(q(\theta)\|p(\theta,\tilde{X}_n)) + \log p(\tilde{X}_n) - \int_\Theta \log R(a,\theta)q(\theta)d\theta.\end{aligned} \tag{3.3}$$

Observe that the set $\mathcal{Q}$ need not be convex. Consequently, this optimization problem is non-convex, in full generality, and practical algorithms for solving (3.3) can only guarantee convergence to local minima. We leave the analysis of these optimization-related issues for future work, and focus instead on the global solution and its associated asymptotics. As we show later in Section 3.2.2 that the optimal value of this loss-calibrated VB objective turns out to be a lower bound to $\log \mathbb{E}_\pi[R(a,\theta)]$, the logarithm of the loss in (3.1).

Loss-calibration was introduced in [10] as a method for approximately computing a generalized Bayesian posterior, where the likelihood is re-weighted or calibrated by a loss function over the parameter space Θ. As with most VB methods, theoretical properties of the approximations present largely unanswered questions. Recently, the theoretical properties of

the variational Bayesian methods have been studied extensively in [27], [29], [39], [61], [102], [103]. [27] established the asymptotic consistency of the VB approximate posterior and also proved a Bernstein-von Mises' type result for the same. Whereas, the authors in [61] studied the convergence rate of the VB approximate posterior. [39] presented a general framework for computing a risk-sensitive VB approximation and also studies the statistical performance of the inferred decision rules using these methods. Furthermore [30], [104]–[106] studied theoretical properties of variational Bayesian methods defined using Hellinger distance, Wasserstein distance, and Rényi divergence respectively instead of Kullback-Liebler (KL) divergence. In this chapter, we study the asymptotic consistency of the loss-calibrated approximate posterior and the optimal decisions computed using the this approximate posterior, as the number of samples $n \to \infty$.

More precisely, in Proposition 3.3.1, we show (for fixed $a \in \mathcal{A}$ and an appropriate subset of distributions $\mathcal{Q}$) that as $n \to \infty$ the optimizer of (3.3) weakly converges to a Dirac delta distribution concentrated on the true parameter θ_0 for almost every sequence generated from the true data generating process. This result shows that the posterior concentrates for any $a \in \mathcal{A}$. The reason for this is manifest: observe that $R(a,\theta)\boldsymbol{\pi}(\theta|\tilde{X}_n) \propto (R(a,\theta)\boldsymbol{\pi}(\theta))\, p(\tilde{X}_n|\theta)$. Thus, the loss function can be seen as only changing the prior distribution in the posterior computation. As the number of samples increases, we should anticipate that any calibration effect is diminished. Extending this result, in Proposition 3.4.3 we show that the optimizers of the approximate decision making problem, computed using the loss calibrated VB posterior, are asymptotically consistent, in the sense that this set of optimizers will necessarily be included in the optimizers of the 'true' objective $R(a,\theta_0)$.

Finally, we illustrate our results on the so-called newsvendor problem, studied extensively in the operations research literature as a prototypical decision-making problem. In this problem, a newsvendor must decide on the number of newspapers to stack up before selling any over a given day. We operate under the assumption that the newsvendor can observe realizations of the demand, but does not know the precise data generation process. The goal is to find the optimal number of newspapers to stack that minimizes losses. We conduct numerical studies to show that both the loss calibrated and naive VB methods on this problem are consistent.

The remainder of the chapter is organized as follows. In Section 3.2 we formally introduce decision-theoretic variational Bayesian methods. In Section 3.3 we prove that the LCVB approximate posterior is asymptotically consistent. We build on this result and prove the consistency of the optimal decisions, using both the LCVB and NVB methods, in Section 3.4. Finally, we present our numerical results in Section 3.5.

3.2 Decision-theoretic Variational Bayes

3.2.1 The Naive Variational Bayes (NVB) Algorithm

The idea behind standard VB is to approximate the intractable posterior $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ with an element $q^*(\theta)$ of a simpler class of distributions $\mathcal{Q}$ known as *variational family*. Popular examples of $\mathcal{Q}$ include the family of Gaussian distributions, or the family of factorized 'mean-field' distributions that discard correlations between components of θ. A natural caveat to the choice of $\mathcal{Q}$ is that these distributions should be absolutely continuous with respect to the posterior (or equivalently, the prior). The variational solution q^* is the element of $\mathcal{Q}$ that is 'closest' to $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ in the sense of the Kullback-Leibler (KL) divergence:

$$\begin{aligned}\min_{q(\theta)\in\mathcal{Q}} \mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)) &= \mathrm{KL}(q\|p(\theta,\tilde{X}_n)) + \log p(\tilde{X}_n) \\ &= \mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta)) - \int_\Theta \log p(\tilde{X}_n|\theta)\ q(\theta)d\theta + \log p(\tilde{X}_n).\end{aligned} \tag{3.4}$$

VB approaches allow practitioners to bring tools from optimization to the challenging problem of Bayesian inference, with expectation-maximization [100] and gradient-based [107] methods being used to minimize equation (3.4). Note that this optimization problem is non-convex, since the constraint set $\mathcal{Q}$ is non-convex in general. Also, observe that the objective $\mathrm{KL}(q(\theta\|\boldsymbol{\pi}(\theta|\tilde{X}_n)))$ in (3.4) only requires the knowledge of posterior distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ up to the proportionality constant, since the normalizing term $\log p(\tilde{X}_n)$ does not depend on q.

The natural variational approximation to the optimization in (3.1) is to calculate the variational approximate expected posterior loss of taking an action a, and then perform the following optimization

$$\mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n) := \text{argmin } \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)]. \tag{3.5}$$

We call this the *naive variational Bayes* (NVB) decision rule. This algorithm involves two optimization steps in sequence, separating the approximation of the posterior in (3.4) from the decision optimization (3.5). This sequential procedure, in general, involves a loss in performance compared to (3.1). This creates the desideratum for a calibrated approach that takes the loss function into consideration in computing an appropriate posterior.

3.2.2 Loss-Calibrated Variational Bayes (LCVB) Algorithm

A more sophisticated approach is to jointly optimize q and a; one would expect this to outperform the naive two-stage NVB algorithm. Assuming that the objective $\inf_{a,\theta} R(a,\theta) > 0$, a loss-calibrated lower bound can be derived by applying Jensen's inequality to the logarithm of the objective in (3.1), obtaining

$$\log \mathbb{E}_{\boldsymbol{\pi}(\theta|\tilde{X}_n)}[R(a,\theta)] = \log \int_\Theta \frac{q(\theta)}{q(\theta)} R(a,\theta)\boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta \geq -\int_\Theta q(\theta) \log \frac{q(\theta)}{R(a,\theta)\boldsymbol{\pi}(\theta|\tilde{X}_n)} d\theta \ \forall a \in \mathcal{A}.$$

In particular, it can be seen that

$$\begin{aligned}\min_{a\in\mathcal{A}} \log \mathbb{E}_{\boldsymbol{\pi}(\theta|\tilde{X}_n)}[R(a,\theta)] \geq \min_{a\in\mathcal{A}} \max_{q\in\mathcal{Q}} &- \mathrm{KL}(q(\theta)||\boldsymbol{\pi}(\theta|\tilde{X}_n)) \\ &+ \int_\Theta \log R(a,\theta) q(\theta) d\theta =: \mathcal{F}(a,q;\tilde{X}_n).\end{aligned} \tag{3.6}$$

We call (3.6) the *loss-calibrated* (LC) variational objective. Since $\log(\cdot)$ is a monotone transformation, minimizing the logarithmic objective on the left hand side above is equivalent

to (3.1). Now, for any given $a \in \mathcal{A}$ we denote the (globally maximal) LCVB approximate posterior as

$$q_a^*(\theta|\tilde{X}_n) := \text{argmax}_{q\in\mathcal{Q}}\mathcal{F}(a, q; \tilde{X}_n). \tag{3.7}$$

If the risk function $R(a, \theta)$ is constant then $q_a^*(\theta|\tilde{X}_n)$, for every $a \in \mathcal{A}$, is the same as $q^*(\theta|\tilde{X}_n)$. Akin to $q^*(\theta|\tilde{X}_n)$ in (3.4), computing $q_a^*(\theta|\tilde{X}_n)$ only requires knowledge of the posterior distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ up to a proportionality constant. The corresponding LCVB decision-rule is defined as

$$\mathsf{a}_{\mathsf{LC}}^*(\tilde{X}_n) := \arg\min_{a\in\mathcal{A}} \max_{q\in\mathcal{Q}} \mathcal{F}(a, q; \tilde{X}_n). \tag{3.8}$$

Observe that the lower bound achieves the log posterior value precisely for q such that $\frac{q(\theta)}{R(a,\theta)}$ is proportional to the posterior $\boldsymbol{\pi}(\theta|\tilde{X}_n)$. Furthermore, (3.6) shows that the maximization in the lower bound computes a 'regularized' approximate posterior. Regularized Bayesian inference [108] views posterior computation as a variational inference problem with constraints on the posterior space represented as bounds on certain expectations with respect to the approximate posterior. The loss-calibrated VB methodology can be viewed as a regularized Bayesian inference procedure where the regularization constraints are imposed through the logarithmic risk term $\int_\Theta \log R(a, \theta)q(\theta)d\theta$. Observe, however, that our setting also involves a minimization over the decisions (which does not exist in the regularized Bayesian inference procedure).

3.3 Consistency of the LCVB Approximate Posterior

Recall the definition of the LCVB approximate posterior $q_a^*(\theta|\tilde{X}_n)$ in (3.7) for any $a \in \mathcal{A}$. In this section, we show regularity conditions on the prior distribution, the risk function, the likelihood model, and the variational family, under which $q_a^*(\theta|\tilde{X}_n)$, for any $a \in \mathcal{A}$, converges weakly to a Dirac-delta distribution at the true parameter θ_0. We first assume that the prior distribution satisfies

Assumption 3.3.1. *The prior density function $\boldsymbol{\pi}(\theta)$ is continuous with non-zero measure in the neighborhood of the true parameter θ_0 and it is bounded by a positive constant M, that is $\boldsymbol{\pi}(\theta) < M, \forall \theta \in \Theta$.*

The prior distribution with bounded density can be chosen from a large class of distribution, like the exponential-family distributions. The first condition, that the prior has positive density at θ_0 is a common assumption in Bayesian consistency analysis, otherwise the posterior will not have any measure in the ball around the true parameter θ_0.

We also assume the expected loss, or risk function, $R(a, \theta)$ satisfies the following

Assumption 3.3.2. *The risk function $R(a, \theta)$ is*

1. *bounded from below by positive number W for any (a, θ),*
2. *measurable and continuous for every $a \in \mathcal{A}$, and $R(\cdot, \theta)$ are continuous for almost every $\theta \in \Theta$.*
3. *$R(\cdot, \theta)$ is locally Lipschitz continuous in a with for almost every $\theta \in \Theta$, such that for a_1, a_2 in compact set $\mathcal{A}$, $|R(a_1, \theta) - R(a_2, \theta)| \leq K_{\mathcal{A}}(\theta)\|x_1 - x_2\|$ for some $K_{\mathcal{A}}(\theta) \leq \bar{K}_{\mathcal{A}}$ for almost every $\theta \in \Theta$.*
4. *uniformly integrable with respect to any q in the variational family $\mathcal{Q}$, that is for any $\epsilon > 0$ and $a \in \mathcal{A}$, there exist a compact set $K_\epsilon \subset \Theta$, such that $\int_{\Theta \setminus K_\epsilon} R(a, \theta) q(\theta) d\theta < \epsilon$.*

In order to analyze the consistency of the decisions in this case, we make a further assumption on the log-likelihood function (which follows [27]):

Assumption 3.3.3. *The likelihood satisfies the* local asymptotic normality *(LAN) condition. In particular, fix $\theta_0 \in \Theta$. The sequence of log-likelihood functions $\{\log P_n(\theta) = \sum_{\mathrm{i}=1}^{n} \log p(\xi_{\mathrm{i}}|\theta)\}$ satisfies a* local asymptotic normality (LAN) *condition, if there exists a sequence of matrices $\{r_n\}$, a matrix $I(\theta_0)$ and a sequence of random vectors $\{\Delta_{n,\theta_0}\}$ weakly converging to $\mathcal{N}(0, I(\theta_0)^{-1})$ as $n \to \infty$, such that for every compact set $K \subset \mathbb{R}^d$*

$$\sup_{h \in K} \left| \log P_n(\theta_0 + r_n^{-1} h) - \log P_n(\theta_0) - h^T I(\theta_0) \Delta_{n,\theta_0} + \frac{1}{2} h^T I(\theta_0) h \right| \xrightarrow{P_0^n} 0 \text{ as } n \to \infty \ .$$

This LAN condition is typical in asymptotic analyses, holding for a wide variety of models and allowing the likelihood to be asymptotically approximated by a scaled Gaussian centered around θ_0 [109]. We use $\Delta_{n,\theta} = \sqrt{n}(\hat{\theta}_n - \theta_0)$ in the proofs of our results, where $\hat{\theta}_n$ is the maximum likelihood estimate of θ_0.

Next, we define the rate of convergence of a sequence of distributions to a Dirac delta distribution.

Definition 3.3.1 (Rate of convergence)**.** *A sequence of distributions $\{q_n(\theta)\}$ converges weakly to δ_{θ_1}, $\forall \theta_1 \in \Theta$ at the rate of γ_n if*

(1) the sequence of means $\{\check{\theta}_n := \int \theta q_n(\theta) d\theta\}$ converges to θ_1 as $n \to \infty$, and

(2) the variance of $\{q_n(\theta)\}$ satisfies

$$E_{q_n(\theta)}[\|\theta - \check{\theta}_n\|^2] = O\left(\frac{1}{\gamma_n^2}\right).$$

We also define rescaled density functions as follows.

Definition 3.3.2 (Rescaled density)**.** *For a random variable ξ distributed as $d(\xi)$ with expectation $\tilde{\xi}$, for any sequence of matrices $\{t_n\}$, the density of the rescaled random variable $\mu := t_n(\xi - \tilde{\xi})$ is*

$$\check{d}_n(\mu) = |det(t_n^{-1})| d(t_n^{-1}\mu + \tilde{\xi}),$$

where $det(\cdot)$ represents the determinant of the matrix.

Next, we place a restriction on the variational family $\mathcal{Q}$:

Assumption 3.3.4.

1. *The variational family $\mathcal{Q}$ must contain distributions that are absolutely continuous with respect to the posterior distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n)$.*

2. *There exists a sequence of distributions $\{q_n(\theta)\}$ in the variational family $\mathcal{Q}$ that converges to a Dirac delta distribution δ_{θ_0} at the rate of $\sqrt{n}$ and with mean $\int \theta q_n(\theta) d\theta = \hat{\theta}_n$, the maximum likelihood estimate.*

3. *The differential entropy of the rescaled density of such sequence of distributions is positive and finite.*

The first condition is necessary, since the KL divergence in (3.4) and (3.6) is undefined for any distribution $q \in \mathcal{Q}$, that is not absolutely continuous with respect to the posterior distribution. The Bernstein von-Mises theorem shows that under mild regularity conditions, the posterior converges to a Dirac delta distribution at the true parameter θ_0 at the rate of $\sqrt{n}$, and the second condition is just to ensure that the KL divergence is well defined for all large enough n. This condition does not, by any means, imply that the LCVB and NVB approximate posterior converges to Dirac delta distribution at the true parameter θ_0 as $n \to \infty$.

The primary result in this section shows that the loss-calibrated approximate posterior $q_a^*(\theta|\tilde{X}_n)$ for any $a \in \mathcal{A}$ is consistent and converges to the Dirac-delta distribution at θ_0. We establish the frequentist consistency of LCVB approximate posterior, extending and building on the results in [27].

Proposition 3.3.1. *Fix $a \in \mathcal{A}$. Then, under Assumptions 3.3.1, 3.3.2, 3.3.3, and 3.3.4*

$$q_a^*(\theta|\tilde{X}_n) \in \arg\min_{q \in \mathcal{Q}} \text{KL}\left(q(\theta) \middle\| \frac{R(a,\theta)\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta}\right) \Rightarrow \delta_{\theta_0} \text{ in } P_0^n - \text{probability as } n \to \infty. \tag{3.9}$$

Some comments are in order for this result. Recall that loss-calibration of the posterior distribution 'weights' it by the risk of taking decision a, $R(a,\theta)$. The optimization then finds the closest density functions in the family $\mathcal{Q}$ to this re-weighted posterior distribution. The posterior re-weighting has the effect of 'directing' the VB optimization to the most informative regions of the parameter sample space for the decision problem of interest. However, $R(a,\theta)$, which does not involve the data $\tilde{X}_n$, effectively serves to change the prior distribution, and in the limit, modulo our regularity assumptions, the consistency of the approximate posterior is to be anticipated. The proof of the proposition is presented in the appendix.

Since for a constant risk function $R(a,\theta)$, the LCVB approximate posterior $q_a^*(\theta|\tilde{X}_n)$ is same as NVB approximate posterior $q^*(\theta|\tilde{X}_n)$, we recover the result obtained in Theorem 5(1) of [27]. We rewrite the result as a corollary for completeness.

Corollary 3.3.1. *Under Assumptions 3.3.1, 3.3.3, and 3.3.4*

$$q^*(\theta|\tilde{X}_n) \in \arg\min_{q\in\mathcal{Q}} \mathrm{KL}\left(q(\theta)\middle\|\boldsymbol{\pi}(\theta|\tilde{X}_n)\right) \Rightarrow \delta_{\theta_0} \text{ in } P_0^n - \text{probability as } n\to\infty. \tag{3.10}$$

3.4 Consistency of Decisions

In this section we prove that the optimal decision estimated by the LCVB and NVB algorithms are consistent, in the sense that for almost every infinite sequence, the optimal decision rules $\mathtt{a}^*_{\mathtt{NV}}$ and $\mathtt{a}^*_{\mathtt{LC}}$ concentrate on the set of 'true' optimizers

$$A^* := \arg\min_{a\in\mathcal{A}} R(a,\theta_0) = \int \ell(y,a)p(y|\theta_0)dy.$$

For brevity, we define $H_q(a) := \mathbb{E}_q[R(a,\theta)]$ for any distribution $q(\cdot)$ on θ and $H_0(a) := R(a,\theta_0)$. We place a typical, but relatively strong condition on the decision space that

Assumption 3.4.1. *The decision space $\mathcal{A}$ is compact.*

Coupled with Assumption 3.3.2, this implies that the risk function is uniformly bounded in the decision space.

Now, suppose that the true posterior $\boldsymbol{\pi}(\theta|\tilde{X}_n))$ is in the set $\mathcal{Q}$. Then, the NVB approximate posterior in (3.4) $q^*(\theta|\tilde{X}_n)$ equals $\boldsymbol{\pi}(\theta|\tilde{X}_n)$, so that the empirical decision-rule $\mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n)$ coincides exactly with the Bayes optimal decision rule $\mathtt{a}^*(\tilde{X}_n)$. The consistency of the true posterior has been well-studied, and under Assumption 3.3.1 it is well known [110], [111] that for any neighborhood U of the true parameter θ_0

$$\boldsymbol{\pi}(U|\tilde{X}_n)) \to 1 \quad P_0 - a.s. \text{ as } n\to\infty, \tag{3.11}$$

where P_0 represents the true data-generation distribution. Then, it follows from Assumption 3.3.2 that

$$\sup_{a\in\mathcal{A}}\left|H_{\pi(\theta|\tilde{X}_n)}(a) - H_0(a)\right| \to 0 \; P_0 - a.s. \text{ as } n \to \infty. \tag{3.12}$$

It is straightforward to see that the limit result follows pointwise, and the uniform convergence result follows from the uniform boundedness of the loss functions. In the following section, we consider the typical case when the posterior $\pi(\theta|\tilde{X}_n)) \notin \mathcal{Q}$.

3.4.1 Analysis of the NVB Decision Rule

The first result of this section proves that the Bayes predictive loss, $H_{q^*}(a)$ is (uniformly) asymptotically consistent as the sample size grows. We relegate the proof to the appendix.

Proposition 3.4.1. *Under the assumptions stated above, we have*

$$\sup_{a\in\mathcal{A}}|H_{q^*}(a) - H_0(a)| \to 0 \quad in \quad P_0^n - probability \; as \; n \to \infty. \tag{3.13}$$

This proposition builds on [27, Theorem 5], which shows that modulo Assumptions 3.3.1, 3.3.3, and 3.3.4, the NVB approximate posterior distribution is asymptotically consistent(see Corollary 3.3.1). Using the consistency of $q^*(\theta|\tilde{X}_n)$ and Assumption 3.3.2(1), we first establish the pointwise convergence of $H_{q^*}(a)$ to $H_0(a)$. Then we argue, using continuity of the risk function $R(a,\theta)$ in a and the compactness of set $\mathcal{A}$, that uniform convergence follows.

A straightforward corollary of Proposition 3.4.1 implies that the optimal value $V_{q^*} := \min_{a\in\mathcal{A}} H_{q^*}(a)$ is asymptotically consistent as well; the proof is in the appendix.

Corollary 3.4.1. *Under Assumptions 3.3.1, 3.3.2, 3.3.3, 3.3.4, and 3.4.1, with* $V_0 := \min_{a\in\mathcal{A}} H_0(a)$, $|V_{q^*} - V_0| \to 0$ *in* P_0^n *– probability as* $n \to \infty$.

The primary question of interest is the asymptotic consistency of the optimal decision-rule $\mathtt{a}^*_{\mathtt{NV}}$. Our main result proves that in the large sample limit $\mathtt{a}^*_{\mathtt{NV}}$ is a subset of the true optimal decisions A^* for almost all samples $\tilde{X}_n$.

Proposition 3.4.2. *Under Assumptions 3.3.1, 3.3.2, 3.3.3, 3.3.4, and 3.4.1 , we have*

$$\left\{\mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n) \subseteq A^*\right\} \quad in \quad P_0^n - probability \ as \ n \to \infty. \tag{3.14}$$

We use the uniform convergence of $H_{q^*}(a)$ to $H_0(a)$ and argue that any decision which is not in the true optimal decision set A^*, must not exist in NVB approximate optimal decision set $\mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n)$ for large enough n. Once again, we relegate the proof to the appendix. Consequently, it follows that NVB optimal actions are asymptotically oracle regret minimizing:

Corollary 3.4.2. *Under Assumptions 3.3.1, 3.3.2, 3.3.3, 3.3.4, and 3.4.1 for any* $\mathbf{a} \in A^*$ *and* $a^* \in \mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n))$, $H_0(a^*) \to H_0(\mathbf{a})$ *in* P_0^n*-probability as* $n \to \infty$.

The result above is a straightforward implication of the continuity of $R(a, \theta_0)$ in a and Proposition 3.4.2 and therefore the proof is omitted.

3.4.2 Analysis of the LCVB decision rule

Now, recall from (3.6) that the LC decision-rule is

$$\mathtt{a}^*_{\mathtt{LC}}(\tilde{X}_n) = \arg\min_{a \in \mathcal{A}} \max_{q \in \mathcal{Q}} -\mathrm{KL}\left(q(\theta) \| \boldsymbol{\pi}(\theta|\tilde{X}_n)\right) + \int_\Theta q(\theta) \log R(a, \theta) d\theta.$$

The next proposition shows that $\mathtt{a}^*_{\mathtt{LC}}(\tilde{X}_n)$ is a subset of the true optimal decision set A^* in the large sample limit for almost all sample sequences. We use similar ideas as used in Section 3.4.1.

Proposition 3.4.3. *Under Assumptions 3.3.1, 3.3.2, 3.3.3, 3.3.4, and 3.4.1, we have*

$$\left\{\mathtt{a}^*_{\mathtt{LC}}(\tilde{X}_n) \subseteq A^*\right\} \ in \ P_0^n - probability \ as \ n \to \infty. \tag{3.15}$$

The proof is in the appendix. This result naturally implies that the loss-calibrated VB optimal decisions are also oracle regret minimizing

Corollary 3.4.3. *For any* $\mathbf{a} \in A^*$ *and* $a^{**} \in \mathtt{a}^*_{\mathtt{LC}}(\tilde{X}_n)$, $H_0(a^{**}) \to H_0(\mathbf{a})$ *as in* P_0^n *– probability as* $n \to \infty$.

3.5 Numerical Example

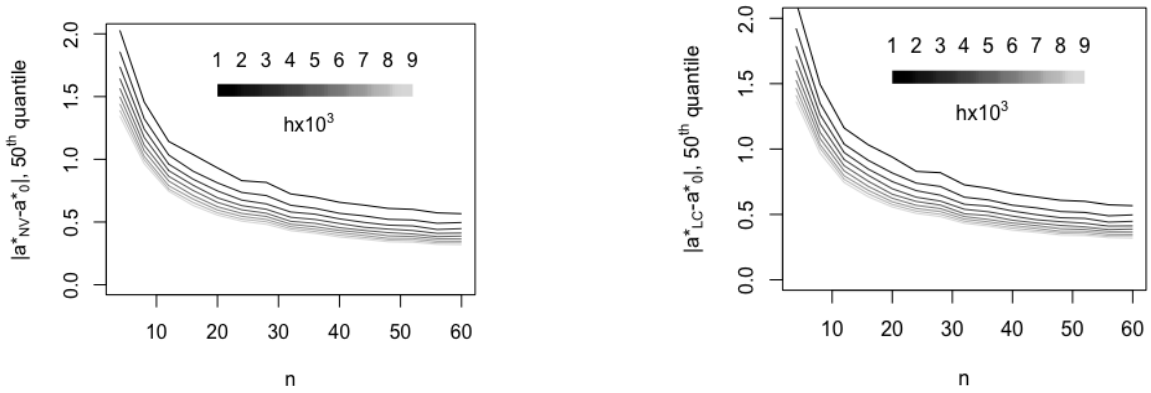
In this section we present a simulation study of a canonical optimal decision making problem called the newsvendor problem. This problem has been extensively studied in the inventory management literature [7]–[9]. Recall that the newsvendor loss function is defined as

$$\ell(a, \xi) := h(a - \xi)^+ + b(\xi - a)^+$$

where $\xi \in [0, \infty)$ is the random demand, a is the inventory or decision variable, and h and b are given positive constants. We assume that the decision variable a take values in a compact decision space $\mathcal{A}$. We also assume that the the random demand ξ is exponentially distributed with unknown rate parameter $\theta_0 \in (0, \infty)$. The model risk can easily be derived as

$$R(a, \theta) = \mathbb{E}_{P_\theta}[\ell(a, \xi)] = ha - \frac{h}{\theta} + (b + h)\frac{\mathrm{e}^{-a\theta}}{\theta}, \tag{3.16}$$

which is convex in a. Let $\tilde{X}_n := \{\xi_1, \xi_2 \dots \xi_n\}$ be n observations of the random demand, assumed to be independent and identically distributed. Next, we posit a non-conjugate inverse-gamma prior distribution over the rate parameter θ with shape and rate parameter α and β respectively. Finally, we run a simulation experiment using the newsvendor model described above for a fix $\theta_0 = 0.68$, $b = 0.1, \alpha = 1$, and $\beta = 4.1$. We use naive VB and LCVB algorithms to obtain the respective optimal decision $\mathtt{a}^*_{\mathtt{NV}}$ and $\mathtt{a}^*_{\mathtt{LC}}$ for 9 different values of $h \in \{0.001, 0.002, \dots 0.009\}$ and repeat the experiment over 1000 sample paths. In Figure 1, we plot the 50^{th} quantile of the $|\mathtt{a}^* - \mathtt{a}^*_0|$, where $\mathtt{a}^* \in \{\mathtt{a}^*_{\mathtt{NV}}, \mathtt{a}^*_{\mathtt{LC}}\}$ for this model. Observe that the optimality gap decreases quite rapidly for both the naive VB (left) and the loss-calibrated VB (right) methods.



(a) Naive VB. (b) LCVB.

Figure 3.1. Optimality gap in decisions (the 50^{th} quantile over 1000 sample paths) against the number of samples (n) for $\mathtt{a}^*_{\mathtt{NV}}$ (left) and $\mathtt{a}^*_{\mathtt{LC}}$ (right).

3.6 Proofs

Proof of Proposition 3.3.1

Lemma 3.6.1. *For any risk function $R(a,\theta)$ that satisfies Assumption 3.3.2 and a given sequence of distributions $\{q_n(\theta)\}$ that converges weakly to any distribution $q(\theta)$ other than the Dirac-delta distribution at θ_0, the* $\mathrm{KL}\left(q_n(\theta)\middle\|\frac{R(a,\theta)\pi(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\pi(\theta)p(\tilde{X}_n|\theta)d\theta}\right)$ *is undefined in the limit as $n\to\infty$ $P_0-a.s.$*

Proof. Using the definition of the posterior distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n)=\frac{\pi(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta \pi(\theta)p(\tilde{X}_n|\theta)d\theta}$, first observe that

$$\begin{aligned}
&\mathrm{KL}\left(q_n(\theta)\middle\|\frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}\right)\\
&\quad=\mathrm{KL}\left(q_n(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)\right)-\int_\Theta \log(R(a,\theta))q_n(\theta)d\theta-\log\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta.\\
&\quad\geq\mathrm{KL}\left(q_n(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)\right)-\int_\Theta R(a,\theta)q_n(\theta)d\theta-\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta, \qquad (3.17)
\end{aligned}$$

where the last inequality uses the fact that $\log x < x$. Now taking the lim inf on either side, we have

$$\begin{aligned}
&\liminf_{n\to\infty} \mathrm{KL}\left(q_n(\theta)\middle\| \frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}\right) \\
&\geq \liminf_{n\to\infty} \mathrm{KL}\left(q_n(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)\right) - \limsup_{n\to\infty}\int_\Theta R(a,\theta)q_n(\theta)d\theta - \limsup_{n\to\infty}\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta.
\end{aligned} \tag{3.18}$$

Recall that the posterior distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ converges weakly to δ_{θ_0} $P_0 - a.s.$ Due to [112, Theorem 16] we know that $\mathrm{KL}(q(\theta)\|p(\theta))$ is a lower semi-continuous function of the pair $(q(\theta), p(\theta))$ in the weak topology on the space of probability measures. Using lower semi-continuity, it follows that the first term in (3.18) satisfies

$$\liminf_{n\to\infty} \mathrm{KL}\left(q_n(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)\right) > \mathrm{KL}(q(\theta)\|\delta_{\theta_0}) = \infty, \tag{3.19}$$

where the last equality is by definition of the KL divergence, since $q(\theta) \neq \delta_{\theta_0}$ (as $q_n(\theta)$ does not weakly converge to δ_{θ_0}) and therefore it is not absolutely continuous with respect to δ_{θ_0}. Since the last two terms are finite due to Assumption 3.3.2, we have shown that for any sequence of distribution $\{q_n(\theta)\}$ that converges weakly to any distribution $q(\theta) \neq \delta_{\theta_0}$, the $\mathrm{KL}\left(q_n(\theta)\middle\| \frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}\right)$ diverges in the limit as $n \to \infty$ $P_0 - a.s.$ □

Lemma 3.6.2. *Let $\{K_n\} \subseteq \Theta$ be a sequence of compact balls such that for all $n \geq 1$, $\theta_0 \in K_n$ and $K_n \to \Theta$ as $n \to \infty$. Then, under Assumption 3.4.1 and for any $\delta > 0$, the sequence of random variables $\left\{\int_{\Theta\setminus K_n} \boldsymbol{\pi}(\theta)R(a,\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta\right\}$ is of order $o_{P_0^n}(1)$; that is*

$$\lim_{n\to\infty} P_0^n\left(\int_{\Theta\setminus K_n} \boldsymbol{\pi}(\theta)R(a,\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta > \delta\right) = 0.$$

Proof. Using Markov's inequality, it follows that,

$$P_0^n\left(\int_{\Theta\setminus K_n} \boldsymbol{\pi}(\theta)R(a,\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta > \delta\right) \leq \frac{1}{\delta}\mathbb{E}_{P_0^n}\left[\int_{\Theta\setminus K_n} \boldsymbol{\pi}(\theta)R(a,\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta\right]. \tag{3.20}$$

Next, using Fubini's Theorem in the RHS above and then the fact that $\mathbb{E}_{P_0^n}\left[\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)\right] \leq 1$, observe that

$$P_0^n\left(\int_{\Theta\backslash K_n} \boldsymbol{\pi}(\theta)R(a,\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta > \delta\right) \leq \frac{1}{\delta}\int_{\Theta\backslash K_n} \boldsymbol{\pi}(\theta)R(a,\theta)d\theta. \tag{3.21}$$

Since $\theta_0 \notin \Theta\backslash K_n$ for all $n \geq 1$ and $\Theta\backslash K_n \to \emptyset$ as $n \to \infty$, $\mathbb{1}_{\Theta\backslash K_n}R(a,\theta)$ is monotonic, and therefore using the monotone convergence theorem, $\int_{\Theta\backslash K_n} R(a,\theta)\boldsymbol{\pi}(\theta)d\theta \to 0$ as $n \to \infty$. Hence, taking limits on either side of (3.21) the result follows. □

Next, we show that for fixed $a \in \mathcal{A}$, the KL divergence between the LC approximate posterior $q_a^*(\theta|\tilde{X}_n)$ and the rescaled posterior $\frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}$ is finite in the limit. Also, the following lemma uses similar proof techniques as used in [27].

Lemma 3.6.3. *Fix $a \in \mathcal{A}$. Then, under Assumptions 3.3.1, 3.3.2, 3.3.3, and 3.3.4,*

$$\lim_{n\to\infty} P_0^n\left(\min_{q\in\mathcal{Q}} \mathrm{KL}\left(q(\theta)\middle\|\frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}\right) < \infty\right) = 1.$$

Furthermore, the LC variational posterior $q_a^*(\theta|\tilde{X}_n)$ can converge only at the rate of $\sqrt{n}$.

Proof. Following Assumption 3.3.4 there exists a sequence of distributions $\{q_n(\theta)\} \in \mathcal{Q}$ that converges to δ_{θ_0} at the rate of $\gamma_n = \sqrt{n}$. Specifically, we consider the sequence where $q_n(\theta)$ has mean $\hat{\theta}_n$, the maximum likelihood estimate. It suffices to show that for such sequence $\{q_n(\theta)\} \subset \mathcal{Q}$,

$$\limsup_{n\to\infty} \mathrm{KL}\left(q_n(\theta)\middle\|\frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}\right) = \mathrm{KL}\left(q_n(\theta)\middle\|\frac{R(a,\theta)\boldsymbol{\pi}(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta}\right) < \infty.$$

For brevity let us denote $\mathrm{KL}\left(q_n(\theta)\middle\|\frac{R(a,\theta)\pi(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)}{\int_\Theta R(a,\theta)\pi(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta}\right)$ as KL. First, observe that for a compact set $K \subset \Theta$ containing the true parameter θ_0, we have

$$\begin{aligned}\mathrm{KL} = \int_\Theta q_n(\theta)\log(q_n(\theta))d\theta - \int_\Theta q_n(\theta)\log(R(a,\theta)\pi(\theta))d\theta - \int_K q_n(\theta)\log\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta \\ - \int_{\Theta\setminus K} q_n(\theta)\log\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta + \log\left(\int_\Theta R(a,\theta)\pi(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta\right). \quad (3.22)\end{aligned}$$

Now we approximate $\int_K q_n(\theta)\log\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta$ using the LAN condition in Assumption 3.3.3. Let $\Delta_{n,\theta_0} := \sqrt{n}(\hat{\theta}_n - \theta_0)$, and reparameterizing the expression with $\theta = \theta_0 + n^{-1/2}h$ and denoting K as the reparameterized set K we have

$$\begin{aligned}&\int_K q_n(\theta)\log\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta \\ &= n^{-1/2}\int_K q_n(\theta_0 + n^{-1/2}h)\log\left(\frac{p(\tilde{X}_n|\theta_0 + n^{-1/2}h)}{p(\tilde{X}_n|\theta_0)}\right)dh \quad (3.23)\\ &= n^{-1/2}\int_K q_n(\theta_0 + n^{-1/2}h)\left(hI(\theta_0)\Delta_{n,\theta_0} - \frac{1}{2}h^2 I(\theta_0) + o_{P_0^n}(1)\right)dh \\ &= \left(o_{P_0^n}(1)\right)\int_K q_n(\theta)d\theta + \int_K q_n(\theta)\left(\sqrt{n}(\theta-\theta_0)I(\theta_0)\Delta_{n,\theta_0} - \frac{1}{2}n(\theta-\theta_0)^2 I(\theta_0)\right)d\theta \\ &= \left(\frac{1}{2}nI(\theta_0)(\hat{\theta}_n - \theta_0)^2 + o_{P_0^n}(1)\right)\int_K q_n(\theta)d\theta - \int_K \frac{1}{2}nI(\theta_0)q_n(\theta)(\theta - \hat{\theta}_n)^2 d\theta. \quad (3.24)\end{aligned}$$

Now consider the last term in (3.22). Let $\{K_n\} \subseteq \Theta$ be a compact sequence of balls such that for all $n \geq 1$, $\theta_0 \in K_n$ and $K_n \to \Theta$ as $n \to \infty$. Next, using the same re-parametrization we obtain,

$$\int_{K_n} R(a,\theta)\pi(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta = \mathrm{e}^{o_{P_0^n}(1)}\mathrm{e}^{\frac{nI(\theta_0)}{2}(\hat{\theta}_n-\theta_0)^2}\int_{K_n} R(a,\theta)\pi(\theta)\mathrm{e}^{-\frac{nI(\theta_0)}{2}(\theta-\hat{\theta}_n)^2}d\theta. \quad (3.25)$$

Now, Lemma 3.6.2 implies that

$$\int_{\Theta\setminus K_n} R(a,\theta)\pi(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta = o_{P_0^n}(1). \quad (3.26)$$

Hence, by the results in (3.25) and (3.26), the last term in (3.22) satisfies

$$\begin{aligned}
&\log\left(\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta\right)\\
&=\log\left(\int_{K_n} R(a,\theta)\boldsymbol{\pi}(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta+\int_{\Theta\setminus K_n} R(a,\theta)\boldsymbol{\pi}(\theta)\left(\frac{p(\tilde{X}_n|\theta)}{p(\tilde{X}_n|\theta_0)}\right)d\theta\right)\\
&\sim\frac{nI(\theta_0)}{2}(\hat{\theta}_n-\theta_0)+\log\int_{K_n} R(a,\theta)\boldsymbol{\pi}(\theta)\mathrm{e}^{-\frac{nI(\theta_0)}{2}(\theta-\hat{\theta}_n)^2}d\theta+o_{P_0^n}(1),
\end{aligned}\tag{3.27}$$

where $a_n\sim b_n$ implies that $\lim_{n\to\infty}\frac{a_n}{b_n}=1$. Now, by substituting (3.24) and (3.27) into (3.22) we obtain,

$$\begin{aligned}
\mathrm{KL}\sim&\int_\Theta q_n(\theta)\log q_n(\theta)d\theta-\int_\Theta q_n(\theta)\log(R(a,\theta)\boldsymbol{\pi}(\theta))d\theta\\
&+\left(\frac{1}{2}nI(\theta_0)(\hat{\theta}_n-\theta_0)^2+o_{P_0^n}(1)\right)\left[1-\int_K q_n(\theta)d\theta\right]\\
&+\log\int_{K_n} R(a,\theta)\boldsymbol{\pi}(\theta)\mathrm{e}^{-\frac{nI(\theta_0)}{2}(\theta-\hat{\theta}_n)^2}d\theta+\frac{1}{2}nI(\theta_0)\int_K(\theta-\hat{\theta}_n)^2q_n(\theta)d\theta+o_{P_0^n}(1).
\end{aligned}$$

Since, the $q_n(\theta)\Rightarrow\delta_{\theta_0}$ as $n\to\infty$ and $\theta_0\in K$,

$$\left(\frac{1}{2}nI(\theta_0)(\hat{\theta}_n-\theta_0)^2+o_{P_0^n}(1)\right)\left[1-\int_K q_n(\theta)d\theta\right]\sim o_{P_0^n}(1),$$

implying that,

$$\begin{aligned}
\mathrm{KL}\sim&\int_\Theta q_n(\theta)\log q_n(\theta)d\theta-\int_\Theta\log(R(a,\theta)\boldsymbol{\pi}(\theta))q_n(\theta)d\theta-\frac{1}{2}\log n+\frac{1}{2}\log\left(\frac{2\boldsymbol{\pi}}{I(\theta_0)}\right)\\
&+\log\int_{K_n}(R(a,\theta)\boldsymbol{\pi}(\theta))\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta+\frac{1}{2}nI(\theta_0)\int_K(\theta-\hat{\theta}_n)^2q_n(\theta)d\theta+o_{P_0^n}(1),
\end{aligned}\tag{3.28}$$

where $\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})$ represents the Gaussian density function. Since $q_n(\theta)$ has mean $\hat{\theta}_n$ and rate of convergence $\sqrt{n}$, then by a change of variable to $\mu=\sqrt{n}(\theta-\hat{\theta}_n)$

$$\int_\Theta q_n(\theta)\log q_n(\theta)d\theta=\frac{1}{\sqrt{n}}\int q_n\left(\frac{\mu}{\sqrt{n}}+\hat{\theta}_n\right)\log q_n\left(\frac{\mu}{\sqrt{n}}+\hat{\theta}_n\right)d\mu=\frac{1}{2}\log n+\int\breve{q}_n(\mu)\log\breve{q}_n(\mu)d\mu,\tag{3.29}$$

where $\check{q}_n(\mu)$ is the rescaled density as defined in Definition 3.3. Substituting (3.29) into (3.28), we obtain

$$\begin{aligned}\mathrm{KL} \sim \int \check{q}_n(\mu) \log \check{q}_n(\mu) d\mu - \int_\Theta \log(R(a,\theta)\boldsymbol{\pi}(\theta)) q_n(\theta) d\theta + \frac{1}{2} \log\left(\frac{2\boldsymbol{\pi}}{I(\theta_0)}\right) \\ + \log \int_{K_n} (R(a,\theta)\boldsymbol{\pi}(\theta)) \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta + \frac{1}{2} nI(\theta_0) \int_K (\theta - \hat{\theta}_n)^2 q_n(\theta) d\theta + o_{P_0^n}(1).\end{aligned} \tag{3.30}$$

Since, $\frac{1}{2} nI(\theta_0) \int_K (\theta - \hat{\theta}_n)^2 q_n(\theta) d\theta \leq \frac{1}{2} nI(\theta_0) \int_\Theta (\theta - \hat{\theta}_n)^2 q_n(\theta) d\theta \leq \frac{1}{2} I(\theta_0)$, due to the specific choice of $q_n(\theta)$ with variance $O(n^{-1})$(see Definition 3.3.1), it follows from (3.30) that for large enough n,

$$\begin{aligned}\mathrm{KL} \lesssim \int \check{q}_n(\mu) \log \check{q}_n(\mu) d\mu - \int_\Theta \log(\boldsymbol{\pi}(\theta)) q_n(\theta) d\theta - \int_\Theta \log(R(a,\theta)) q_n(\theta) d\theta + \frac{1}{2} \log\left(\frac{2\boldsymbol{\pi}}{I(\theta_0)}\right) \\ + \log \int_{K_n} (R(a,\theta)\boldsymbol{\pi}(\theta)) \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta + \frac{1}{2} I(\theta_0) + o_{P_0^n}(1).\end{aligned} \tag{3.31}$$

Now take limsup on either side of the above equation. Observe that the first term is finite by Assumption 3.3.4. The second term is finite since the prior distribution is continuous in θ bounded due to Assumption 3.3.1, therefore using the the defintion of weak convergence, $\lim_{n\to\infty} \int_\Theta \log(\boldsymbol{\pi}(\theta)) q_n(\theta) d\theta = \log \boldsymbol{\pi}(\theta_0)$ and $\boldsymbol{\pi}(\theta)$ is non-zero in the neighbourhood of θ_0. The third term is bounded by $-\log(W)$ due to Assumption 3.3.2(1). Since, $\theta_0 \in K_n \forall n \geq 1$, the fifth term is bounded by Laplace's approximation,

$$\int_{K_n} (R(a,\theta)\boldsymbol{\pi}(\theta)) \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta \sim R(a,\theta_0)\boldsymbol{\pi}(\theta_0).$$

Since the last term is finite in P_0^n-probability, therefore it follows that,

$$\begin{aligned}\lim_{n\to\infty} P_0^n \left(\min_{q\in\mathcal{Q}} \mathrm{KL}\left(q(\theta) \middle\| \frac{R(a,\theta)\boldsymbol{\pi}(\theta) p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta) p(\tilde{X}_n|\theta) d\theta}\right) < \infty \right) \\ \geq \lim_{n\to\infty} P_0^n \left(\mathrm{KL}\left(q_n(\theta) \middle\| \frac{R(a,\theta)\boldsymbol{\pi}(\theta) p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta) p(\tilde{X}_n|\theta) d\theta}\right) < \infty \right) = 1.\end{aligned}$$

□

Proof of Proposition 3.3.1. Recall from the Lemma 3.6.1 that for any risk function $R(a,\theta)$ that satisfies Assumption 3.3.2 and for a given sequence of distributions $\{q_n(\theta)\}$ that converges weakly to any distribution $q(\theta)$ other than δ_{θ_0}, $\mathrm{KL}\left(q_n(\theta)\middle\|\frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}\right)$ diverges as $n\to\infty$ $P_0-a.s.$ On the other hand, Lemma 3.6.3 shows that for any $a\in\mathcal{A}$,

$$\begin{aligned}\lim_{n\to\infty} P_0^n\left(\min_{q\in\mathcal{Q}}\mathrm{KL}\left(q(\theta)\middle\|\frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}\right)<\infty\right)\\ \geq \lim_{n\to\infty} P_0^n\left(\mathrm{KL}\left(q_n(\theta)\middle\|\frac{R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)}{\int_\Theta R(a,\theta)\boldsymbol{\pi}(\theta)p(\tilde{X}_n|\theta)d\theta}\right)<\infty\right)=1.\end{aligned}$$

Therefore, Lemma 3.6.1 and 3.6.3 combined together imply that for any $a\in\mathcal{A}$, and for any risk function $R(a,\theta)$ that satisfies Assumption 3.3.2, the LC approximate posterior must converge weakly to δ_{θ_0} in P_0^n-probability as $n\to\infty$; that is $q_a^*(\theta|\tilde{X}_n)\Rightarrow\delta_{\theta_0}$ in P_0^n- probability as $n\to\infty$. □

Proof of Proposition 3.4.1

Proof. First, we establish point-wise convergence using similar ideas as used in the proof of [113, Theorem 3.7]. Fix $a\in\mathcal{A}$. Due to Assumption 3.3.2(3), $R(a,\theta)$ is uniformly integrable with respect to any $q\in\mathcal{Q}$, which implies that for $q^*(\theta|\tilde{X}_n)$ and for any $\epsilon>0$, there exists a compact set K_ϵ such that for all $n\geq 1$ $\int_{\Theta\setminus K_\epsilon}|R(a,\theta)|q^*(\theta|\tilde{X}_n)d\theta<\epsilon$.

Now fix $\gamma_\epsilon:=\max_{\theta\in K_\epsilon}|R(a,\theta)|$. Note that $\gamma_\epsilon<+\infty$, since K_ϵ is compact and $R(a,\cdot)$ is a continuous mapping for any $a\in\mathcal{A}$. Define $R_\epsilon(a,\theta)$ be the truncation of $R(a,\theta)$, that is

$$R_\epsilon(a,\theta)=\begin{cases}R(a,\theta) & \text{if } |R(a,\theta)|<\gamma_\epsilon\\ \gamma_\epsilon & \text{if } R(a,\theta)>\gamma_\epsilon\\ -\gamma_\epsilon & \text{if } R(a,\theta)<-\gamma_\epsilon.\end{cases}\tag{3.32}$$

It follows from the definition above that $|R_\epsilon(a,\theta)|\leq|R(a,\theta)|$, which implies that

$$\int_{\Theta\setminus K_\epsilon}|R_\epsilon(a,\theta)|q^*(\theta|\tilde{X}_n)d\theta<\epsilon\tag{3.33}$$

Note the $R_\epsilon(a,\theta)$ is bounded and continuous in θ, therefore, it follows using the definition of weak convergence and Corollary 3.3.1 that

$$\lim_{n\to\infty} \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] \overset{P_0^n}{=} R_\epsilon(a,\theta_0). \tag{3.34}$$

Next observe that

$$\begin{aligned}
&|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)| \\
&\quad= \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] + \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0) \\
&\qquad\qquad + R_\epsilon(a,\theta_0) - R(a,\theta_0)\Big| \\
&\quad\le \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)]\Big| + \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0)\Big| \\
&\qquad\qquad + |R_\epsilon(a,\theta_0) - R(a,\theta_0)| \\
&\quad= \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)]\Big| + \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0)\Big| \\
&\qquad\qquad + |R_\epsilon(a,\theta_0) - R(a,\theta_0)|\,.
\end{aligned} \tag{3.35}$$

Now using the definition of $R_\epsilon(a,\theta)$ note that

$$\begin{aligned}
\Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)]\Big| &= \left|\int_{\Theta\setminus K_\epsilon} (R(a,\theta) - R_\epsilon(a,\theta)) q^*(\theta|\tilde{X}_n) d\theta\right| \\
&\le \int_{\Theta\setminus K_\epsilon} |R(a,\theta)| q^*(\theta|\tilde{X}_n) d\theta + \int_{\Theta\setminus K_\epsilon} |R_\epsilon(a,\theta)| q^*(\theta|\tilde{X}_n) d\theta \quad \le 2\epsilon.
\end{aligned}$$

Similarly, $|R_\epsilon(a,\theta_0) - R(a,\theta_0)| \le 2\epsilon$, since due to Assumption 3.3.2(4) $\int_{\Theta\setminus K_\epsilon} |R(a,\theta)| q^*(\theta|\tilde{X}_n) d\theta < \epsilon$ is true for all $n \ge 1$ and consequently for δ_{θ_0} as well. Hence, substituting the above two observations into (3.35) yields

$$|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)| \le 4\epsilon + \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0)\Big|\,.$$

Consequently, it follows for any $\epsilon > 0$ that,

$$P_0^n\left(|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)| > 5\epsilon\right) \le P_0^n\left(|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0)| > \epsilon\right). \tag{3.36}$$

Now taking limits $n \to \infty$ on either side of the inequality above, the result follows straightforwardly using the observation in (3.34).

Since $\mathcal{A}$ is compact and $R(a, \theta_0)$ is continuous in a, using Corollary 2.2 in [114] the uniform convergence follows from point-wise convergence if there exist a bounded sequence B_n and for all $a_1, a_2 \in \mathcal{A}$, $|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a_1, \theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a_2, \theta)]| \leq B_n \|a_1 - a_2\|$. Since, $R(a, \theta)$ is locally Lipschitz in a due to Assumption 3.3.2(3), therefore for $a_1, a_2 \in \mathcal{A}$,

$$\begin{aligned} |\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a_1, \theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a_2, \theta)]| &\leq \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[|R(a_1, \theta) - R(a_2, \theta)|] \\ &\leq \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[K_{\mathcal{A}}(\theta)] \|a_1 - a_2\|. \end{aligned} \tag{3.37}$$

The uniform convergence follows since by Assumption 3.3.2(3) $\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[K_{\mathcal{A}}(\theta)] \leq \bar{K}_{\mathcal{A}}$. □

Proof of Corollary 3.4.1

Proof. Let $a_q \in \mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n)$ and $a_0 \in A^*$ then, by definition, $V_{q^*} = H_{q^*}(a_q)$ and $V_0 = H_0(a_0)$. Then,

$$V_{q^*} - V_0 = [H_{q^*}(a_q) - H_0(a_0)] \leq [H_{q^*}(a_0) - H_0(a_0)] \leq \sup_{a \in \mathcal{A}} |H_{q^*}(a) - H_0(a)|. \tag{3.38}$$

On the other hand, observe that

$$V_{q^*} - V_0 \geq [H_{q^*}(a_q) - H_0(a_q)] \geq -|H_{q^*}(a_q) - H_0(a_q)| \geq -\sup_{a \in \mathcal{A}} |H_{q^*}(a) - H_0(a)|. \tag{3.39}$$

Therefore from (3.38), (3.39), and Proposition 3.4.1, it follows that

$$\lim_{n \to \infty} |V_{q^*} - V_0| \leq \lim_{n \to \infty} \sup_{a \in \mathcal{A}} |H_q(a, X) - H_0(a)| \overset{P_0^n}{=} 0 \text{ in},$$

and the result follows. □

Proof of Proposition 3.4.2

Proof. Equivalently, we can show that $a \in \mathcal{A} \setminus A^*$ implies that $a \notin \mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n)$ in P_0^n − probability as $n \to \infty$. Fix $a \in \mathcal{A} \setminus A^*$, then we have $H_0(a) > V_0$. Next define $\epsilon := \inf_{a \in \mathcal{A} \setminus A^*} H_0(a) - V_0$. Using Proposition 3.4.1, for any $\delta > 0$, there exists an $n_0 \geq 1$ (depending on ϵ) such that $\forall n \geq n_0$, $P_0^n\left(|V_{q^*} - V_0| < \frac{\epsilon}{2}\right) \geq P_0^n\left(\sup_{a \in \mathcal{A}} |H_{q^*}(a) - H_0(a)| < \frac{\epsilon}{2}\right) \geq 1 - \delta$. Therefore, we have $P_0^n\left(V_{q^*} < V_0 + \frac{\epsilon}{2}\right) \geq 1 - \delta$ for all $n \geq n_0$. Using the definition of ϵ and Proposition 3.4.1, it also follows that for any $a \in \mathcal{A} \setminus A^*$ and for all $n \geq n_0$

$$\begin{aligned} P_0^n\left(V_0 + \epsilon < H_{q^*}(a) + \frac{\epsilon}{2}\right) &= P_0^n\left(\inf_{a \in \mathcal{A} \setminus A^*} H_0(a) < H_{q^*}(a) + \frac{\epsilon}{2}\right) \\ &\geq P_0^n\left(H_0(a) < H_{q^*}(a) + \frac{\epsilon}{2}\right) \\ &\geq P_0^n\left(\sup_{a \in \mathcal{A}} |H_{q^*}(a) - H_0(a)| < \frac{\epsilon}{2}\right) \\ &\geq 1 - \delta. \end{aligned}$$

Therefore for any $a \in \mathcal{A} \setminus A^*$,

$$P_0^n(a \notin \mathtt{a}^*_{\mathtt{NV}}(\tilde{X}_n)) \geq P_0^n\left(V_{q^*} < H_{q^*}(a)\right) \geq P_0^n\left(\left\{V_0 + \frac{\epsilon}{2} < H_{q^*}(a)\right\} \cap \left\{V_{q^*} < V_0 + \frac{\epsilon}{2}\right\}\right) \geq 1 - \delta.$$

Hence the proposition follows. □

Proof of Proposition 3.4.3

Proof. Fix $\bar{a} \in \mathcal{A}$ and recall from (3.6) that

$$\mathcal{F}(a, q; \tilde{X}_n) = -\mathrm{KL}(q(\theta) \| \boldsymbol{\pi}(\theta | \tilde{X}_n)) + \int_\Theta \log R(a, \theta) q(\theta) d\theta.$$

Also recall that the LC approximate posterior $q^*_{\bar{a}}(\theta|\tilde{X}_n)$ converges weakly to a Dirac delta distribution at θ_0 due to Propostion 3.3.1. It now follows that, due to Assumption 3.3.2 (2) on $R(a,\theta)$ and using the definition of weak convergence

$$\lim_{n\to\infty} \int_\Theta \log R(a,\theta) q^*_{\bar{a}}(\theta|\tilde{X}_n) d\theta \overset{P_0^n}{=} \log R(a,\theta_0) \ \forall a \in \mathcal{A}. \tag{3.40}$$

Now since the set $\mathcal{A}$ is compact, logarithm function is continuous, and $R(a,\theta)$ is continuous in $\forall a \in \mathcal{A}$, it follows using similar arguments as used in Proposition 3.4.1 that for any $\bar{a} \in \mathcal{A}$,

$$\sup_{a\in\mathcal{A}} \left| \int_\Theta \log R(a,\theta) q^*_{\bar{a}}(\theta|\tilde{X}_n) d\theta - \log R(a,\theta_0) \right| \overset{P_0^n}{\to} 0 \ \text{ as } n \to \infty. \tag{3.41}$$

Now again using similar arguments as in Proposition 3.4.2 and monotonicity of logarithm function, we can show that the LC approximate decision rule for any $\bar{a} \in \mathcal{A}$, that is

$$\mathsf{a}^*_{\mathsf{LC}}(\tilde{X}_n, \bar{a}) := \text{argmin}_{a\in\mathcal{A}} \int_\Theta \log R(a,\theta) q^*_{\bar{a}}(\theta|\tilde{X}_n) d\theta$$

is subset of the true decision set A^* in $P_0^n -$ probability as $n \to \infty$. Since the result is true for any $\bar{a} \in \mathcal{A}$, it is true for any a that lies in LC approximate decision set $\mathsf{a}^*_{\mathsf{LC}}$ and therefore the proposition follows. □

4. BAYESIAN JOINT CHANCE CONSTRAINED OPTIMIZATION

In this chapter, we consider data-driven chance-constrained stochastic optimization problems in a Bayesian framework. Bayesian posteriors afford a principled mechanism to incorporate data and prior knowledge into stochastic optimization problems. However, the computation of Bayesian posteriors is typically an intractable problem, and has spawned a large literature on approximate Bayesian computation. Here, in the context of chance-constrained optimization, we focus on the question of statistical consistency (in an appropriate sense) of the optimal value, computed using an approximate posterior distribution. To this end, we rigorously prove a frequentist consistency result demonstrating the convergence of the optimal value to the optimal value of a fixed, parameterized constrained optimization problem. We augment this by also establishing a probabilistic rate of convergence of the optimal value. We also prove the convex feasibility of the approximate Bayesian stochastic optimization problem. Finally, we demonstrate the utility of our approach on an optimal staffing problem for an M/M/c queueing model.

4.1 Introduction

Consider a constrained optimization problem,

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & R(a, \theta_0) \\ \text{s.t.} \quad & g_{\mathrm{i}}(a, \theta_0) \leq 0, \ \mathrm{i} \in \{1, 2, 3, \ldots, m\}, \end{aligned} \tag{TP}$$

where $a \in \mathcal{A} \subseteq \mathbb{R}^p$ is a decision vector in some convex set $\mathcal{A}$ and $\theta_0 \in \mathbb{R}^q$ parametrizes the problem. The function $R : \mathcal{A} \times \mathbb{R}^q \mapsto \mathbb{R}$ encodes the cost/risk and the functions $g_{\mathrm{i}} : \mathcal{A} \times \mathbb{R}^q \mapsto \mathbb{R}$ define the constraints. We assume that such a *nominal* optimization problem and its solution(s) exists, under suitable regularity conditions.

In practice, the parameter is often unknown beyond lying in some set $\Theta \subseteq \mathbb{R}^q$. It is natural, therefore, to assume the existence of a probability distribution $P(\cdot)$ with support

Θ that quantifies the decision-maker's (DM) epistemic uncertainty about the parameter, leading to a *joint chance constrained* optimization problem

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & E_P[R(a,\theta)] \\ \text{s.t.} \quad & P\left(g_{\mathrm{i}}(a,\theta)\leq 0,\ \mathrm{i}\in\{1,2,3,\ldots,m\}\right)\geq\beta. \end{aligned} \tag{JCCP}$$

Note that a solution to (JCCP) is feasible for (TP) with probability at least β. Joint chance constrained problems have been used extensively to model a range of constrained optimization problems with parametric uncertainty [2], [115].

In this chapter we are interested in data-driven settings where only a dataset of n samples – so-called 'covariates' – is available, and whose joint distribution $P^n_{\theta_0}(\cdot)$ depends on the 'true' parameter θ_0. For instance, consider a staffing problem in a queueing system, where the goal is to compute the minimal number of servers required to ensure, with high probability, that the typical customer applying for service waits no more than a fixed amount of time to be served. The waiting time distribution for the typical customer depends on the arrival and service rates, which are unknown in a data-driven setting. Datasets here might include waiting times, inter-arrival and service times, whose distributions depend on the (unknown) rates. Problems of this type are prevalent across operations management [18], [44], [116], finance [117], and engineering [118].

In this data-driven setting, one might expect the epistemic uncertainty to diminish with an increasing number of samples, with each additional sample providing 'new information' about the true parameter θ_0. Bayesian methods provide a coherent way to quantify the devolution of the epistemic uncertainty through a *posterior* density $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ over the parameters $\theta\in\Theta$.The latter is computed by combining a *prior* density, quantifying *a priori* information (and biases) about the parameters, and a likelihood function, quantifying the probability of observed data under any parameter θ. Specifically, from Bayes' formula, it is well known that

$$\boldsymbol{\pi}(\theta|\tilde{X}_n)=\frac{p^n_\theta(\tilde{X}_n)\boldsymbol{\pi}(\theta)}{\int p^n_\theta(\tilde{X}_n)\boldsymbol{\pi}(\theta)d\theta}, \tag{4.1}$$

where $\boldsymbol{\pi}(\theta)$ is the prior density, $p_\theta^n(\tilde{X}_n)$ is the likelihood of observing $\tilde{X}_n$, and the denominator is the so-called data evidence. Bayesian methods have the advantage of calibrating uncertainty about hidden variables given partial observations. Further, in many applications, incorporating prior knowledge is preferable to straight empirics. For example, in the queueing system design problem the prior distribution maybe specified by a modeler based on expert input and require that the arrival rate be strictly less than the total system capacity (ensuring that the system is stochastically stable). Of course, in the absence of such knowledge, uninformative priors (such as Jeffrey's prior or uniform priors) can be used, but the same calculus holds.

this chapter focuses on the formulation of a *Bayesian joint chance constrained program* (BJCCP) model, wherein a posterior distribution is used as the measure of epistemic uncertainty in (JCCP) to obtain,

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & E_{\boldsymbol{\pi}(\theta|\tilde{X}_n)}[R(a,\theta)] \\ \text{s.t.} \quad & \Pi\left(g_{\mathrm{i}}(a,\theta)\leq 0,\ \mathrm{i}\in\{1,2,3,\ldots,m\}|\tilde{X}_n\right)\geq\beta, \end{aligned} \tag{BJCCP}$$

where, for any set $A\subseteq\Theta$, $\Pi(A|\tilde{X}_n)=\int_A\boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta$. The (BJCCP) formulation provides a principled way to combine data with parametric models of the uncertainty in (JCCP). To the best of our knowledge, this formulation has not been considered in the literature on data-driven chance constrained optimization before and, we believe, a useful addition to the growing toolbox of methodology for solving such problems; see Section 4.1.1.

The posterior can be computed in closed-form under conjugacy assumptions. However, these assumptions are restrictive and untenable for many application settings. The computation of the posterior under more general conditions is intractable, since the evidence cannot be easily calculated. Consequently, there is a substantial body of work on *approximate Bayesian computation* focused on the question of efficiently and accurately approximating the posterior distribution. Broadly speaking, there are two classes of methods in approximate Bayesian computation: sampling methods and optimization-based methods. Markov chain Monte Carlo (MCMC) is the canonical sampling method, where the objective is to design a stationary Markov chain whose invariant distribution is precisely the posterior distribution.

Initializing the Markov chain in an arbitrary initial state, after a 'burn-in' period the state of the designed Markov chain is (roughly speaking) a sample from a distribution that closely approximates the invariant/posterior distribution (where closeness is typically measured in terms of the total variation distance). MCMC, however, is known to suffer from high variance, complex diagnostics, and has poor scaling properties with the problem dimension [25]. Furthermore, as we will show below, sample-based methods in chance constrained settings can produce non-convex feasible sets, even when the 'true' problem is convex feasible. Coupled with the high variance of the methods, it may not be appropriate to use MCMC (or other sampling methods) to solve data-driven chance constrained problems like (BJCCP).

Variational Bayesian (VB) methods [25], in contrast, use optimization to compute an approximation to the posterior distribution from a class of 'simpler' distribution functions (that does not, necessarily, contain the posterior) called the *variational family*, by minimizing divergence from the posterior distribution. Importantly, the posterior distribution being intractable, VB methods optimize a surrogate objective that lower bounds the divergence measure, and the optimizer of the surrogate is precisely the posterior distribution when the class of distributions includes it. The Kullback-Leibler divergence is a standard choice in VB methods [25], though there is increasing interest in α-Rényi divergence as well [26] which yield approximations that have better support coverage. Broadly speaking, VB methods trade variance for bias; specifically, there is no sampling variance, but since the variational family does not contain the 'true' posterior, there is often an unavoidable bias that is introduced. From the perspective of solving data-driven chance constrained stochastic optimization problems, this trade-off may be appropriate, since the approximation (under very general conditions, as we show) is often necessarily convex feasible. Consequently, we focus on Kullback-Leibler divergence-based VB methods and consider the question of asymptotic consistency (in the large sample limit) of the variational approximation (VBJCCP) to (BJCCP).

Besides proposing (BJCCP) and (VBJCCP) (see Section 4.3 below), our primary contributions are to

1. Demonstrate the convex feasibility of the joint chance constraint (VBJCCP) when the posterior distribution belongs to a 'nice' class of distributions.

2. Establish the 'frequentist' statistical consistency of the value of both (BJCCP) and (VBJCCP) in the limit of a large data-set and a single chance constraint.

3. Quantify the consistency results for the value of both (BJCCP) and (VBJCCP), by establishing a probabilistic rate of convergence for a single chance constraint.

Frequentist consistency of Bayesian methods demonstrate that the Bayesian posterior concentrates on the 'true' parameter θ_0 of the data generating distribution in the large sample limit. Typically this is demonstrated by showing that the posterior converges weakly to a Dirac delta distribution concentrated at θ_0 in probability or almost surely under the data-generating distribution [111]. Here, we consider the frequentist consistency of the value of (VBJCCP), and establish convergence in probability results demonstrating the consistency of VB approximations in Theorem 4.4.3 and a probabilistic rate of convergence in Theorem 4.4.2. Furthermore, as direct corollaries, we can easily recover consistency and rates of convergence for (BJCCP).

4.1.1 Relevant Literature

To the best of our knowledge, Bayesian models of data-driven chance constrained optimization have not been considered before in the literature. At the outset, we note that there is precedence for Bayesian formulations of data-driven stochastic optimization problems – for instance, [15] develop the so-called Bayesian risk optimization (BRO) decision-making framework and establish frequentist consistency of the optimal value in the large sample limit; see recent follow-on work [37], [38] as well. In [39], an approximate Bayesian formulation of the risk-sensitive decision-making problem is considered and, again, frequentist consistency results are established. Neither of these papers consider the chance constrained setting of this chapter.

Nonetheless, there is an extensive literature on data-driven methods for solving chance constrained optimization problems, specifically scenario-based (SB) approaches [40]–[42],

distributionally robust optimization (DRO) [40], [43]–[45] and sample average approximation (SAA) [46], [47]. This is by no means a comprehensive literature review, but highlights the range of approaches that have been explored. We direct the reader to the excellent recent review paper [48] for a comprehensive overview of the literature on data-driven chance constrained optimization.

Both scenario and SAA approaches use samples from the uncertainty probability measure in (TP) to compute an estimate of the solution. However, this presumes that it is possible to access the uncertainty measure, which may not be possible in practice. In the DRO approach, the uncertainty measure is assumed to belong to a pre-defined class of probability measures and the chance constraints are required to be satisfied by every probability measure in this 'ambiguity set.' In the data-driven setting, the ambiguity set is constructed 'centered' (defined, for instance, through the Wasserstein metric) around the empirical measure computed using samples of the parameter, which in the large sample limit converges to the true uncertainty measure; see [49] which establishes the consistency of chance-constrained DRO with Wasserstein ambiguity sets. This highlights an important difference with our current setting, where the posterior distribution (or its approximation) is used as a quantification of the epistemic uncertainty about the 'true' parameter θ_0, and is shown to weakly converge to a Dirac delta distribution concentrated at θ_0, in the limit of a large covariate sample size $\tilde{X}_n$.

The rest of the chapter is laid out as follows. In the next section we introduce necessary notation and definitions that will be used throughout the chapter. In Section 4.3 we detail both (BJCCP) and (VBJCCP) providing a clean rationale for the modeling framework, and demonstrate the convex feasibility of (VBJCCP). Next, in Section 4.4 we first establish the asymptotic consistency of the optimal value and the optimizers of (VBJCCP) under general conditions on the objective and constraint functions and then establish convergence rates for values of (VBJCCP) and (BJCCP). We end in Section 4.5 with a simulation result demonstrating the efficacy of our approach in solving an optimal staffing problem.

4.2 Notations and Definitions

In this section, we introduce important notations and definitions used throughout the chapter. We define an indicator function for any arbitrary set A as $\mathbb{I}_A(t) := 1$ if $t \in A$ or 0 if $t \notin A$. Let $\|\cdot\|$ denote the Euclidean norm. Let δ_θ represent the Dirac delta distribution function, or singularity, concentrated at the parameter θ. Given an ensemble of random variables $\tilde{X}_n$ distributed as P_0^n for any $n \geq 1$, following [91] we define the convergence of a sequence of random mappings $\{f_n : \tilde{X}_n \to \mathbb{R}\}$ to f in P_0^n- probability as $\lim_{n\to\infty} P_0^n(|f_n - f| > \epsilon) = 0$ for any $\epsilon > 0$. We also use the notation $\lim_{n\to\infty} f_n \overset{P_0^n}{=} f$ or $f_n \overset{P_0^n}{\to} f$ as $n \to \infty$ to denote convergence in P_0^n- probability. Next, we define degenerate distributions as

Definition 4.2.1 (Degenerate distributions)**.** *A sequence of distributions $\{q_n(\theta)\}$ converges weakly to δ_θ that is, $q_n(\theta) \Rightarrow \delta_\theta$ for a $\theta \in \Theta$, if and only if $\forall \eta > 0$ $\lim_{n\to\infty} \int_{\{\|\theta-\theta\|>\eta\}} q_n(\theta)d\theta = 0$.*

Definition 4.2.2 (Rate of convergence)**.** *A sequence of distributions $\{q_n(\theta)\}$ converges weakly to δ_{θ_1}, $\forall \theta_1 \in \Theta$ at the rate of γ_n if*

(1) the sequence of means $\{\breve{\theta}_n := \int \theta q_n(\theta)d\theta\}$ converges to θ_1 as $n \to \infty$, and

(2) the variance of $\{q_n(\theta)\}$ satisfies $E_{q_n(\theta)}[\|\theta - \breve{\theta}_n\|^2] = O\left(\frac{1}{\gamma_n^2}\right)$.

We also define rescaled density functions as follows.

Definition 4.2.3 (Rescaled density)**.** *For a random variable ξ distributed as $d(\xi)$ with expectation $\tilde{\xi}$, for any sequence of matrices $\{t_n\}$, the density of the rescaled random variable $\mu := t_n(\xi - \tilde{\xi})$ is $\breve{d}_n(\mu) = |det(t_n^{-1})|d(t_n^{-1}\mu + \tilde{\xi})$, where $det(\cdot)$ represents the determinant of the matrix.*

Next, recall the definition of a test function [110].

Definition 4.2.4 (Test function)**.** *Let $\tilde{X}_n$ be a sequence of random variables on measurable space $(\mathbb{R}^{q\times n}, \mathcal{S}^n)$. Then any $\mathcal{S}^n$-measurable sequence of functions $\{\phi_n\}$, $\phi_n : \tilde{X}_n \mapsto [0, 1]$ $\forall n \in \mathbb{N}$, is a test of a hypothesis that a probability measure on $\mathcal{S}^n$ belongs to a given set against*

the hypothesis that it belongs to an alternative set. The test ϕ_n is consistent for hypothesis P_0^n against the alternative $P^n \in \{P_\theta^n : \theta \in \Theta \backslash \{\theta_0\}\}$ if $\mathbb{E}_{P^n}[\phi_n] \to \mathbb{I}_{\{\theta \in \Theta \backslash \{\theta_0\}\}}(\theta), \forall \theta \in \Theta$ as $n \to \infty$, where $\mathbb{I}_{\{.\}}$ is an indicator function.

A classic example of a test function is $\phi_n^{\text{KS}} = \mathbb{I}_{\{\text{KS}_n > K_\nu\}}(\theta)$ that is constructed using the Kolmogorov-Smirnov statistic $\text{KS}_n := \sup_t |\mathbb{F}_n(t) - \mathbb{F}_\theta(t)|$, where $\mathbb{F}_n(t)$ and $\mathbb{F}_\theta(t)$ are the empirical and true distribution respectively, and K_ν is the confidence level. If the null hypothesis is true, the Glivenko-Cantelli theorem [109, Theorem 19.1] shows that the KS statistic converges to zero as the number of samples increases to infinity.

4.3 Variational Bayesian Chance Constrained Optimization

Consider a parameterized joint probability distribution P_θ^n over $\mathbb{R}^{d \times n}$, where $\theta \in \mathbb{R}^q$ and let $p_\theta^n(\cdot)$ represent the corresponding density. We observe a random sample $\tilde{X}_n = \{X_1, X_2, \ldots, X_n\}$ drawn from $P_{\theta_0}^n \equiv P_0^n$. Note that $\tilde{X}_n$ need *not* be an independent and identically distributed (IID) sequence. Recall from (4.1) that the Bayesian approach computes a posterior over the unknown 'true' parameter θ_0, giving rise to the Bayesian joint chance-constrained optimization problem (BJCCP).

As noted in the introduction, there are the two significant challenges in solving (BJCCP):

(i) *Computing the posterior distribution.* While in some cases conjugate priors can be used, this is not appropriate in most problems. In general, posterior computation is intractable, and it is the common motivation for using approximate Bayesian inference methods [25] .

(ii) *Convexity of the feasible set.* Observe that, even if the posterior distribution is computable, to qualify (BJCCP) as a convex program, the feasible set,

$$\{a \in \mathcal{A} : \Pi\left(g_{\text{i}}(a, \theta) \leq 0,\ \text{i} \in \{1, 2, 3, \ldots, m\} | \tilde{X}_n\right) \geq \beta\} \tag{4.2}$$

must be convex. However, it is possible that this set is not convex, even when the underlying constraint functions $g_{\text{i}}(a, \theta), \text{i} \in \{1, 2, \ldots m\}$ are (in a) and, thus, finding a

global optimum becomes challenging [119]. This raises the canonical question of when (VBJCCP) and (BJCCP) are convex feasible.

Note that, if the constraint function has some structural regularity and the posterior distribution belongs to an appropriate class of distributions, then it can be shown that the feasible set in (4.2) is convex. For instance,

Proposition 4.3.1. *[2, Theorem 2.5] If the constraint functions $g_{\mathrm{i}}(a, \mathbf{y}), \mathrm{i} \in \{1, 2, \dots m\}$ for $a \in \mathcal{X}$ and $\mathbf{y} \in \mathbb{R}^q$ are quasi-convex in $(a, \mathbf{y})$ and θ is a random variable with log-concave probability distribution, then the feasible set in (BJCCP) is convex.*

Proof. The proof is a direct consequence of the result in Theorem 2.5 in [2]. □

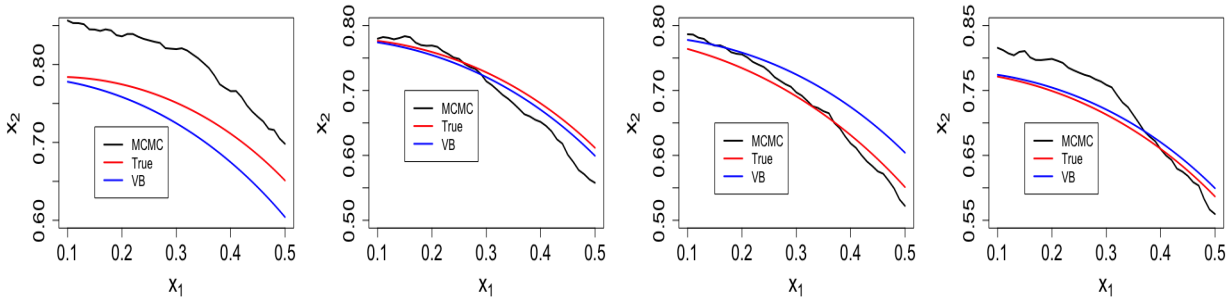
Furthermore, [120] showed that if the constraint function $g_{\mathrm{i}}(a, \theta)$ is of the form $\{\mathbf{a}^T a \leq \mathbf{b}\}$, where $\theta = (\mathbf{a}^T, \mathbf{b})^T$ and has a symmetric log-concave density then with $\beta > \frac{1}{2}$ the feasible set in (BJCCP) is convex.

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 [24], [121]. Apart from these common issues, there is another important reason due to which any sampling-based method cannot 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 feasible set in (BJCCP), results in a non-convex feasible set [51]. To illustrate this, consider the following simple example of a chance-constraint feasible set motivated by [51].

Example 4.3.1. *Figure 4.1(a) plots the chance-constraint feasible set*

$$\left\{a \in R^2 : \mathcal{N}\left(\theta^T a - 1 \leq 0 | \mu = [0, 0]^T, \mathbf{\Sigma}_A = [1, -0.1; -0.1, 1]\right) > \beta\right\}, \tag{4.3}$$

and its empirical approximator using 8000 MCMC samples (Metropolis-Hastings with a 'burn-in' of 3000 samples) generated from the underlying correlated multivariate Gaussian



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

Figure 4.1. Feasible Region : True Distribution vs Monte Carlo Approximation (5000 samples) vs. VB (mean field approximation).

distribution. We fix $\beta = 0.9$. *We observe that the resulting MC approximate feasible set is non-convex.*

Next, we show that using the popular 'mean-field variational family' [25] to approximate the correlated multivariate Gaussian distribution in the same example in (4.3)*, we obtain a smooth and convex approximation to the (BJCCP) feasible set. First, we compute mean-field approximation* $q_A(\theta)$ *and* $q_B(\theta)$ *of* $\mathcal{N}\left(\theta|\mu = [0,0]^T, \mathbf{\Sigma}\right)$ *for four different covariance matrices* $\mathbf{\Sigma}$*, 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 4.1. 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.*

4.3.1 Variational Bayes

Variational Bayes (VB) methods are an alternative method for computing an approximate posterior. Standard VB minimizes the Kullback-Leibler (KL) divergence measure to compute q^*, the element in a given class of distributions $\mathcal{Q}$ that is 'closest' to the posterior $\boldsymbol{\pi}(\theta|\tilde{X}_n)$:

$$q^*(\theta|\tilde{X}_n) \in \text{argmin}_{q\in\mathcal{Q}}\ \text{KL}(q(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)) := \int q(\theta)\log\frac{q(\theta)}{\boldsymbol{\pi}(\theta|\tilde{X}_n)}d\theta. \tag{4.4}$$

Using this, we approximate (BJCCP) with,

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & \mathbb{E}_{Q^*(\theta|\tilde{X}_n)}[R(a,\theta)] \\ \text{s.t.} \quad & Q^*\left(g_{\mathrm{i}}(a,\theta)\leq 0,\ \mathrm{i}\in\{1,2,3,\ldots,m\}|\tilde{X}_n\right)\geq\beta, \end{aligned} \tag{VBJCCP}$$

where β is the confidence level and for any set $A\subseteq\Theta$, $Q^*(A|\tilde{X}_n)=\int_A q^*(\theta|\tilde{X}_n)d\theta$. Observe that the optimization problem (4.4) is infeasible, since the posterior is unknown. However, unpacking the KL divergence, we see that

$$\mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n))=\int q(\theta)\log\frac{q(\theta)}{\boldsymbol{\pi}(\theta,\tilde{X}_n)}d\theta+\log\int p_\theta^n(\tilde{X}_n)\boldsymbol{\pi}(\theta)d\theta. \tag{4.5}$$

Since, $\log\int p_\theta^n(\tilde{X}_n)\boldsymbol{\pi}(\theta)d\theta$ is a constant (with respect to q), minimizing the KL divergence is equivalent to maximizing $\int q(\theta)\log\frac{\boldsymbol{\pi}(\theta,\tilde{X}_n)}{q(\theta)}d\theta$. Since, KL divergence is non-negative, it follows that the log-evidence satisfies

$$\begin{aligned} \log\int p_\theta^n(\tilde{X}_n)\boldsymbol{\pi}(\theta)d\theta &\geq \int q(\theta)\log\frac{\boldsymbol{\pi}(\theta,\tilde{X}_n)}{q(\theta)}d\theta \\ &= -\mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta))+\int\log p_\theta^n(\tilde{X}_n)\ q(\theta)d\theta, \end{aligned} \tag{ELBO}$$

and the bound is tight if and only if the optimizer $q^*(\cdot)$ is the 'true' posterior distribution. Thus, an approximate posterior can be computed by maximizing the so-called *evidence lower bound* (ELBO) in the final expression above:

$$q^*(\theta|\tilde{X}_n)\in\arg\max_{q\in\mathcal{Q}}\int\log p_\theta^n(\tilde{X}_n)\ q(\theta)d\theta-\mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta)). \tag{4.6}$$

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 [2] and [120] the choice of a log-concave family of distributions as the approximating family could retain the convexity of the feasible set, if the constraint functions have certain structural regularity.

As Example 4.3.1 shows, the VB approximation of the feasbility set could include infeasible points, in general. This raises the question of whether the VB approximation can be consistent (in some appropriate sense) when the sample size n is large. In other words, is there a notion of 'frequentist' consistency of the feasbility set, the optimal values and solutions? We address this question in the remainder of the chapter.

4.4 Asymptotic Analysis

In this section, we first identify regularity conditions on the prior distribution, the likelihood model, the variational family, and the risk and constraint functions to establish the rate at which the feasible region of (VBJCCP) coincides with the true feasible region. Then, under similar regularity conditions, we derive the convergence rate of the optimal value of (VBJCCP) to that of (TP), in the setting with a single constraint function (i.e., $m = 1$). We derive the convergence rate result under very mild conditions on the prior distribution and the likelihood models that are, nonetheless, hard to verify in practice for many problems of interest. Therefore, under more restrictive, but easily verifiable, regularity conditions we show that the the optimal values V^*_{VB} of (VBJCCP) converges to the optimal value V^* of (TP) at $\theta = \theta_0$ (respectively), in P_0^n-probability as the number of samples converges to infinity, again in the setting with a single constraint function. Note that it follows from the definition of the VB posterior $q^*(\theta|\tilde{X}_n)$ in (4.4) that when the variational family $\mathcal{Q}$ consists of all possible distributions then $q^*(\theta|\tilde{X}_n)$ coincides with the true posterior distribution. Consequently, all of our theoretical results for (VBJCCP) trivially extend to (BJCCP).

4.4.1 Convergence rate and feasibility guarantee

We state the assumptions under which we establish the rate of convergence and feasibility guarantee results. Let $L_n : \Theta \times \Theta \mapsto [0, \infty)$ be an arbitrary loss function that measures the distance between parameters and also depends on n.

Assumption 4.4.1. *Let* $\{\epsilon_n\} \subset (0,\infty)$ *be a sequence such that* $\epsilon_n \to \infty$ *and* $n\epsilon_n^2 \geq 1$ *as* $n \to \infty$. *Fix* $n \geq 1$. *Then, for* $L_n(\theta,\theta_0) \geq 0$ *and any* $\epsilon > \epsilon_n$, $\exists$ *a test function* $\phi_{n,\epsilon} : \tilde{X}_n \mapsto [0,1]$ *and sieve set* $\Theta_n(\epsilon) \subseteq \Theta$ *such that*

(i) $\mathbb{E}_{P_0^n}[\phi_{n,\epsilon}] \leq C_0 \exp(-Cn\epsilon^2)$, *and*

(ii) $\sup_{\{\theta \in \Theta_n(\epsilon): L_n(\theta,\theta_0) \geq C_1 n\epsilon^2\}} \mathbb{E}_{P_\theta^n}[1-\phi_{n,\epsilon}] \leq \exp(-Cn\epsilon^2)$.

Assumption 4.4.1(i) quantifies the rate at which a Type-1 error diminishes with the sample size, while the condition in Assumption 4.4.1(ii) quantifies that of a Type-2 error. Assumption 4.4.2 below ensures the prior distribution places 'sufficient' mass on the sieve set $\Theta_n(\epsilon)$ defined in Assumption 4.4.1.

Assumption 4.4.2. *Let* $\{\epsilon_n\} \subset (0,\infty)$ *be a sequence such that* $\epsilon_n \to \infty$ *and* $n\epsilon_n^2 \geq 1$ *as* $n \to \infty$. *Fix* $n \geq 1$. *Then, the prior distribution satisfies* $\mathbb{E}_\Pi[\mathbb{I}_{\{\Theta_n^c(\epsilon)\}}] \leq \exp(-Cn\epsilon^2)$.

Notice that Assumption 4.4.2 is trivially satisfied if $\Theta_n(\epsilon) = \Theta$. The next assumption ensures that the prior distribution places sufficient mass around a neighborhood A_n, defined using the Rényi divergence, of the true parameter θ_0.

Assumption 4.4.3. *Fix* $n \geq 1$ *and a constant* $\lambda > 0$. *Let* $A_n := \{\theta \in \Theta : D_{1+\lambda}(P_0^n \| P_\theta^n) \leq C_3 n\epsilon_n^2\}$, *where* $D_{1+\lambda}(P_0^n \| P_\theta^n) := \frac{1}{\lambda}\log \int \left(\frac{dP_0^n}{dP_\theta^n}\right)^\lambda dP_0^n$ *is the Rényi divergence between* P_0^n *and* P_θ^n, *assuming* P_0^n *is absolutely continuous with respect to* P_θ^n. *The prior distribution satisfies* $\mathbb{E}_\Pi[\mathbb{I}_{\{A_n\}}] \geq \exp(-nC_2\epsilon_n^2)$.

Observe that the set A_n defines a neighborhood of the distribution corresponding to θ_0. If Assumption 4.4.3 is violated then the posterior too will place no mass in this neighborhood of θ_0, implying asymptotic inconsistency. Assumptions 4.4.1, 4.4.2, and 4.4.3 are adopted from [91] and has also been used in [122] to prove convergence rates of variational posteriors.

Our main result demonstrating the rate of convergence follows a series of lemmas. All the proofs (except main results) can be found in Section 4.6. We first recall the following result from [122],

Lemma 4.4.1 (Theorem 2.1 [122]). *For any* $L_n(\theta,\theta_0) \geq 0$ *and* $\delta > 0$, *under Assumptions 4.4.1, 4.4.2, and, 4.4.3, and for* $C > C_2 + C_3 + 2$ *and* $\eta_n^2 :=$

$\frac{1}{n}\inf_{q\in\mathcal{Q}}\mathbb{E}_{P_0^n}\left[\int_\theta q(\theta)\log\frac{q(\theta)}{\pi(\theta|\tilde{X}_n)}d\theta\right]$, *the VB approximator of the true posterior,* $q^*(\theta|\tilde{X}_n)$, *satisfies,*

$$P_0^n\left[\int_\theta L_n(\theta,\theta_0)q^*(\theta|\tilde{X}_n)d\theta > n\delta\right] \leq \frac{M}{\delta}(\epsilon_n^2+\eta_n^2) \tag{4.7}$$

for some constant M *that depends on the* $C, C_1, C_2,$ *and* C_3.

As noted before in Assumption 4.4.1, the distance function $L_n(\theta,\theta_0)$ is arbitrary and it quantifies the distance between model P_θ^n and P_0^n. For instance, $L_n(\theta,\theta_0)$ could be chosen to be $n\|\theta-\theta_0\|$. Also, note that the rate comprises of two sequences ϵ_n^2 and η_n^2. The sequence ϵ_n is the rate of convergence of the true posterior. In particular, [91] established ϵ_n as the rate of convergence of the true posterior under Assumptions 4.4.1, 4.4.2, and 4.4.3. On the other hand, evident from its definition, the second sequence in the VB convergence rate is due to the variational approximation. Moreover, it is straightforward to observe that when $\mathcal{Q}$ is the family of all possible distributions, η_n^2 is 0. Furthermore, under certain conditions on the variational family $\mathcal{Q}$ (see Assumption 4.4.4), it can be shown that η_n^2 is bounded above by another convergent sequence ϵ_n^2. In fact, in Lemma 4.5.3 we show that $\epsilon_n = \epsilon_n$ for the prior, the likelihood and the variational family chosen for the optimal staffing problem discussed in Section 4.5.

We first use the result above to prove the finite sample feasibility guarantee of the (VB-JCCP) solution. Let us define the set where the true constraint $\mathrm{i}\in\{1,2,\dots m\}$ is satisfied as $F_0^{\mathrm{i}} := \{a\in\mathcal{A} : \{g_{\mathrm{i}}(a,\theta_0)\leq 0\},\}$, and VB-approximate feasible set is denoted as $\hat{F}_{VB}(\tilde{X}_n) := \{a\in\mathcal{A} : Q^*\left(g_{\mathrm{i}}(a,\theta)\leq 0,\ \mathrm{i}\in\{1,2,3,\dots,m\}|\tilde{X}_n\right)\geq\beta\}$. We show that the solutions obtained for (VBJCCP) are feasible for (TP) with high probability. In particular, we show that if a point does not satisfy any of the constraints, then the probability of that point being in the VB approximate feasible set decays at a certain rate. We quantify that rate in the following result.

Theorem 4.4.1. *For any* $\mathrm{i} \in \{1, 2, \ldots, m\}$ *let* $a \in \mathcal{A} \backslash F_0^{\mathrm{i}}$ *and* $L_n^{\mathrm{i}}(\theta, \theta_0) := n \sup_{a \in \mathcal{A}} \mathbb{I}_{(0,\infty)}(g_{\mathrm{i}}(a, \theta_0) - g_{\mathrm{i}}(a, \theta))$ *satisfies Assumption 4.4.1. Then under Assumptions 4.4.2 and 4.4.3, there exists a constant* $C_{\mathrm{i}} > 0$ *for each* $\mathrm{i} \in \{1, 2, \ldots m\}$*, such that*

$$P_0^n[a \in \hat{F}_{VB}(\tilde{X}_n)] \leq \frac{C_{\mathrm{i}}}{\beta}(\epsilon_n^2 + \eta_n^2),$$

where $\epsilon_n^2 \to 0$ *as* $n \to \infty$ *and* $\eta_n^2 = \frac{1}{n} \inf_{q \in \mathcal{Q}} \mathbb{E}_{P_0} \left[\mathrm{KL}(q(\theta) \| \boldsymbol{\pi}(\theta | \tilde{X}_n)) \right]$.

Proof. Using Markov's inequality observe that for any $a \in \mathcal{A}$,

$$\begin{aligned} P_0^n[Q^*\left(g_{\mathrm{i}}(a, \theta) \leq 0,\ \mathrm{i} \in \{1, 2, 3, \ldots, m\} | \tilde{X}_n\right) \geq \beta] &\leq \frac{1}{\beta} \mathbb{E}_0[Q^*\left(\cap_{\mathrm{i}=1}^m \{g_{\mathrm{i}}(a, \theta) \leq 0\} | \tilde{X}_n\right)] \\ &\leq \frac{1}{\beta} \mathbb{E}_0[Q^*\left(\{g_{\mathrm{i}}(a, \theta) \leq 0\} | \tilde{X}_n\right)] \end{aligned} \tag{4.8}$$

for any $\mathrm{i} \in \{1, \ldots, m\}$. Fixing $\mathrm{i} \in \{1, \ldots, m\}$, since $a \in \mathcal{A} \backslash F_0^{\mathrm{i}}$ implies that $a \in \{g_{\mathrm{i}}(a, \theta_0) > 0\}$, it follows that $\{g_{\mathrm{i}}(a, \theta) \leq 0\} \subseteq \{g_{\mathrm{i}}(a, \theta) < g_{\mathrm{i}}(a, \theta_0)\}$. Therefore, for all $a \in \mathcal{A} \backslash F_0^{\mathrm{i}}$, it follows from (4.8) that

$$P_0^n[Q^*\left(g_{\mathrm{i}}(a, \theta) \leq 0,\ \mathrm{i} \in \{1, 2, 3, \ldots, m\} | \tilde{X}_n\right) \geq \beta] \leq \frac{1}{\beta} \mathbb{E}_0[Q^*\left(\{g_{\mathrm{i}}(a, \theta) < g_{\mathrm{i}}(a, \theta_0)\} | \tilde{X}_n\right)]. \tag{4.9}$$

Now using [122, Theorem 2.1], it follows that if $L_n^{\mathrm{i}}(\theta, \theta_0) := n \sup_{a \in \mathcal{A}} \mathbb{I}_{(0,\infty)}(g_{\mathrm{i}}(a, \theta_0) - g_{\mathrm{i}}(a, \theta))$ satisfies Assumption 4.4.1, then there exists a constant C_{i} such that $\mathbb{E}_0[Q^*\left(\{g_{\mathrm{i}}(a, \theta) < g_{\mathrm{i}}(a, \theta_0)\} | \tilde{X}_n\right)] \leq C_{\mathrm{i}}(\epsilon_n^2 + \eta_n^2)$, where $\eta_n^2 := \frac{1}{n} \inf_{q \in \mathcal{Q}} \mathbb{E}_{P_0} \left[\int_\theta q(\theta) \log \frac{q(\theta)}{\boldsymbol{\pi}(\theta | \tilde{X}_n)} d\theta \right]$. Finally, using Theorem 4.4.1 in (4.9), the assertion follows immediately. □

Now, we state a straightforward corollary of the result above establishing feasibility guarantee of the (BJCCP) solution.

Corollary 4.4.1. *For any* $\mathrm{i} \in \{1, 2, \ldots, m\}$ *let* $a \in \mathcal{A} \backslash F_0^{\mathrm{i}}$ *and* $L_n^{\mathrm{i}}(\theta, \theta_0) := n \sup_{a \in \mathcal{A}} \mathbb{I}_{(0,\infty)}(g_{\mathrm{i}}(a, \theta_0) - g_{\mathrm{i}}(a, \theta))$ *satisfies Assumption 4.4.1. Then under Assumptions 4.4.2 and 4.4.3, there exists a constant* $C_{\mathrm{i}} > 0$ *for each* $\mathrm{i} \in \{1, 2, \ldots m\}$*, such that*

$$P_0^n[a \in \hat{F}_B(\tilde{X}_n)] \leq \frac{C_{\mathrm{i}}}{\beta} \epsilon_n^2,$$

where $\hat{F}_B(\tilde{X}_n) := \{a \in \mathcal{A} : \Pi\left(g_{\mathrm{i}}(a, \theta) \leq 0,\ \mathrm{i} \in \{1, 2, 3, \ldots, m\} | \tilde{X}_n\right) \geq \beta\}$*,* $\epsilon_n^2 \to 0$ *as* $n \to \infty$*.*

Proof. The proof follows straightforwardly from Theorem 4.4.1 and the fact that $q^*(\theta|\tilde{X}_n)$ is the same as the true posterior distribution and $\eta_n^2 = 0$, when the variational family $\mathcal{Q}$ is fixed to the set of all possible distributions on Θ. □

To leverage the result in Lemma 4.4.1 in establishing the rate of convergence of the optimal value of (VBJCCP), we now fix $L_n(\theta, \theta_0)$ to specific positive distance functions in the following two lemmas. Lemma 4.4.2 establishes a rate of convergence of the VB posterior constraint set to the true constraint set.

Lemma 4.4.2. *If* $L_n^1(\theta, \theta_0) = n \sup_{a \in \mathcal{A}} |\mathbb{I}_{(-\infty,0]}(g(a, \theta)) - \mathbb{I}_{(-\infty,0]}(g(a, \theta_0))|$ *satisfies Assumption 4.4.1, then under the conditions of Lemma 4.4.1, for any* $\delta > 0$*, we have*

$$P_0^n \left[\sup_{a \in \mathcal{A}} |Q^*\left(g(a, \theta) \leq 0 | \tilde{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(a, \theta_0))| > \delta\right] \leq \frac{M_1}{\delta}(\epsilon_n^2 + \eta_n^2), \tag{4.10}$$

for a positive constant M_1*.*

In the following lemma, we establish the rate of convergence of the expected cost function under VB posterior to the true cost function.

Lemma 4.4.3. *If* $L_n^2(\theta, \theta_0) = n \sup_{a \in \mathcal{A}} |R(a, \theta) - R(a, \theta_0)|$ *satisfies Assumption 4.4.1, then under conditions of Lemma 4.4.1 for any* $\delta > 0$*,*

$$P_0^n[\sup_{a \in \mathcal{A}} |\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a, \theta)] - R(a, \theta_0)| > \delta] \leq \frac{M_2}{\delta}(\epsilon_n^2 + \eta_n^2). \tag{4.11}$$

The next theorem proves a rate of convergence on the optimal value of (VBJCCP) as a consequence of the lemmas above.

Theorem 4.4.2. *If* $L_n^1(\theta,\theta_0) = n\sup_{a\in\mathcal{A}}|\mathbb{I}_{(-\infty,0]}(g(a,\theta)) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))|$ *and* $L_n^2(\theta,\theta_0) = n\sup_{a\in\mathcal{A}}|R(a,\theta) - R(a,\theta_0)|$ *satisfy Assumption 4.4.1, then under Assumption 4.4.2 and 4.4.3, and when* $\mathcal{X}$ *is compact, for (fixed) constants* $M_1 > 0$ *and* $M_2 > 0$ *, we have for any* $\eta > 0$ *and* $\delta \in (0,\beta)$

$$P_0^n[|V_{VB}^*(\tilde{X}_n) - V^*| > 2\eta] \leq \left[\frac{M_1}{\min(\delta, 1-\beta)} + \frac{M_2}{\eta}\right](\epsilon_n^2 + \eta_n^2),$$

where $\epsilon_n^2 \to 0$ *as* $n \to \infty$ *and* $\eta_n^2 := \frac{1}{n}\inf_{q\in\mathcal{Q}} \mathbb{E}_{P_0}\left[\int_\theta q(\xi)\log\frac{q(\xi)}{\pi(\xi|\tilde{X}_n)}d\xi\right]$.

Proof. Recall $\mathcal{S}_{VB}^*(\tilde{X}_n)$ is the solution of (VBJCCP) and $\mathcal{S}^*$ is the solution of (TP) with $\theta = \theta_0$. Observe that, since both $Q^*\left(g(a,\theta) \leq 0|\tilde{X}_n\right)$ and $\mathbb{I}_{(-\infty,0]}(g(a,\theta_0))$ are upper- semi-continuous their corresponding super-level sets are closed, and since $\mathcal{A}$ is compact the corresponding feasible sets are also compact. Also, if the corresponding feasible sets are non-empty then the corresponding optimal sets $\mathcal{S}_{VB}^*(\tilde{X}_n)$ and $\mathcal{S}^*$ are too.

Next fix a point a^* in the true solution set of (TP). Since $\mathcal{A}$ is compact, for any $\epsilon > 0$, there is $a \in \mathcal{A}$ such that for any $\epsilon > 0$, there exists $a \in \mathcal{A}$ such that $\|a - a^*\| < \epsilon$ and $g(a,\theta_0) \leq 0$. This implies that there exists a sequence $\{a_k\} \subset \mathcal{A}$ such that $a_k \to a^*$ as $k \to \infty$ and $g(a_k,\theta_0) \leq 0$ for all $k \geq 1$ Now fix $a \in \mathcal{A}$ such that $g(a,\theta_0) \leq 0$ and, using Lemma 4.4.2, observe that for all $n \geq n_0$

$$\begin{aligned} &P_0^n[|Q^*\left(g(a,\theta) \leq 0|\tilde{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))| > \delta] \\ &\qquad = P_0^n[|Q^*\left(g(a,\theta) \leq 0|\tilde{X}_n\right) - 1| > \delta] \leq \frac{M_1}{\delta}(\epsilon_n^2 + \eta_n^2). \end{aligned}$$

Now, fix $\beta \in (0,1)$ and let $\delta = 1-\beta$. It follows from the above inequality that, for all $n > n_0$

$$P_0^n[Q^*\left(g(a,\theta) \leq 0|\tilde{X}_n\right) < 1-\delta] = P_0^n[Q^*\left(g(a,\theta) \leq 0|\tilde{X}_n\right) \leq \beta] \leq \frac{M_1}{1-\beta}(\epsilon_n^2 + \eta_n^2).$$

Notice that for $a \in \mathcal{A}$ such that $g(a,\theta_0) \leq 0$, $\{a \in \mathcal{X} : Q^*\left(g(a,\theta) \leq 0|\tilde{X}_n\right) > \beta\} \subseteq \{a \in \mathcal{X} : \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] \geq V_{VB}^*(\tilde{X}_n)\}$. Hence, for all $n \geq n_0$,

$$P_0^n[\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] < V_{VB}^*(\tilde{X}_n)] \leq \frac{M_1}{1-\beta}(\epsilon_n^2 + \eta_n^2). \tag{4.12}$$

Next, using the result in part(1) of Lemma 4.4.3, for all $n \geq n_0$, any $x \in \mathcal{A}$, and $\delta > 0$

$$P_0^n[|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)| > \delta] \leq \frac{M_2}{\delta}(\epsilon_n^2 + \eta_n^2). \tag{4.13}$$

Observe that, for any $\eta > 0$

$$\begin{aligned}
&P_0^n[R(a,\theta_0) - V_{VB}^*(\tilde{X}_n) < -2\eta] \\
&\leq P_0^n[\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - V_{VB}^*(\tilde{X}_n) < -\eta] + P_0^n[R(a,\theta_0) - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] < -\eta] \\
&\leq P_0^n[\{\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - V_{VB}^*(\tilde{X}_n) < -\eta\}] + \frac{M_2}{\eta}(\epsilon_n^2 + \eta_n^2) \\
&\leq P_0^n[\{\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - V_{VB}^*(\tilde{X}_n) < 0\}] + \frac{M_2}{\eta}(\epsilon_n^2 + \eta_n^2) \\
&\leq \frac{M_1}{1-\beta}(\epsilon_n^2 + \eta_n^2) + \frac{M_2}{\eta}(\epsilon_n^2 + \eta_n^2) = \left[\frac{M_1}{1-\beta} + \frac{M_2}{\eta}\right](\epsilon_n^2 + \eta_n^2),
\end{aligned}$$

where the second inequality follows from (4.13) and the last inequality uses (4.12). Now, since a can be chosen arbitrarily close to a^*, it follows that

$$P_0^n[V^* - V_{VB}^*(\tilde{X}_n) < -2\eta] = P_0^n[V^* - V_{VB}^*(\tilde{X}_n) < -2\eta] \leq \left[\frac{M_1}{1-\beta} + \frac{M_2}{\eta}\right](\epsilon_n^2 + \eta_n^2). \tag{4.14}$$

Next, let $\hat{a}_n \in \mathcal{S}_{VB}^*$; that is $\hat{a}_n \in \mathcal{A}$, $Q^*\left(g(\hat{a}_n,\theta) \leq 0|\tilde{X}_n\right) \geq \beta$ and $V_{VB}^*(\tilde{X}_n) = \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(\hat{a}_n,\theta)]$. Since $\mathcal{A}$ is compact, we assume that $\hat{a}_n \to a_0$ (the limit point of the sequence $\{\hat{a}_n\} \subseteq \mathcal{A}$).

Recall that Lemma 4.4.2 holds uniformly over any $a \in \mathcal{A}$, therefore using the fact that $Q^*\left(g(\hat{a}_n,\theta) \leq 0|\mathbf{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0)) \leq |Q^*\left(g(\hat{a}_n,\theta) \leq 0|\mathbf{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0))| \leq \sup_{a\in\mathcal{A}}|Q^*\left(g(a,\theta) \leq 0|\mathbf{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))|$, we have for all $n \geq n_0$ and $\delta > 0$,

$$P_0^n\left[Q^*\left(g(\hat{a}_n,\theta) \leq 0|\mathbf{X}_n\right) \leq \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0)) + \delta\right] \geq 1 - \frac{M_1}{\delta}(\epsilon_n^2 + \eta_n^2). \tag{4.15}$$

Next using the fact that $Q^*\left(g(\hat{a}_n,\theta)\leq 0|\tilde{X}_n\right)\geq\beta$ for every $n\geq 1$, it follows that $\hat{a}_n$ is a feasible point of (TP) for $\delta\leq\beta$, that is $\left\{a\in\mathcal{X}:Q^*\left(g(\hat{a}_n,\theta)\leq 0|\mathbf{X}_n\right)\leq\mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0))+\delta\right\}\subset\{a\in\mathcal{X}:\mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0))+\delta\geq\beta\}$. Therefore, it follows that

$$\begin{aligned}\left\{Q^*\left(g(\hat{a}_n,\theta)\leq 0|\mathbf{X}_n\right)\leq\mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0))+\delta\right\}&\subseteq\{\mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0))+\delta\geq\beta\}\\&\subseteq\{R(\hat{a}_n,\theta_0)\geq V^*\},\end{aligned}\tag{4.16}$$

since the penultimate condition implies that the $\hat{a}_n$ is a feasible point of (TP). Therefore, for any $\delta\leq\beta$, $P_0^n\left[R(\hat{a}_n,\theta_0)\leq V^*\right]\leq\frac{M_1}{\delta}(\epsilon_n^2+\eta_n^2)$. Since Lemma 4.4.3 holds uniformly over all a and therefore using the fact that $R(\hat{a}_n,\theta_0)-\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(\hat{a}_n,\theta)]\leq|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(\hat{a}_n,\theta)]-R(\hat{a}_n,\theta_0)|\leq\sup_{a\in\mathcal{A}}|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)]-R(a,\theta_0)|$, for any $\delta>0$, we have $P_0^n\left[\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(\hat{a}_n,\theta)]+\delta\geq R(\hat{a}_n,\theta_0)\right]=P_0^n\left[V_{VB}^*(\mathbf{X}_n)+\delta\geq R(\hat{a}_n,\theta_0)\right]\geq 1-\frac{M_2}{\delta}(\epsilon_n^2+\eta_n^2)$, and therefore $P_0^n\left[V_{VB}^*(\tilde{X}_n)+\delta\leq R(\hat{a}_n,\theta_0)\right]\leq\frac{M_2}{\delta}(\epsilon_n^2+\eta_n^2)$. Observe that for any $\eta>0$

$$\begin{aligned}&P_0^n\left[V^*-V_{VB}^*(\mathbf{X}_n)\geq 2\eta\right]\\&\quad\leq P_0^n\left[V^*-R(\hat{a}_n,\theta_0)\geq\eta\right]+P_0^n\left[R(\hat{a}_n,\theta_0)-V_{VB}^*(\mathbf{X}_n)\geq\eta\right]\\&\quad\leq P_0^n\left[V^*-R(\hat{a}_n,\theta_0)\geq 0\right]+P_0^n\left[R(\hat{a}_n,\theta_0)-V_{VB}^*(\mathbf{X}_n)\geq\eta\right]\\&\quad\leq\frac{M_1}{\delta}(\epsilon_n^2+\eta_n^2)+\frac{M_2}{\eta}(\epsilon_n^2+\eta_n^2)=\left[\frac{M_1}{\delta}+\frac{M_2}{\eta}\right](\epsilon_n^2+\eta_n^2),\end{aligned}\tag{4.17}$$

where $\delta<\beta$.

Combining equation (4.14) and (4.17), we obtain

$$\begin{aligned}P_0^n\left[|V^*-V_{VB}^*(\mathbf{X}_n)|\geq 2\eta\right]&\leq\max\left(\left[\frac{M_1}{\delta}+\frac{M_2}{\eta}\right],\left[\frac{M_1}{1-\beta}+\frac{M_2}{\eta}\right]\right)(\epsilon_n^2+\eta_n^2)\\&=\left[\frac{M_1}{\min(\delta,1-\beta)}+\frac{M_2}{\eta}\right](\epsilon_n^2+\eta_n^2).\end{aligned}\tag{4.18}$$

□

The next result establishes the convergence rate of the optimal value of (BJCCP) with single constraint.

Corollary 4.4.2. *If* $L_n^1(\theta,\theta_0) = n\sup_{a\in\mathcal{A}} |\mathbb{I}_{(-\infty,0]}(g(a,\theta)) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))|$ *and* $L_n^2(\theta,\theta_0) = n\sup_{a\in\mathcal{A}} |R(a,\theta) - R(a,\theta_0)|$ *satisfy Assumption 4.4.1, then under Assumption 4.4.2 and 4.4.3, and when* $\mathcal{X}$ *is compact, for (fixed) constants* $M_1 > 0$ *and* $M_2 > 0$ *, we have for any* $\eta > 0$ *and* $\delta \in (0,\beta)$

$$P_0^n[|V_B^*(\tilde{X}_n) - V^*| > 2\eta] \leq \left[\frac{M_1}{\min(\delta, 1-\beta)} + \frac{M_2}{\eta}\right]\epsilon_n^2,$$

where $V_B^*(\tilde{X}_n)$ *is the optimal value of (BJCCP) with single constraint and* $\epsilon_n^2 \to 0$ *as* $n \to \infty$.

Proof. The proof is a direct consequence of Theorem 4.4.2 and the fact that V_{VB}^* is the same as V_B^* and $\eta_n^2 = 0$, when the variational family $\mathcal{Q}$ is fixed to the set of all possible distributions on Θ. □

Characterizing η_n^2

In order to characterize η_n^2, we specify conditions on variational family $\mathcal{Q}$ such that $\eta_n^2 = O(\epsilon_n^2)$, for some $\epsilon_n \geq \frac{1}{\sqrt{n}}$ and $\epsilon_n \to 0$ as $n \to \infty$. We impose following condition on the variational family $\mathcal{Q}$ that lets us obtain a bound on η_n^2.

Assumption 4.4.4. *There exists a sequence of distributions* $\{q_n(\cdot)\} \subset \mathcal{Q}$ *such that for a positive constant* C_1, $\frac{1}{n}\left[\mathrm{KL}\left(q_n(\theta)\|\boldsymbol{\pi}(\theta)\right) + \mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n(\tilde{X}_n)\|dP_\theta^n(\tilde{X}_n)\right)\right]\right] \leq C_1\epsilon_n^2$.

If the observations in $\tilde{X}_n$ are i.i.d, then observe that $\frac{1}{n}\mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n(\tilde{X}_n))\|dP_\theta^n(\tilde{X}_n)\right)\right] = \mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_{\lambda_0})\|dP_\theta(\xi)\right)\right]$. Intuitively, this assumption implies that the variational family must contain a sequence of distributions that converges weakly to a Dirac delta distribution concentrated at the true parameter θ_0 otherwise the second term in the LHS of Assumption 4.4.4 will be non-zero. We demonstrate the satisfaction of Assumption 4.4.4 for a specific variational family in Lemma 4.5.3.

Proposition 4.4.1. *Under Assumption 4.4.4 and* $C_9 > 0$, $\eta_n^2 \leq C_9\epsilon_n^2$.

Existence of Tests

Recall that our convergence rates and finite sample feasibility guarantee depend on existence of certain tests for the specified distance functions. We prove a general result which is applicable to distance functions for which the set $\{\theta \in \Theta : L_n(\theta, \theta_0) > n\epsilon^2\}$ is fixed for any $\epsilon \in (0, 1]$ and is a null set for any $\epsilon > 1$ (for example such distance functions should satisfy $n^{-1}L_n(\theta, \theta_0) \in \{0, 1\}$). Notice that the distance functions $L_n^1(\theta, \theta_0)$ in Theorem 4.4.2 and $L_n^{\mathrm{i}}(\theta, \theta_0)$ in Theorem 4.4.1 satisfy these conditions.

We recall the following result from [91, Lemma 7.2] which is due to Le Cam.

Lemma 4.4.4. *Suppose that there exist tests ω_n such that for fixed sets $\mathcal{P}_0$ and $\mathcal{P}_1$, of probability measures*

$$\sup_{P_0^n \in \mathcal{P}_0} \mathbb{E}_{P_0^n}[\omega_n] \to 0 \text{ and } \sup_{P^n \in \mathcal{P}_1} \mathbb{E}_{P^n}[1 - \omega_n] \to 0 \text{ as } n \to \infty,$$

then there exist tests ϕ_n and constants $K > 0$ such that

$$\sup_{P_0^n \in \mathcal{P}_0} \mathbb{E}_{P_0^n}[\phi_n] \leq \mathrm{e}^{-Kn} \text{ and } \sup_{P^n \in \mathcal{P}_1} \mathbb{E}_{P^n}[1 - \phi_n] \leq \mathrm{e}^{-Kn}.$$

Proposition 4.4.2. *Given $\Theta \subseteq \mathbb{R}^d$, if there exists a sequence of test function $\phi_{n,\epsilon}$ for any $\epsilon > 1$, such that $\mathbb{E}_{P_0^n}[\phi_{n,\epsilon}] \leq \mathrm{e}^{-Kn\epsilon^2}$, then the distance functions $L_n^{\mathrm{i}}(\theta, \theta_0)$ in Theorem 4.4.1 for any $\mathrm{i} \in \{1, \ldots, m\}$ and $L_n^1(\theta, \theta_0)$ in Theorem 4.4.2 satisfy Assumption 4.4.1.*

For the distance function $L_n^2(\theta, \theta_0)$ in Theorem 4.4.2, we have to use [91, Lemma 7.1] or construct an explicit test function to satisfy Assumption 4.4.1. Interested readers may refer to [39], [91], [122] for further discussions on existence of tests and/or constructing bespoke test functions.

4.4.2 Asymptotic consistency

Although, the rate of convergence result implies asymptotic consistency, it will be evident from the application presented in Section 4.5 that the regularity conditions required to compute the rate are difficult to verify in practice. Consequently, in this section, we identify

slightly more restrictive, but more easily verifiable, conditions on the prior, likelihood, and the variational family to guarantee asymptotic consistency of the optimal value and solution of (VBJCCP). We assume that $m = 1$ in the remainder of this section.

First, we impose the following conditions on the prior distribution.

Assumption 4.4.5 (Prior Density)**.**

(1) The prior density function $\boldsymbol{\pi}(\theta)$ is continuous with non-zero measure in the neighborhood of the true parameter θ_0, and

(2) there exists a constant $M_p > 0$ such that $\boldsymbol{\pi}(\theta) \leq M_p \ \forall \theta \in \Theta$ and $\mathbb{E}_{\boldsymbol{\pi}(\theta)}[|\theta|] < \infty$.

Assumption 4.4.5 is satisfied by a large class of prior distributions. Next, we assume that the likelihood function satisfies the following asymptotic normality property. Recall that $P_0^n \equiv P_{\theta_0}^n$.

Assumption 4.4.6 (Local Asymptotic Normality)**.** *Fix $\theta_0 \in \Theta$. The sequence of log-likelihood functions $\{\log P_\theta^n(\tilde{X}_n)\}$ satisfies a* local asymptotic normality (LAN) *condition, if there exists a sequence of matrices $\{r_n\}$, a matrix $I(\theta_0)$ and a sequence of random vectors $\{\Delta_{n,\theta_0}\}$ weakly converging to $\mathcal{N}(0, I(\theta_0)^{-1})$ as $n \to \infty$, such that for every compact set $K \subset \mathbb{R}^d$*

$$\sup_{h \in K} \left| \log P^n_{\theta_0 + r_n^{-1}h}(\tilde{X}_n) - \log P^n_{\theta_0}(\tilde{X}_n) - h^T I(\theta_0)\Delta_{n,\theta_0} + \frac{1}{2} h^T I(\theta_0) h \right| \xrightarrow{P_0^n} 0 \text{ as } n \to \infty \ .$$

The LAN condition is standard, and holds for a wide variety of models. The assumption affords significant flexibility in the analysis by allowing the likelihood to be asymptotically approximated by a scaled Gaussian centered around θ_0 [109]. Any likelihood model that is twice-continuously differentiable satisfies the LAN condition [109, Eq. 7.15].

Next, we place a restriction on the variational family $\mathcal{Q}$:

Assumption 4.4.7.

1. *The variational family $\mathcal{Q}$ must contain distributions that are absolutely continuous with respect to the prior distribution.*

2. *There exists a sequence of distributions* $\{q_n(\theta)\}$ *in the variational family* $\mathcal{Q}$ *that converges to a Dirac delta distribution* δ_{θ_0} *at the rate of* $\sqrt{n}$ *and with mean* $\int \theta q_n(\theta) d\theta = \hat{\theta}_n$*, the maximum likelihood estimate.*

3. *The differential entropy of the rescaled density (Definition 4.2) of such sequence of distributions is positive and finite.*

The first condition ensures that the KL divergence in (4.4) is not undefined for all distributions in $\mathcal{Q}$, that is not absolutely continuous with respect to the posterior distribution. The Bernstein von-Mises theorem [109] shows that under mild regularity conditions, the posterior converges to a Dirac delta distribution at the true parameter θ_0 at the rate of $\sqrt{n}$, and the second condition ensures that the KL divergence is well defined for all large enough n. These three assumptions together imply that the VB approximate posterior weakly converges to δ_{θ_0} as number of samples increases.

Lemma 4.4.5 ([27], [123])**.** *Under Assumptions 4.4.5, 4.4.6, and 4.4.7*

$$q^*(\theta|\tilde{X}_n) \in \arg\min_{q\in\mathcal{Q}} \mathrm{KL}\left(q(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)\right) \Rightarrow \delta_{\theta_0} \text{ in } P_0^n - \text{probability as } n \to \infty. \tag{4.19}$$

Proof. See [27, Theorem 5(1)] or [123, Corollary 1] for a proof. □

Now to establish asymptotic properties of the optimal value and optimal solution to (VBJCCP), we assume that the following regularity conditions are satisfied by the cost and the constraint functions.

Assumption 4.4.8. *We assume that*

1. $R(a,\cdot)$ *and* $g_i(a,\cdot)$ *are measurable and continuous for every* $a \in \mathcal{A}$*, and* $R(\cdot,\theta)$ *and* $g(\cdot,\theta)$ *are continuous for almost every* $\theta \in \Theta$*.*

2. $R(\cdot,\theta)$ *is locally Lipschitz continuous in* a *with for almost every* $\theta \in \Theta$*, such that for* a_1, a_2 *in compact set* $\mathcal{A}$*,* $|R(a_1,\theta) - R(a_2,\theta)| \le K_{\mathcal{A}}(\theta)\|x_1 - x_2\|$ *for some* $K_{\mathcal{A}}(\theta) \le \bar{K}_{\mathcal{A}}$ *for almost every* $\theta \in \Theta$*.*

3. *$R(a,\cdot)$ is uniformly integrable with respect to any q in the variational family $\mathcal{Q}$, that is for any $\epsilon > 0$ and $a \in \mathcal{A}$, there exist a compact set $K_\epsilon \subset \Theta$, such that $\int_{\Theta \setminus K_\epsilon} R(a,\theta) q(\theta) d\theta < \epsilon$.*

We first establish consistency of the constraint function, under the 'true' data generating distribution.

Lemma 4.4.6. *Under Assumptions 4.4.5, 4.4.6,and 4.4.7, we show that for any $\delta > 0$*

$$\lim_{n\to\infty} P_0^n \left(\sup_{x\in\mathcal{A}} \left| \mathbb{E}_{q^*(\theta|\tilde{X}_n)} \left[\prod_{\mathrm{i}=1}^{m} \mathbb{I}_{(-\infty,0]}(g_{\mathrm{i}}(a,\theta) \right] - \prod_{\mathrm{i}=1}^{m} \mathbb{I}_{(-\infty,0]}(g_{\mathrm{i}}(a,\theta_0)) \right| > \delta \right) = 0$$

The next lemma establishes the point-wise and uniform convergence of the expected cost.

Lemma 4.4.7. *Under Assumptions 4.4.5, 4.4.6, 4.4.7, and 4.4.8, we show that,*

1. *For each $a \in \mathcal{A}$, $\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] \to R(a,\theta_0)$ in $P_0^n -$ probability as $n \to \infty$.*
2. *Suppose $\mathcal{A}$ is compact, then $\sup_{x\in\mathcal{A}} |\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)|$ converges to 0 in $P_0^n -$ probability as $n \to \infty$; that is for any $\delta > 0$*

$$\lim_{n\to\infty} P_0^n \left(\sup_{x\in\mathcal{A}} \left| \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0) \right| > \delta \right) = 0.$$

Using the results in Lemma 4.4.6 and 4.4.7, Theorem 4.4.3 establishes the asymptotic consistency of the optimal values of (VBJCCP) and, as a consequence, (BJCCP) with single constraint.

Theorem 4.4.3. *Under Assumptions 4.4.5, 4.4.6, 4.4.7, and 4.4.8 and when $\mathcal{A}$ is a compact set, we have $V_{VB}^*(\tilde{X}_n) \xrightarrow{P_0^n} V^*$ as $n \to \infty$.*

Proof. Recall $\mathcal{S}_{VB}^*(\tilde{X}_n)$ is the solution set of (VBJCCP) and $\mathcal{S}^*$ is the solution set of (TP). Observe that since both $Q^*\left(g(a,\theta) \leq 0 | \tilde{X}_n\right)$ and $\mathbb{I}_{(-\infty,0]}(g(a,\theta_0))$ are upper-semicontinuous, their corresponding super-level sets are closed and, since $\mathcal{A}$ is compact, the corresponding feasible sets are compact. Furthermore, if the corresponding feasible sets are non-empty then the corresponding optimal sets $\mathcal{S}_{VB}^*(\tilde{X}_n)$ and $\mathcal{S}^*$ are also non-empty.

Next fix a point a^* in the true solution set $\mathcal{S}^*$ of (TP). Since $\mathcal{A}$ is compact, for any $\epsilon > 0$, there is $a \in \mathcal{A}$ such that $\|a - a^*\| < \epsilon$ and $g(a, \theta_0) \leq 0$. It follows that there exists a sequence $\{a_k\} \subset \mathcal{A}$ such that $a_k \to a^*$ as $k \to \infty$ and $g(a_k, \theta_0) \leq 0$ for all $k \geq 1$. Now fix $a \in \mathcal{A}$ such that $g(a, \theta_0) \leq 0$. By Lemma 4.4.6, $Q^*\left(g(a,\theta) \leq 0 | \tilde{X}_n\right) \overset{P_0^n}{\to} \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))$ as $n \to \infty$, and therefore there exists an n_0 depending on $\epsilon > 0$ such that for all $n \geq n_0$ and any $\eta > 0$, we have for a given confidence level $\beta \in (0,1)$,

$$\begin{aligned} P_0^n\left(Q^*\left(g(a,\theta) \leq 0 | \tilde{X}_n\right) \geq \beta\right) &\geq P_0^n\left(Q^*\left(g(a,\theta) \leq 0 | \tilde{X}_n\right) \geq 1\right) \\ &\geq P_0^n\left(\mathbb{I}_{(-\infty,0]}(g(a,\theta_0)) - Q^*\left(g(a,\theta) \leq 0 | \tilde{X}_n\right) \leq 0\right) \\ &\geq P_0^n\left(\mathbb{I}_{(-\infty,0]}(g(a,\theta_0)) - Q^*\left(g(a,\theta) \leq 0 | \tilde{X}_n\right) \leq -\eta\right) \geq 1 - \epsilon. \end{aligned}$$

Hence for all $n \geq n_0$, a is a feasible solution of (VBJCCP) with P_0^n-probability of at least $1 - \epsilon$, and therefore

$$P_0^n\left(\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] \geq V_{VB}^*(\tilde{X}_n)\right) \geq P_0^n\left(Q^*\left(g(a,\theta) \leq 0 | \tilde{X}_n\right) \geq \beta\right) \geq 1 - \epsilon.$$

Now, since a can be chosen arbitrarily close to a^*, it follows from the equation above and the bounded convergence theorem that

$$P_0^n\left(\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a^*,\theta)] \geq V_{VB}^*(\tilde{X}_n)\right) \geq 1 - \epsilon. \tag{4.20}$$

for all $n \geq n_0$. For any $\delta > 0$ observe that

$$\begin{aligned} &P_0^n\left(V_{VB}^*(\tilde{X}_n) - R(a^*,\theta_0) > \delta\right) \\ &\quad = P_0^n\left(V_{VB}^*(\tilde{X}_n) - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a^*,\theta)] + \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a^*,\theta)] - R(a^*,\theta_0) > \delta\right) \\ &\quad \leq P_0^n\left(V_{VB}^*(\tilde{X}_n) - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a^*,\theta)] > \delta/2\right) + P_0^n\left(\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a^*,\theta)] - R(a^*,\theta_0) > \delta/2\right) \\ &\quad \leq P_0^n\left(V_{VB}^*(\tilde{X}_n) - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a^*,\theta)] > 0\right) + P_0^n\left(\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a^*,\theta)] - R(a^*,\theta_0) > \delta/2\right). \end{aligned}$$

By Lemma 4.4.7(1), for every $a \in \mathcal{A}$, $\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] \xrightarrow{P_0^n} R(a,\theta_0)$ as $n \to \infty$. Therefore it follows from the inequality above and (4.20) that

$$\lim_{n\to\infty} P_0^n \left(V_{VB}^*(\tilde{X}_n) - R(a^*,\theta_0) > \delta\right) = \lim_{n\to\infty} P_0^n \left(V_{VB}^*(\tilde{X}_n) - V^* > \delta\right) = 0. \tag{4.21}$$

We are left to show that $\lim_{n\to\infty} P_0^n \left(V_{VB}^*(\tilde{X}_n) - R(a^*,\theta_0) < -\delta\right) = \lim_{n\to\infty} P_0^n \left(V_{VB}^*(\tilde{X}_n) - V^* < -\delta\right) = 0$ for any $\delta > 0$. Let $\hat{a}_n \in \mathcal{S}_{VB}^*$; that is $Q^*\left(g(\hat{a}_n,\theta) \le 0|\tilde{X}_n\right) \ge \beta$ and $V_{VB}^*(\tilde{X}_n) = \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(\hat{a}_n,\theta)]$. Since $\mathcal{A}$ is compact, we assume that as $n \to \infty$ $\hat{a}_n \to a_0 \in \mathcal{A}$ (the limit point of the sequence $\{\hat{a}_n\} \subseteq \mathcal{A}$).

Recall that Lemma 4.4.6 holds uniformly over all $a \in \mathcal{A}$. Therefore using the fact that $Q^*\left(g(\hat{a}_n,\theta) \le 0|\mathbf{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0)) \le \left|Q^*\left(g(\hat{a}_n,\theta) \le 0|\mathbf{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0))\right| \le \sup_{a\in\mathcal{A}} \left|Q^*\left(g(a,\theta) \le 0|\mathbf{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))\right|$, we have for any $\eta > 0$,

$$\lim_{n\to\infty} P_0^n \left[Q^*\left(g(\hat{a}_n,\theta) \le 0|\mathbf{X}_n\right) \le \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0)) + \eta\right] = 1. \tag{4.22}$$

Next using the fact that $Q^*\left(g(\hat{a}_n,\theta) \le 0|\tilde{X}_n\right) \ge \beta$ for every $n \ge 1$, it follows that $\hat{a}_n$ is a feasible point of (TP) for $\eta \le \beta$; that is, $\left\{a \in \mathcal{X} : Q^*\left(g(\hat{a}_n,\theta) \le 0|\mathbf{X}_n\right) \le \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0)) + \eta\right\} \subset \{a \in \mathcal{X} : \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0)) + \eta \ge \beta\}$. Therefore, it follows that

$$\begin{aligned}\left\{Q^*\left(g(\hat{a}_n,\theta) \le 0|\mathbf{X}_n\right) \le \mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0)) + \eta\right\} &\subseteq \left\{\mathbb{I}_{(-\infty,0]}(g(\hat{a}_n,\theta_0)) + \eta \ge \beta\right\} \\ &\subseteq \left\{R(\hat{a}_n,\theta_0) \ge V^*\right\}, \end{aligned} \tag{4.23}$$

since the penultimate condition implies that $\hat{a}_n$ is a feasible point of (TP). Therefore, for any $\eta \le \beta$, $\lim_{n\to\infty} P_0^n \left[R(\hat{a}_n,\theta_0) \le V^*\right] = 0$. Using the fact that $R(\hat{a}_n,\theta_0) - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(\hat{a}_n,\theta)] \le \left|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(\hat{a}_n,\theta)] - R(\hat{a}_n,\theta_0)\right| \le \sup_{a\in\mathcal{A}} \left|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)\right|$, for any $\delta > 0$ Lemma 4.4.7(2) implies that

$$\lim_{n\to\infty} P_0^n \left[\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(\hat{a}_n,\theta)] + \delta \ge R(\hat{a}_n,\theta_0)\right] = \lim_{n\to\infty} P_0^n \left[V_{VB}^*(\mathbf{X}_n) + \delta \ge R(\hat{a}_n,\theta_0)\right] = 1,$$

and therefore $\lim_{n\to\infty} P_0^n \left[V_{VB}^*(\tilde{X}_n) + \delta \leq R(\hat{a}_n, \theta_0)\right] = 0$. Observe that for any $\delta > 0$

$$\begin{aligned} P_0^n \left[V^* - V_{VB}^*(\mathbf{X}_n) \geq \delta\right] &\leq P_0^n \left[V^* - R(\hat{a}_n, \theta_0) \geq \delta/2\right] + P_0^n \left[R(\hat{a}_n, \theta_0) - V_{VB}^*(\mathbf{X}_n) \geq \delta/2\right] \\ &\leq P_0^n \left[V^* - R(\hat{a}_n, \theta_0) \geq 0\right] + P_0^n \left[R(\hat{a}_n, \theta_0) - V_{VB}^*(\mathbf{X}_n) \geq \delta/2\right]. \end{aligned}$$

Taking limit $n \to \infty$ on either side of the inequality above, we have

$$\lim_{n\to\infty} P_0^n \left[V^* - V_{VB}^*(\mathbf{X}_n) \geq \delta\right] = 0. \tag{4.24}$$

Combining equation (4.21) and (4.24), we conclude that for any $\delta > 0$,

$$\lim_{n\to\infty} P_0^n \left[|V^* - V_{VB}^*(\mathbf{X}_n)| \geq \delta\right] = 0.$$

□

Next, we state the corollary of the result above that guarantees asymptotic consistency of the optimal value $V_B^*(\tilde{X}_n)$ of (BJCCP) with a single constraint.

Corollary 4.4.3. *Under Assumptions 4.4.5, 4.4.6, 4.4.7, and 4.4.8 and when $\mathcal{A}$ is a compact set, we have $V_B^*(\tilde{X}_n) \xrightarrow{P_0^n} V^*$ as $n \to \infty$.*

Proof. The proof follows straightforwardly from Theorem 4.4.3 and the fact that $V_{VB}^*(\tilde{X}_n)$ is the same as $V_B^*(\tilde{X}_n)$ when the variational family $\mathcal{Q}$ is fixed to the set of all possible distributions on Θ. □

4.5 Application

Data-driven chance constrained optimization problems abound throughout operations research, finance, engineering and the sciences. In this section we present an example application of Bayesian chance constrained optimization to solving a staffing problem in a queueing system.

4.5.1 Optimal Staffing

Consider a situation where a decision maker (DM) has to decide the optimal number of servers in a multi-server $M/M/c$ queueing system, using arrival times and service time data. We assume that the rate parameters of the exponentially distributed inter-arrival and service time distributions, denoted as λ and μ respectively, are unknown. Note that λ and μ, together constitute the system parameter $\xi = \{\lambda, \mu\}$ and the number of servers c is the decision/input variable. The DM collects n realizations of the random vector $\mathcal{V} := \{T, S, E\}$, denoted as $\tilde{X}_n := \{\mathcal{V}_1, \ldots \mathcal{V}_n\}$ where T, S, and E are the random variables denoting the arrival, service-start, and service-end time of each customer $\mathrm{i} \in \{1, 2, \ldots n\}$ respectively. We also assume that the inter-arrival and service times are independent, that is $T_\mathrm{i} - T_{\mathrm{i}-1}$ is independent of $E_\mathrm{i} - S_\mathrm{i}$ for each $\mathrm{i} \geq 1$. The joint likelihood of the arrival and departure times for n customers is $dP_\theta(\tilde{X}_n) := \prod_{\mathrm{i}=1}^{n} \lambda \mathrm{e}^{-\lambda(T_\mathrm{i} - T_{\mathrm{i}-1})} \mu \mathrm{e}^{-\mu(E_\mathrm{i} - S_\mathrm{i})}$.

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 steady-state probability that the customer did not wait in the queue. A closed-form expression for $1 - W_q(c, \lambda, \mu)$ for an $M/M/c$ queue is known to be (see [5])

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

where $r = \frac{\lambda}{\mu}$ and $\rho = \frac{r}{c}$ with $\rho < 1$. ρ is also known as *traffic intensity* and $\rho < 1$ is a necessary and sufficient condition for an $M/M/c$ queue to be in steady-state (or stable).

The DM fixes α, the desired maximum fraction of customers delayed in the queue and the smallest c is chosen that satisfies $(\alpha - \{1 - W_q(c, \lambda, \mu)\}) > 0$ and $(c\mu - \lambda) > 0$. Referring to the queueing literature, we will use the term the quality of service(QoS) constraint for

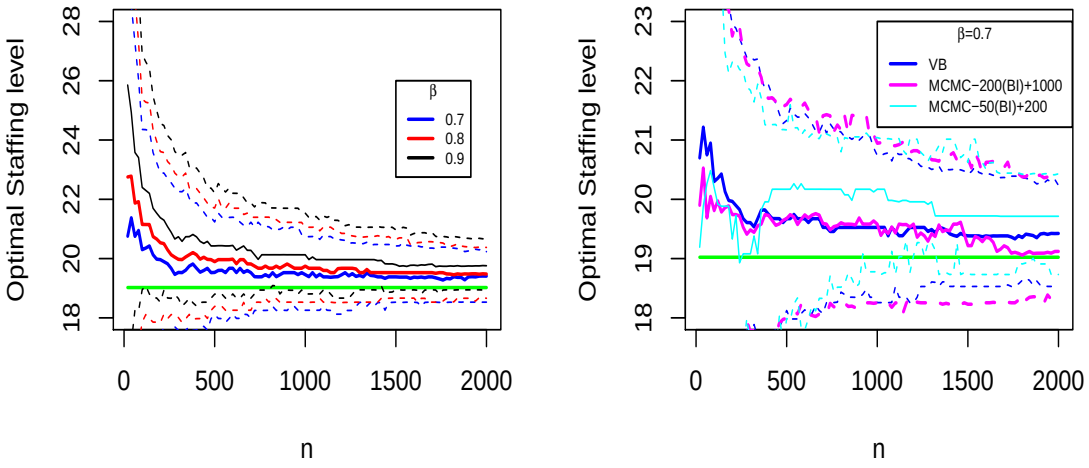
the first constraint. In fact, the QoS constraint is only valid when $\rho < 1$. The corresponding constraint optimization problem is

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

This so-called staffing problem and its variants are well studied in the queueing literature. As noted before, we are interested in the data-driven setting where the parameters of the problem are unknown. This data-driven staffing problem has been considered as well and the interested reader may referred to [4] and [6].

Next, we fix a non-conjugate inverse Gamma ($\text{Inv} - \Gamma(\cdot)$) distribution prior on both λ and μ, that is $d\Pi(\lambda, \mu) = \text{Inv} - \Gamma_\lambda(\lambda; \alpha_q, \beta_q)\text{Inv} - \Gamma_\mu(\mu; \alpha_s, \beta_s)d\lambda d\mu$. In our experiments, we fix $\alpha_q = \alpha_s = 1$ and $\beta_q = \beta_s = 1$. We fix the variational family $\mathcal{Q} = \Big\{q(\lambda, \mu) : q(\lambda, \mu; a_q, b_q, a_s, b_s) = \Gamma(\lambda; a_q, b_q)\Gamma(\mu; a_s, b_s)\Big\}$, where $\Gamma(\cdot; a_{(\cdot)}, b_{(\cdot)})$ denotes the Gamma distribution with rate $b_{(\cdot)}$ and shape $a_{(\cdot)}$. In the simulation experiment, we fix $\lambda_0 = 16$ and $\mu_0 = 1$ and generate 2000 samples of service and inter-arrival times. We then solve the (VBJCCP) for 250 sample paths and denote its solution as C^*_{VB}. We then solve the corresponding (BJCCP) using a sample average approximation (SAA) of the chance constrained problem, by generating samples from the posterior distribution using MCMC. We denote the optimal staffing level computed using MCMC as C^*_{MCMC}.

The results of this simulation experiment are summarized in Figure 4.2. We observe in Figure 4.2(a) that C^*_{VB} is consistent and moreover, for larger confidence level β, C^*_{VB} is more conservative (i.e., the optimal number of servers is larger) as expected. In Figure 4.2(b), we compare C^*_{VB} and C^*_{MCMC} for $\beta = 0.7$. We compute C^*_{MCMC} at each n using two sequences of MCMC samples from the 'true' posterior distribution generated using Metropolis–Hastings algorithm [124]: 1) 1000 samples with 200 burn-in (magenta) and 2) 200 samples with 50 burn-in (cyan). Observe that, as n increases both C^*_{VB} and C^*_{MCMC} (magenta) converges to the true solution almost at the same rate and there is no significant difference between the two approaches. In fact, we will later show in Theorem 4.5.1 and Corollary 4.5.1 that the



(a) C^*_{VB} (b) C^*_{VB} vs. C^*_{MCMC}

Figure 4.2. $\lambda_0 = 16, \mu_0 = 1$, (a) Optimal Staffing Level (5^{th}, 50^{th}, and 95^{th} quantile over 250 sample paths) for $\beta = \{0.7, 0.8, 0.9\}$ (b) C^*_{VB} vs. C^*_{MCMC} -Optimal Staffing Level (5^{th}, 50^{th}, and 95^{th} quantile over 250 sample paths) against the number of samples (n) , green line is the solution of (TP-Q) at $\{\lambda_0\, \mu_0\}$.

optimal staffing levels computed using the (VBJCCP) and (BJCCP) approaches converge at the same rate. Moreover, the average computation time taken by the VB and MCMC (magenta) approaches to compute an optimal staffing level at a given n are of the same order (30 seconds (average) on Sky Lake CPU @ 2.60GHz). Unsurprisingly, the computation time in an MCMC approach can be reduced by reducing the number of samples; however, it may result in computing a suboptimal solution. We observe that computing C^*_{MCMC} (cyan) is faster (8 seconds (on average) on Sky Lake CPU @ 2.60GHz) but suboptimal.

Next, we verify the conditions on the prior, the likelihood model and the variational family to compute the convergence rate of C^*_{VB}. First note that the risk function $f(c, \theta) = c$ in the optimal staffing problem, therefore $L^2_n(\theta, \theta_0)$ is 0. Hence, Lemma 4.4.5 is trivially true even without existence of tests conditions (Assumption 4.4.1) defined using $L_n(\theta, \theta_0) = L^2_n(\theta, \theta_0)$. Next, we consider $L^1_n(\theta, \theta_0)$ and $L^{\mathrm{i}}_n(\theta, \theta_0)$ for $\mathrm{i} \in \{1, \ldots, m\}$ and recall Proposition 4.4.2. We satisfy the conditions of Proposition 4.4.2 in the following result so that these distance functions satisfy Assumption 4.4.1.

Lemma 4.5.1. *For the sequence of tests*

$$\phi_{n,\epsilon} = \mathbb{I}_{\left\{\tilde{X}_n : \left| \frac{n}{\sum_{i=1}^{n} T_i - T_{i-1}} - \lambda_0 \right| > \lambda_0 \sqrt{\frac{n+2}{(n-2)^2}} e^{Cn\epsilon^2} \right\} \cap \left\{ \tilde{X}_n : \left| \frac{n}{\sum_{i=1}^{n} E_i - S_i} - \mu_0 \right| > \mu_0 \sqrt{\frac{n+2}{(n-2)^2}} e^{Cn\epsilon^2} \right\}},$$

it can be shown that

$$\mathbb{E}_{P_0^n}[\phi_{n,\epsilon}] \leq e^{-Kn\epsilon^2},$$

for $C = K/2$.

We assume that $\Theta_n(\epsilon) = \Theta = (0, \infty)^2$. Observe that Assumption 4.4.2 is trivially satisfied by the product of Inverse Gamma priors on λ and μ. Next, we show that the prior and the likelihood model satisfy Assumption 4.4.3.

Lemma 4.5.2. *Fix* $n_2 \geq 2$ *and any* $\rho > 1$. *Let* $A_n := \{\theta \in \Theta : D_{1+\rho}(P_0^n \| P_\theta^n) \leq C_3 n \epsilon_n^2\}$, *where* $D_{1+\rho}(P_0^n \| P_\theta^n)$ *is the Rényi divergence between* P_0^n *and* P_θ^n. *Then for* $\epsilon_n^2 = \frac{\log n}{n}$ *the prior satisfies*

$$\Pi\{A_n\} \geq \exp(-nC_2\epsilon_n^2), \forall n \geq n_2$$

with $C_3 > 4 \max\{\alpha_s^{-1}, \alpha_q^{-1}\}$ *and* $C_2 = 0.5(\alpha_s + \alpha_q)C_3$.

The results above verify the conditions required to establish the convergence rate of the optimal staffing level computed using (VBJCCP). However, to explicitly quantify the rate of convergence, we also need to identify a bound on η_n^2 using Proposition 4.4.1. Therefore, in the next result, we identify a sequence of distribution in $\mathcal{Q}$ that satisfies Assumption 4.4.4 required for Proposition 4.4.1 to hold.

Lemma 4.5.3. *Let* $\{Q_n(\lambda, \mu)\}$ *be a sequence of distributions defined as* $\Gamma(\lambda; n, n/\lambda_0)\Gamma(\mu; n, n/\mu_0)$, *then*

$$\frac{1}{n}\left[\mathrm{KL}\left(Q_n(\lambda,\mu) \| \Pi(\theta)\right) + \mathbb{E}_{Q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n(\tilde{X}_n)) \| dP_\theta^n(\tilde{X}_n)\right)\right]\right] \leq C_9 \epsilon_n^2,$$

where $\epsilon_n^2 = \frac{\log n}{n}$ *and* $C_9 = 1 + \max\left(0, 2 + \frac{2\beta_q}{\lambda_0} - \log\sqrt{2\pi} - \log\left(\frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\right) + \alpha_q \log\lambda_0\right) + \max\left(0, 2 + \frac{2\beta_s}{\mu_0} - \log\sqrt{2\pi} - \log\left(\frac{\beta_s^{\alpha_s}}{\Gamma(\alpha_s)}\right) + \alpha_s \log\mu_0\right)$ *and the parameters of the prior distribution are such that* $C_9 > 0$.

Lemmas 4.5.2 and 4.5.3, combined together, identify that the optimal staffing level computed using (VBJCCP) converges at the rate of $\epsilon_n = \sqrt{\frac{\log n}{n}}$. More formally,

Theorem 4.5.1. *For* $L_n^1(\theta, \theta_0) = n \sup_{c \in \mathcal{A}} |\mathbb{I}_{(-\infty,0]}(1 - W_q(c, \lambda, \mu) - \alpha) - \mathbb{I}_{(-\infty,0]}(1 - W_q(c, \lambda_0, \mu_0) - \alpha)|$ *and* $L_n^2(\theta, \theta_0) = n \sup_{c \in \mathcal{A}} |c - c| = 0$, *where* $\mathcal{A}$ *is a finite set of positive integers, there exists a constant* $M > 0$ *(that depends on all the fixed hyper-parameters), such that for any* $\eta > 0$,

$$P_0^n[|C_{VB}^*(\tilde{X}_n) - C^*| > 2\eta] \leq M\epsilon_n^2,$$

where $\epsilon_n^2 = \frac{\log n}{n}$.

Proof. The proof is a direct consequence of Lemmas 4.5.1, 4.5.2, 4.5.3, Propositions 4.4.1, 4.4.2, and Theorem 4.4.2. □

Using the result above, we can directly establish the following result that quantifies the convergence rate of optimal staffing level computed using (BJCCP) approach.

Corollary 4.5.1. *For* $L_n^1(\theta, \theta_0) = n \sup_{c \in \mathcal{A}} |\mathbb{I}_{(-\infty,0]}(1 - W_q(c, \lambda, \mu) - \alpha) - \mathbb{I}_{(-\infty,0]}(1 - W_q(c, \lambda_0, \mu_0) - \alpha)|$ *and* $L_n^2(\theta, \theta_0) = n \sup_{c \in \mathcal{A}} |c - c| = 0$, *where* $\mathcal{A}$ *is a finite set of positive integers, there exists a constant* $\bar{M} > 0$ *(that depends on all the fixed hyper parameters), such that for any* $\eta > 0$,

$$P_0^n[|C_B^*(\tilde{X}_n) - C^*| > 2\eta] \leq \bar{M}\epsilon_n^2,$$

where C_B^* *is the optimal staffing level computed using (BJCCP) and* $\epsilon_n^2 = \frac{\log n}{n}$.

Proof. The proof follows straightforwardly from Theorem 4.5.1 and the fact that $q^*(\theta|\tilde{X}_n)$ is the same as the true posterior distribution when the variational family $\mathcal{Q}$ is fixed to all possible distributions. □

Next, we discuss that the prior, the likelihood model, and the variational family easily satisfy Assumptions 4.4.5, 4.4.6, and 4.4.7, that are required to show consistency of C_{VB}^*.

Notice that the prior density $\Pi(\lambda,\mu) = \text{Inv}-\Gamma_\lambda(\lambda;\alpha_q,\beta_q)\text{Inv}-\Gamma_\mu(\mu;\alpha_s,\beta_s)$ is continuous in $\theta = \{\lambda,\mu\}$ and places positive mass in the neighbourhood of the true parameter θ_0 and moreover it is bounded, therefore it satisfies Assumption 4.4.5. The exponential models are twice continuously differentiable therefore it satisfies the LAN condition in Assumption 4.4.6. Moreover, the variational family, the product of Gamma distributions on λ and μ, is absolutely continuous with respect to the prior distribution and also consists of a sequence of distribution that converges at the true parameter at the rate of $\sqrt{n}$ (refer the construction in Lemma 4.5.3). Therefore, the $\mathcal{Q}$ satisfies Assumption 4.4.7. Under these assumptions, it can be shown using the result in Theorem 4.4.3 that the optimal number of servers computed using (VBJCCP) (and (BJCCP)) are consistent.

4.6 Proofs

Proof of Lemma 4.4.2. First observe that

$$\begin{aligned}
\int_\Theta L_n^1(\theta,\theta_0)q^*(\theta|\tilde{X}_n)d\theta &= n\int_\Theta \sup_{a\in\mathcal{A}} |\mathbb{I}_{(-\infty,0]}(g(a,\theta)) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))|q^*(\theta|\tilde{X}_n)d\theta \\
&\geq n\sup_{a\in\mathcal{A}}\int_\Theta |\mathbb{I}_{(-\infty,0]}(g(a,\theta)) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))|q^*(\theta|\tilde{X}_n)d\theta \\
&\geq n\sup_{a\in\mathcal{A}}\left|\int_\Theta \left(\mathbb{I}_{(-\infty,0]}(g(a,\theta)) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))\right) q^*(\theta|\tilde{X}_n)d\theta\right| \\
&= n\sup_{a\in\mathcal{A}}\left|\int_\Theta \mathbb{I}_{(-\infty,0]}(g(a,\theta))q^*(\theta|\tilde{X}_n)d\theta - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))\right| \\
&= n\sup_{a\in\mathcal{A}}\left|Q^*\left(g(a,\theta)\leq 0|\tilde{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))\right|.
\end{aligned}$$

Now using Theorem 4.4.1 and the inequality above, it is straightforward to observe that

$$\begin{aligned}
\mathbb{P}_0[\sup_{a\in\mathcal{A}}|Q^*\left(g(a,\theta)\leq 0|\tilde{X}_n\right) - \mathbb{I}_{(-\infty,0]}(g(a,\theta_0))| > \delta] &\leq \mathbb{P}_0\left[\int_\Theta L_n^1(\theta,\theta_0)q^*(\theta|\tilde{X}_n)d\theta > n\delta\right] \\
&\leq \frac{M_1}{\delta}(\epsilon_n^2+\eta_n^2).
\end{aligned}$$

□

Proof of Lemma 4.4.3. Proof is similar to Lemma 4.4.2 hence omitted. □

Proof of Proposition 4.4.1. The proof follows straightforwardly using the definition of η_n^2 and Assumption 4.4.4. □

Proof of Proposition 4.4.2. Note that consistent tests always exist for finite-dimensional models on fixed null and alternate sets; for instance, the Kolmogorov-Smirnov test statistic [109, Theorem 19.1]. Therefore, the condition of Lemma 4.4.4 is always satisfied for finite dimensional (or parametric) models. Now for distance functions $L_n^1(\theta, \theta_0)$ in Theorem 4.4.2 and $L_n^{\mathrm{i}}(\theta, \theta_0)$ in Theorem 4.4.1 fix $\mathcal{P}_0 = P_0$ and $\mathcal{P}_1 = \{P_\theta : L_n^{(\cdot)}(\theta, \theta_0) > n\epsilon^2\}$, where we use $L_n^{(\cdot)}$ to reference either $L_n^1(\theta, \theta_0)$ or $L_n^{\mathrm{i}}(\theta, \theta_0)$ for brevity. Note that for any $\epsilon \in (0, 1]$, $\mathcal{P}_1$ is fixed. Therefore, it follows from Lemma 4.4.4 that for any $\epsilon \in (\epsilon_n, 1]$,

$$\mathbb{E}_{P_0}[\phi_n] \leq \mathrm{e}^{-Kn} \leq \mathrm{e}^{-Kn\epsilon^2} \text{ and } \sup_{P \in \mathcal{P}_1} \mathbb{E}_P[1 - \phi_n] \leq \mathrm{e}^{-Kn} \leq \mathrm{e}^{-Kn\epsilon^2}.$$

For $\epsilon > 1$, by assumption in the assertion of the proposition we have,

$$\mathbb{E}_{P_0}[\phi_{n,\epsilon}] \leq \mathrm{e}^{-Kn\epsilon^2} \text{ and } \sup_{P \in \mathcal{P}_1} \mathbb{E}_P[1 - \phi_{n,\epsilon}] = 0 \leq \mathrm{e}^{-Kn\epsilon^2},$$

where the second equality follows since $\mathcal{P}_1$ is null set for $\epsilon > 1$. Therefore, it follows that there exists a test $\phi_{n,\epsilon} = \phi_n \mathbb{I}_{\{\epsilon \in (0,1])\}} + \phi_{n,\epsilon} \mathbb{I}_{\{\epsilon \in (1,\infty)\}}$ such that distance function $L_n^{(\cdot)}$ satisfies Assumption 4.4.1. □

Proof of Lemma 4.4.6. Lemma 4.4.5 implies that the VB approximate posterior $q^*(\theta|\tilde{X}_n)$ is consistent, and it follows from Definition 4.2.1 that for every $\eta > 0$

$$\int_{\|\theta - \theta_0\| > \eta} q^*(\theta|\tilde{X}_n) d\theta \overset{P_0^n}{\to} 0 \text{ as } n \to \infty. \tag{4.25}$$

In fact, $q^*(\theta|\tilde{X}_n)$ converges pointwise to δ_{θ_0} almost everywhere with respect to Lebesgue measure. Consequently, Scheffé's lemma [109, Corollary 2.30] implies that $q^*(\theta|\tilde{X}_n)$ converges to δ_{θ_0} in total-variation distance, that is

$$d_{TV}(q^*(\theta|\tilde{X}_n), \delta_{\theta_0}) = \sup_{A \subseteq \Theta} |Q^*(A|\tilde{X}_n) - \delta_{\theta_0}(A)| \xrightarrow{P_0^n} 0 \text{ as } n \to \infty, \tag{4.26}$$

where for any set $A \subseteq \Theta$, $Q^*(A|\tilde{X}_n) = \int_A q^*(\theta|\tilde{X}_n)d\theta$. Using this observation note that

$$\begin{aligned}
&\sup_{a \in \mathcal{A}} \left| \int_\Theta \prod_{\mathrm{i}=1}^m \mathbb{I}_{(-\infty,0]}(g_{\mathrm{i}}(a,\theta)) q^*(\theta|\tilde{X}_n)d\theta - \prod_{\mathrm{i}=1}^m \mathbb{I}_{(-\infty,0]}(g_{\mathrm{i}}(a,\theta_0)) \right| \\
&\quad = \sup_{a \in \mathcal{A}} |Q^*(\cap_{\mathrm{i}=1}^m \{g_{\mathrm{i}}(a,\theta) < 0\} - \delta_{\theta_0}(\cap_{\mathrm{i}=1}^m \{g_{\mathrm{i}}(a,\theta) < 0\}| \\
&\quad = |Q^*(\cap_{\mathrm{i}=1}^m \{g_{\mathrm{i}}(\bar{a},\theta) < 0\} - \delta_{\theta_0}(\cap_{\mathrm{i}=1}^m \{g_{\mathrm{i}}(\bar{a},\theta) < 0\}| \\
&\quad \leq d_{TV}(q^*(\theta|\tilde{X}_n), \delta_{\theta_0}),
\end{aligned} \tag{4.27}$$

for some $\bar{a} \in \mathcal{A}$ at which supremum is attained in the RHS of the first equality above. Now the result follows straightforwardly from (4.26).

□

Proof of Lemma 4.4.7. **Part 1: Point-wise convergence** The proof uses similar ideas as used in the proof of [113, Theorem 3.7]. Fix $a \in \mathcal{A}$. Due to Assumption 4.4.8(3), $R(a,\theta)$ is uniformly integrable with respect to any $q \in \mathcal{Q}$, which implies that for $q^*(\theta|\tilde{X}_n)$ and for any $\epsilon > 0$, there exists a compact set K_ϵ such that for all $n \geq 1$ $\int_{\Theta \setminus K_\epsilon} |R(a,\theta)| q^*(\theta|\tilde{X}_n)d\theta < \epsilon$.

Now fix $\gamma_\epsilon := \max_{\theta \in K_\epsilon} |R(a,\theta)|$. Note that $\gamma_\epsilon < +\infty$, since K_ϵ is compact and $R(a,\cdot)$ is a continuous mapping for any $x \in \mathcal{A}$. Define $R_\epsilon(a,\theta)$ be the truncation of $R(a,\theta)$, that is

$$R_\epsilon(a,\theta) = \begin{cases} R(a,\theta) & \text{if } |R(a,\theta)| < \gamma_\epsilon \\ \gamma_\epsilon & \text{if } R(a,\theta) > \gamma_\epsilon \\ -\gamma_\epsilon & \text{if } R(a,\theta) < -\gamma_\epsilon. \end{cases} \tag{4.28}$$

It follows from the definition above that $|R_\epsilon(a,\theta)| \leq |R(a,\theta)|$, which implies that

$$\int_{\Theta \setminus K_\epsilon} |R_\epsilon(a,\theta)| q^*(\theta|\tilde{X}_n) d\theta < \epsilon \tag{4.29}$$

Note the $R_\epsilon(a,\theta)$ is bounded and continuous in θ, therefore, it follows using the definition of weak convergence and Lemma 4.4.5 that

$$\lim_{n\to\infty} \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] \overset{P_0^n}{=} R_\epsilon(a,\theta_0). \tag{4.30}$$

Next observe that

$$\begin{aligned}
&|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)| \\
&\quad = \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] + \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0) \\
&\qquad\qquad + R_\epsilon(a,\theta_0) - R(a,\theta_0)\Big| \\
&\quad \leq \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)]\Big| + \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0)\Big| \\
&\qquad\qquad + |R_\epsilon(a,\theta_0) - R(a,\theta_0)| \\
&\quad = \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)]\Big| + \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0)\Big| \\
&\qquad\qquad + |R_\epsilon(a,\theta_0) - R(a,\theta_0)|\,.
\end{aligned} \tag{4.31}$$

Now using the definition of $R_\epsilon(a,\theta)$ note that

$$\begin{aligned}
\Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)]\Big| &= \left|\int_{\Theta\setminus K_\epsilon} (R(a,\theta) - R_\epsilon(a,\theta)) q^*(\theta|\tilde{X}_n) d\theta\right| \\
&\leq \int_{\Theta\setminus K_\epsilon} |R(a,\theta)| q^*(\theta|\tilde{X}_n) d\theta + \int_{\Theta\setminus K_\epsilon} |R_\epsilon(a,\theta)| q^*(\theta|\tilde{X}_n) d\theta \quad \leq 2\epsilon.
\end{aligned}$$

Similarly, $|R_\epsilon(a,\theta_0) - R(a,\theta_0)| \leq 2\epsilon$, since due to Assumption 4.4.8(3) $\int_{\Theta\setminus K_\epsilon} |R(a,\theta)| q^*(\theta|\tilde{X}_n) d\theta < \epsilon$ is true for all $n \geq 1$ and consequently for δ_{θ_0} as well. Hence, substituting the above two observations into (4.31) yields

$$|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)| \leq 4\epsilon + \Big|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0)\Big|\,.$$

Consequently, it follows for any $\epsilon > 0$ that,

$$P_0^n\left(|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a,\theta)] - R(a,\theta_0)| > 5\epsilon\right) \leq P_0^n\left(|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R_\epsilon(a,\theta)] - R_\epsilon(a,\theta_0)| > \epsilon\right). \quad (4.32)$$

Now taking limits $n \to \infty$ on either side of the inequality above, the result follows straightforwardly using the observation in (4.30).

Part 2: Uniform convergence:

Since $\mathcal{A}$ is compact and $R(a,\theta_0)$ is continuous in a, using Corollary 2.2 in [114] the uniform convergence follows from point-wise convergence (Part 1) if there exist a bounded sequence B_n and for all $a_1, a_2 \in \mathcal{A}$, $|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a_1,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a_2,\theta)]| \leq B_n\|a_1 - a_2\|$. Since, $R(a,\theta)$ is locally Lipschitz in a due to Assumption 4.4.8(2), therefore for $a_1, a_2 \in \mathcal{A}$,

$$\begin{aligned}|\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a_1,\theta)] - \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[R(a_2,\theta)]| &\leq \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[|R(a_1,\theta) - R(a_2,\theta)|] \\ &\leq \mathbb{E}_{q^*(\theta|\tilde{X}_n)}[K_{\mathcal{A}}(\theta)]\|a_1 - a_2\|. \quad (4.33)\end{aligned}$$

The uniform convergence follows since by Assumption 4.4.8(2) $\mathbb{E}_{q^*(\theta|\tilde{X}_n)}[K_{\mathcal{A}}(\theta)] \leq \bar{K}_{\mathcal{A}}$. □

Proof of Lemma 4.5.1. Due to independence of arrival and service time distributions, first note that

$$\begin{aligned}\mathbb{E}_{P_0^n}[\phi_{n,\epsilon}] =& P_0^n\left(\tilde{X}_n : \left|\frac{n}{\sum_{\mathrm{i}=1}^n T_\mathrm{i} - T_{\mathrm{i}-1}} - \lambda_0\right| > \lambda_0\sqrt{\frac{n+2}{(n-2)^2}}\mathrm{e}^{Cn\epsilon^2}\right) \\ &\times P_0^n\left(\tilde{X}_n : \left|\frac{n}{\sum_{\mathrm{i}=1}^n E_\mathrm{i} - S_\mathrm{i}} - \mu_0\right| > \mu_0\sqrt{\frac{n+2}{(n-2)^2}}\mathrm{e}^{Cn\epsilon^2}\right)\end{aligned}$$

Denote $\xi_\mathrm{i} = T_\mathrm{i} - T_{\mathrm{i}-1}$. Using Chebyschev's inequality observe that

$$\begin{aligned}&P_0^n\left(\left|\frac{n}{\sum_{\mathrm{i}=1}^n \xi_\mathrm{i}} - \lambda_0\right| > \lambda_0\sqrt{\frac{n+2}{(n-2)^2}}\mathrm{e}^{Cn\epsilon^2}\right) \\ &\leq \frac{(n-2)^2}{\lambda_0^2(n+2)}\mathrm{e}^{-2Cn\epsilon^2}\mathbb{E}_{P_0^n}\left[\left|\frac{n}{\sum_{\mathrm{i}=1}^n \xi_\mathrm{i}} - \lambda_0\right|^2\right] \\ &= \frac{(n-2)^2}{\lambda_0^2(n+2)}\mathrm{e}^{-2Cn\epsilon^2}\mathbb{E}_{P_0^n}\left[\left(\frac{n}{\sum_{\mathrm{i}=1}^n \xi_\mathrm{i}}\right)^2 + \lambda_0^2 - \left(\frac{2n\lambda_0}{\sum_{\mathrm{i}=1}^n \xi_\mathrm{i}}\right)\right].\end{aligned}$$

Now using the fact that the sum of n i.i.d exponential random variable with rate parameter λ_0 is Gamma distributed with rate and shape parameter λ_0 and n (respectively), we obtain that the RHS in the equation above is bounded above by

$$\frac{(n-2)^2}{\theta_0^2(n+2)}\mathrm{e}^{-2Cn\epsilon^2}\theta_0^2\left[\frac{n^2}{(n-1)(n-2)}+1-\frac{2n}{n-2}\right]=\frac{(n-2)^2}{n+2}\mathrm{e}^{-2Cn\epsilon^2}\left[\frac{n+2}{(n-1)(n-2)}\right]$$
$$\leq \mathrm{e}^{-2Cn\epsilon^2}. \tag{4.34}$$

Now, choosing $C = K/2$, we have

$$\mathbb{E}_{P_0^n}[\phi_{n,\epsilon}] \leq \mathrm{e}^{-Kn\epsilon^2},$$

and the proposition follows. □

Proof of Lemma 4.5.2. First, we write the Rényi divergence between P_0^n and P_θ^n,

$$D_{1+\rho}\left(P_0^n\|P_\theta^n\right)=\frac{1}{\rho}\log\int\left(\frac{dP_0^n}{dP_\theta^n}\right)^\rho dP_0^n = n\frac{1}{\rho}\log\int\left(\frac{dP_{\lambda_0}}{dP_\lambda}\right)^\rho dP_0+n\frac{1}{\rho}\log\int\left(\frac{dP_{\mu_0}}{dP_\mu}\right)^\rho dP_{\mu_0}$$
$$= n\left(\log\frac{\lambda_0}{\lambda}+\frac{1}{\rho}\log\frac{\lambda_0}{(\rho+1)\lambda_0-\rho\lambda}\right)$$
$$+n\left(\log\frac{\mu_0}{\mu}+\frac{1}{\rho}\log\frac{\mu_0}{(\rho+1)\mu_0-\rho\mu}\right),$$

when $((\rho+1)\lambda_0-\rho\lambda)>0$ and $((\rho+1)\lambda_0-\rho\lambda)>0$, otherwise $D_{1+\rho}\left(P_0^n\|P_\theta^n\right)=\infty$. Using the straightforward inequality for two independent random variables A and B that $P(A+B\leq 2c)\geq P(\{A\leq c\}\cup\{B\leq c\})=P(\{A\leq c\})P(\{B\leq c\})$, it follows that

$$\Pi(D_{1+\rho}\left(P_0^n\|P_\theta^n\right)\leq C_3n\epsilon_n^2)\geq \mathrm{Inv}-\Gamma_\lambda(D_{1+\rho}\left(P_{\lambda_0}^n\|P_\theta^n\right)\leq 0.5C_3n\epsilon_n^2)\times$$
$$\mathrm{Inv}-\Gamma_\mu(D_{1+\rho}\left(P_{\mu_0}^n\|P_\theta^n\right)\leq 0.5C_3n\epsilon_n^2). \tag{4.35}$$

Now consider the first term of the product in the RHS of the equation above. Observe that, $D_{1+\rho}\left(P_0^n \| P_\lambda^n\right)$ is non-decreasing in ρ (this also follows from non-decreasing property of the Rényi divergence with respect to ρ). Therefore, observe that

$$\begin{aligned}\mathrm{Inv}-\Gamma_\lambda(D_{1+\rho}\left(P_0^n \| P_\lambda^n\right) \leq 0.5C_3 n\epsilon_n^2) &\geq \mathrm{Inv}-\Gamma_\lambda(D_{\infty}\left(P_0^n \| P_\lambda^n\right) \leq 0.5C_3 n\epsilon_n^2) \\ &= \mathrm{Inv}-\Gamma_\lambda\left(0 \leq \log\frac{\lambda_0}{\lambda} \leq 0.5C_3\epsilon_n^2\right) \\ &= \mathrm{Inv}-\Gamma_\lambda\left(\lambda_0 \mathrm{e}^{-0.5C_3\epsilon_n^2} \leq \lambda \leq \lambda_0\right).\end{aligned}$$

The cumulative distribution function of inverse-gamma distribution is $\mathrm{Inv}-\Gamma_\lambda(\{\lambda < t\}) := \frac{\Gamma\left(\alpha_q, \frac{\beta_q}{t}\right)}{\Gamma(\alpha_q)}$, where $\alpha_q(>0)$ is the shape parameter, $\beta_q(>0)$ is the scale parameter, $\Gamma(\cdot)$ is the Gamma function, and $\Gamma(\cdot,\cdot)$ is the incomplete Gamma function. Therefore, it follows for $\alpha \geq 1$ that

$$\begin{aligned}\mathrm{Inv}-\Gamma_\lambda\left(\lambda_0 \mathrm{e}^{-0.5C_3\epsilon_n^2} \leq \lambda \leq \lambda_0\right) &= \frac{\Gamma\left(\alpha_q, \beta_q/\lambda_0\right) - \Gamma\left(\alpha_q, \beta_q/\lambda_0 \mathrm{e}^{0.5C_3\epsilon_n^2}\right)}{\Gamma(\alpha_q)} \\ &= \frac{\int_{\beta_q/\lambda_0}^{\beta_q/\lambda_0 \mathrm{e}^{0.5C_3\epsilon_n^2}} \mathrm{e}^{-x} x^{\alpha_q - 1} dx}{\Gamma(\alpha_q)} \\ &\geq \frac{\mathrm{e}^{-\beta_q/\lambda_0 \mathrm{e}^{0.5C_3\epsilon_n^2} + \alpha_q 0.5C_3\epsilon_n^2}}{\alpha_q \Gamma(\alpha_q)} \left(\frac{\beta_q}{\lambda_0}\right)^{\alpha_q} \left[1 - \mathrm{e}^{-\alpha_q 0.5C_3\epsilon_n^2}\right] \\ &\geq \frac{\mathrm{e}^{-\beta_q/\lambda_0 \mathrm{e}^{0.5C_3}}}{\alpha_q \Gamma(\alpha_q)} \left(\frac{\beta_q}{\lambda_0}\right)^{\alpha_q} \left[\mathrm{e}^{-\alpha_q 0.5C_3 n\epsilon_n^2}\right]\end{aligned}$$

where the penultimate inequality follows since $0 < \epsilon_n^2 < 1$ and the last inequality follows from the fact that, $1-\mathrm{e}^{-\alpha_q 0.5C_3\epsilon_n^2} \geq \mathrm{e}^{-\alpha_q 0.5C_3 n\epsilon_n^2}$, for large enough n. Also note that, $1-\mathrm{e}^{-\alpha_q 0.5C_3\epsilon_n^2} \geq \mathrm{e}^{-\alpha_q 0.5C_3 n\epsilon_n^2}$ can't hold true for $\epsilon_n^2 = 1/n$. However, for $\epsilon_n^2 = \frac{\log n}{n}$ it holds for any $n \geq 2$ when $\alpha_q C_3 > 4$. Using similar steps as above we can also bound

$$\mathrm{Inv}-\Gamma_\mu(D_{1+\rho}\left(P_{\mu_0}^n \| P_\mu^n\right) \leq 0.5C_3 n\epsilon_n^2) \geq \frac{\mathrm{e}^{-\beta_s/\mu_0 \mathrm{e}^{0.5C_3}}}{\alpha_s \Gamma(\alpha_s)} \left(\frac{\beta_s}{\mu_0}\right)^{\alpha_s} \left[\mathrm{e}^{-\alpha_s 0.5C_3 n\epsilon_n^2}\right],$$

for $\alpha_s C_3 > 4$ Therefore, substituting the above two results we have for the prior distribution defined as the product of two inverse-Gamma priors on λ and μ, $C_3 > 4\max(\alpha_s^{-1}, \alpha_q^{-1})$, $C_2 = 0.5(\alpha_q + \alpha_s)C_3$ and any $\rho > 1$ the result follows for sufficiently large n. □

Proof of Lemma 4.5.3. Since family $\mathcal{Q}$ contains all product Gamma distributions, observe that $\{q_n(\cdot) \in \mathcal{Q}\}\forall n \geq 1$. First, due to independence of queue and server data observe that

$$\begin{aligned}
&\mathrm{KL}\left(q_n(\lambda,\mu)\|\boldsymbol{\pi}(\theta)\right) + \mathbb{E}_{q_n(\theta)}\left[\mathrm{KL}\left(dP_0^n(\tilde{X}_n))\|dP_\theta^n(\tilde{X}_n)\right)\right] \\
&\quad = \mathrm{KL}\left(q_n(\lambda)\|\boldsymbol{\pi}(\lambda)\right) + \mathbb{E}_{q_n(\lambda)}\left[\mathrm{KL}\left(dP_{\lambda_0}^n(\tilde{X}_n(q)))\|dP_\lambda^n(\tilde{X}_n(q))\right)\right] \qquad (4.36) \\
&\qquad + \mathrm{KL}\left(q_n(\mu)\|\boldsymbol{\pi}(\mu)\right) + \mathbb{E}_{q_n(\mu)}\left[\mathrm{KL}\left(dP_{\mu_0}^n(\tilde{X}_n(s)))\|dP_\mu^n(\tilde{X}_n(s))\right)\right], \qquad (4.37)
\end{aligned}$$

where $q_n(\cdot) = \frac{n^n}{(\cdot)_0^n\Gamma(n)}(\cdot)^{n-1}\mathrm{e}^{-n\frac{(\cdot)}{(\cdot)_0}}$, $\tilde{X}_n(q)$ and $\tilde{X}_n(s)$ denote the data pertaining to arrival and service times respectively, $\boldsymbol{\pi}(\cdot)$ denote the $\mathrm{Inv}-\Gamma.$ prior. Now consider the first term in (4.36); using the definition of the KL divergence it follows that

$$\mathrm{KL}(q_n(\lambda)\|\boldsymbol{\pi}(\lambda)) = \int q_n(\lambda)\log(q_n(\lambda))d\lambda - \int q_n(\lambda)\log(\boldsymbol{\pi}(\lambda))d\lambda. \qquad (4.38)$$

Substituting $q_n(\lambda)$ in the first term of the equation above and expanding the logarithm term, we obtain

$$\begin{aligned}
\int q_n(\lambda)\log(q_n(\lambda))d\lambda &= (n-1)\int \log\lambda \frac{n^n}{\lambda_0^n\Gamma(n)}\lambda^{n-1}\mathrm{e}^{-n\frac{\lambda}{\lambda_0}}d\lambda - n + \log\left(\frac{n^n}{\lambda_0^n\Gamma(n)}\right) \\
&= -\log\lambda_0 + (n-1)\int \log\frac{\lambda}{\lambda_0}\frac{n^n}{\lambda_0^n\Gamma(n)}\lambda^{n-1}\mathrm{e}^{-n\frac{\lambda}{\lambda_0}}d\lambda - n + \log\left(\frac{n^n}{\Gamma(n)}\right)
\end{aligned} \qquad (4.39)$$

Now consider the second term in the equation above. Substitute $\lambda = \frac{t\lambda_0}{n}$ into the integral, we have

$$\begin{aligned}
\int \log\frac{\lambda}{\lambda_0}\frac{n^n}{\lambda_0^n\Gamma(n)}\lambda^{n-1}\mathrm{e}^{-n\frac{\lambda}{\lambda_0}}d\lambda &= \int \log\frac{t}{n}\frac{1}{\Gamma(n)}t^{n-1}\mathrm{e}^{-t}dt \\
&\leq \int\left(\frac{t}{n}-1\right)\frac{1}{\Gamma(n)}t^{n-1}\mathrm{e}^{-t}dt = 0.
\end{aligned} \qquad (4.40)$$

Substituting the above result into (4.39), we get

$$\begin{aligned}\int q_n(\lambda)\log(q_n(\lambda))d\lambda &\leq -\log\lambda_0 - n + \log\left(\frac{n^n}{\Gamma(n)}\right)\\ &\leq -\log\lambda_0 - n + \log\left(\frac{n^n}{\sqrt{2\pi n}n^{n-1}\mathrm{e}^{-n}}\right)\\ &= -\log\sqrt{2\pi}\lambda_0 + \frac{1}{2}\log n, \end{aligned} \tag{4.41}$$

where the second inequality uses the fact that $\sqrt{2\pi n}n^n\mathrm{e}^{-n} \leq n\Gamma(n)$. Recall $\boldsymbol{\pi}(\lambda) = \frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\lambda^{-\alpha_q-1}\mathrm{e}^{-\frac{\beta_q}{\lambda}}$. Now consider the second term in (4.38). Using the definition of inverse-gamma prior and expanding the logarithm function, we have

$$\begin{aligned}-\int q_n(\lambda)\log(\boldsymbol{\pi}(\lambda))d\lambda &= -\log\left(\frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\right) + (\alpha_q+1)\int \log\lambda \frac{n^n}{\lambda_0^n\Gamma(n)}\lambda^{n-1}\mathrm{e}^{-n\frac{\lambda}{\lambda_0}}d\lambda\\ &\qquad + \beta_q\frac{n}{(n-1)\lambda_0}\\ &= -\log\left(\frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\right) + (\alpha_q+1)\int \log\frac{\lambda}{\lambda_0} \frac{n^n}{\lambda_0^n\Gamma(n)}\lambda^{n-1}\mathrm{e}^{-n\frac{\lambda}{\lambda_0}}d\lambda\\ &\quad + \beta_q\frac{n}{(n-1)\lambda_0} + (\alpha_q+1)\log\lambda_0\\ &\leq -\log\left(\frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\right) + \beta_q\frac{n}{(n-1)\lambda_0} + (\alpha_q+1)\log\lambda_0, \end{aligned} \tag{4.42}$$

where the last inequality follows from the observation in (4.40). Substituting (4.42) and (4.41) into (4.38) and dividing either sides by n, we obtain

$$\begin{aligned}&\frac{1}{n}\mathrm{KL}(q_n(\lambda)\|\boldsymbol{\pi}(\lambda))\\ &\qquad \leq \frac{1}{n}\left(-\log\sqrt{2\pi}\lambda_0 + \frac{1}{2}\log n - \log\left(\frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\right) + \beta_q\frac{n}{(n-1)\lambda_0} + (\alpha_q+1)\log\lambda_0\right)\\ &\qquad = \frac{1}{2}\frac{\log n}{n} + \frac{\beta_q}{(n-1)\lambda_0} + \frac{1}{n}\left(-\log\sqrt{2\pi} - \log\left(\frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\right) + (\alpha_q)\log\lambda_0\right). \end{aligned} \tag{4.43}$$

Now, consider the second term in (4.36). Since the observations are independent and identically distributed, we obtain

$$\frac{1}{n}\mathbb{E}_{q(\lambda)}\left[\mathrm{KL}\left(dP^n_{\lambda_0}\|p(\tilde{X}_n|\lambda)\right)\right] = \mathbb{E}_{q_n(\lambda)}\left[\mathrm{KL}\left(dP_{\lambda_0}\|p(\xi|\lambda)\right)\right]$$

Now using the expression for KL divergence between the two exponential distributions, we have

$$\begin{aligned}\frac{1}{n}\mathbb{E}_{q(\lambda)}\left[\mathrm{KL}\left(dP^n_{\lambda_0}\|p(\tilde{X}_n|\lambda)\right)\right] &= \int\left(\log\frac{\lambda_0}{\lambda}+\frac{\lambda}{\lambda_0}-1\right)\frac{n^n}{\lambda_0^n\Gamma(n)}\lambda^{n-1}\mathrm{e}^{-n\frac{\lambda}{\lambda_0}}d\lambda \\ &\leq \frac{n}{n-1}+1-2=\frac{1}{n-1},\end{aligned} \tag{4.44}$$

where second inequality uses the fact that $\log x \leq x-1$. Combined together (4.45) and (4.43) for $n \geq 2$ implies that

$$\begin{aligned}&\frac{1}{n}\left[\mathrm{KL}\left(q(\lambda)\|\boldsymbol{\pi}(\lambda)\right)+\mathbb{E}_{q(\lambda)}\left[\mathrm{KL}\left(dP^n_{\lambda_0})\|p(\tilde{X}_n|\lambda)\right)\right]\right] \\ &\quad\leq \frac{1}{2}\frac{\log n}{n}+\frac{1}{n}\left(2+\frac{2\beta_q}{\lambda_0}-\log\sqrt{2\boldsymbol{\pi}}-\log\left(\frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\right)+\alpha_q\log\lambda_0\right)\leq C_9\frac{\log n}{n}.\end{aligned} \tag{4.45}$$

where $C_9 := \frac{1}{2}+\max\left(0, 2+\frac{2\beta_q}{\lambda_0}-\log\sqrt{2\boldsymbol{\pi}}-\log\left(\frac{{\beta_q}^{\alpha_q}}{\Gamma(\alpha_q)}\right)+\alpha_q\log\lambda_0\right)$. Now using similar arguments as used for (4.36), we can bound (4.36) as

$$\begin{aligned}&\frac{1}{n}\left[\mathrm{KL}\left(q(\mu)\|\boldsymbol{\pi}(\mu)\right)+\mathbb{E}_{q(\mu)}\left[\mathrm{KL}\left(dP^n_{\mu_0})\|p(\tilde{X}_n|\mu)\right)\right]\right] \\ &\quad\leq \frac{1}{2}\frac{\log n}{n}+\frac{1}{n}\left(2+\frac{2\beta_s}{\mu_0}-\log\sqrt{2\boldsymbol{\pi}}-\log\left(\frac{{\beta_s}^{\alpha_s}}{\Gamma(\alpha_s)}\right)+\alpha_s\log\mu_0\right)\leq C_9\frac{\log n}{n}.\end{aligned} \tag{4.46}$$

where $C_9 := \frac{1}{2}+\max\left(0, 2+\frac{2\beta_s}{\mu_0}-\log\sqrt{2\boldsymbol{\pi}}-\log\left(\frac{{\beta_s}^{\alpha_s}}{\Gamma(\alpha_s)}\right)+\alpha_s\log\mu_0\right)$. Combining the above two results the proposition follows with $\epsilon_n = \frac{\log n}{n}$, and $C_9 = C_9 + C_9$. □

5. ASYMPTOTIC CONSISTENCY OF $\alpha-$ RÉNYI-APPROXIMATE POSTERIORS

In this chapter, we study the asymptotic consistency properties of α-Rényi approximate posteriors, a class of variational Bayesian methods that approximate an intractable Bayesian posterior with a member of a tractable family of distributions, the member chosen to minimize the α-Rényi divergence from the true posterior. Unique to our work is that we consider settings with $\alpha > 1$, resulting in approximations that upperbound the log-likelihood, and consequently have wider spread than traditional variational approaches that minimize the KL divergence from the posterior. Our primary result identifies sufficient conditions under which consistency holds, centering around the existence of a 'good' sequence of distributions in the approximating family that possesses, among other properties, the right rate of convergence to a limit distribution. We further characterize the good sequence by demonstrating that a sequence of distributions that converges too quickly cannot be a good sequence. We also extend our analysis to the setting where α equals one, corresponding to the minimizer of the reverse KL divergence, and to models with local latent variables. We also illustrate the existence of good sequence with a number of examples. Our results complement a growing body of work focused on the frequentist properties of variational Bayesian methods.

5.1 Introduction

Recall that the idea behind VB is to approximate the intractable posterior $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ with an element $q(\theta)$ of some simpler class of distributions $\mathcal{Q}$. 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 θ. The variational solution q is the element of $\mathcal{Q}$ that is closest to $\boldsymbol{\pi}(\theta|\tilde{X}_n)$, where closeness is measured in terms of the Kullback-Leibler (KL) divergence. Thus, q is the solution to:

$$q(\theta) = \text{argmin}_{\tilde{q}\in\mathcal{Q}}\text{KL}(\tilde{q}(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)). \tag{5.1}$$

We term this as the KL-VB method. From the non-negativity of the Kullback-Leibler divergence, we can view this as maximizing a lower-bound to the logarithm of the model *evidence*, $\log p(\tilde{X}_n) = \log\left(\int p(\tilde{X}_n, \theta)\mathrm{d}\theta\right)$. This lower-bound, called the variational lower-bound or evidence lower bound (ELBO) is defined as

$$\mathrm{ELBO}(\tilde{q}(\theta)) = \log p(\tilde{X}_n) - \mathrm{KL}(\tilde{q}(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)). \tag{5.2}$$

Optimizing the two equations above with respect to q does not involve either calculating expectations with respect to the intractable posterior $\boldsymbol{\pi}(\theta|\tilde{X}_n)$, or evaluating the posterior normalization constant. As a consequence, a number of standard optimization algorithms can be used to select the best approximation $q(\theta)$ to the posterior distribution, examples including expectation-maximization [100] and gradient-based [107] methods. This has allowed the application of Bayesian methods to increasingly large datasets and high-dimensional settings. Despite their widespread popularity in the machine learning, and more recently, the statistics communities, it is only recently that variational Bayesian methods have been studied theoretically [27], [29], [61], [102], [103].

5.1.1 Rényi divergence minimization

Despite its popularity, variational Bayes has a number of well-documented limitations. An important one is its tendency to produce approximations that underestimate the spread of the posterior distribution [52], [53]: in essence, the variational Bayes solution tends to match closely with the dominant mode of the posterior. This arises from the choice of the divergence measure $\mathrm{KL}(q(\theta)\|\boldsymbol{\pi}(\theta|\tilde{X}_n)) = \mathbb{E}_q[\log(q(\theta)/\boldsymbol{\pi}(\theta|\tilde{X}_n))]$, which does not penalize solutions where $q(\theta)$ is small while $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ is large. While many statistical applications only focus on the mode of the distribution, definite calculations of the variance and higher moments are critical in predictive and decision-making problems.

A natural solution is to consider different divergence measures than those used in variational Bayes. Expectation propagation (EP) [54] was developed to minimize $\mathbb{E}_{\boldsymbol{\pi}}[\log(\boldsymbol{\pi}/q)]$ instead, though this requires an expectation with respect to the intractable posterior. Consequently, EP can only minimize an approximation of this objective.

More recently, Rényi's α-divergence [59] has been used as a family of parametrized divergence measures for variational inference [52], [60]. The α-Rényi divergence is defined as

$$D_\alpha\left(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q(\theta)\right) := \frac{1}{\alpha-1}\log\int_\Theta q(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q(\theta)}\right)^\alpha d\theta.$$

The parameter α spans a number of divergence measures and, in particular, we note that as $\alpha \to 1$ we recover the EP objective $\mathrm{KL}(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q(\theta))$, we will call its minimizer $1-$Rényi approximate posterior. Settings of $\alpha > 1$ are particularly interesting since, in contrast to VB which lower-bounds the log-likelihood of the data (equation (5.2)), one obtains tractable upper bounds. Precisely, using Jensen's inequality,

$$p(\tilde{X}_n)^\alpha = \left(\int p(\theta,\tilde{X}_n)\,\frac{q(\theta)}{q(\theta)}\,d\theta\right)^\alpha \le \mathbb{E}_q\left[\left(\frac{p(\theta,\tilde{X}_n)}{q(\theta)}\right)^\alpha\right].$$

Applying the logarithm function on either side,

$$\alpha\log p(\tilde{X}_n) \le \log\mathbb{E}_q\left[\left(\frac{p(\theta,\tilde{X}_n)}{q(\theta)}\right)^\alpha\right] \tag{5.3}$$

$$= \alpha\log p(\tilde{X}_n) + \log\mathbb{E}_q\left[\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q(\theta)}\right)^\alpha\right] := \mathcal{F}_2(q). \tag{5.4}$$

Observe that the second term in the expression for $\mathcal{F}_2(q)$ is just $(\alpha-1)D_\alpha(p(\theta|\tilde{X}_n)\|q(\theta))$. Like with the ELBO lower bound, evaluating this upperbound only involves expectations with respect to $q(\theta)$, and only requires evaluating $p(\theta,\tilde{X}_n)$, the unnormalized posterior distribution. Optimizing this upper bound over some class of distributions $\mathcal{Q}$, we obtain the *α-Rényi* approximation. As noted before, standard variational Bayes, which optimizes a lower-bound, tends to produce approximating distributions that underestimate the posterior variance, resulting in predictions that are overconfident and ignore high-risk regions in the support of the posterior. We illustrate this fact in Figure 5.1 below that reproduces a result from Li and Turner [52]. The true posterior distribution is an anisotropic Gaussian distribution and the variational family consists of isotropic (or mean field) Gaussian distributions. Standard KL-VB, represented by the green curve titled ($\alpha = 0$), clearly fits the mode of the posterior, but completely underestimates the dominant eigen-direction. On the other hand,

for large values of α (the teal shows $\alpha \to +\infty$), the α-Rényi approximate posterior matches the mode and does a better job of capturing the spread of the posterior. The figure also presents results for the $\alpha = 1$ (or EP) and the $\alpha \to -\infty$ cases. As an aside, we observe that our parametrization of the Rényi divergence is different from Li and Turner [52], where the upper-bounds considered in Li and Turner [52] emerge as $\alpha \to -\infty$. We note, furthermore,

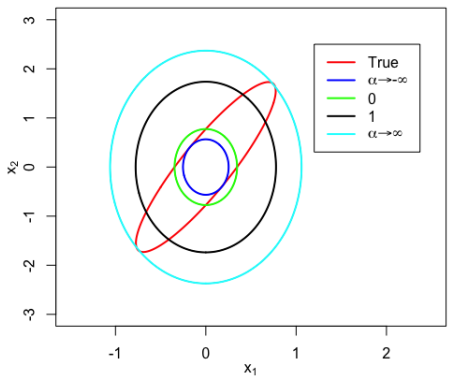


Figure 5.1. Isotropic variational α-Rényi approximations to an anisotropic Gaussian, for different values of α (see also Li and Turner [52])

that in tasks such as model selection, the marginal likelihood of the data is of fundamental interest [125], and the α-Rényi upper bound provides an approximation that complements the VB lower bound. Recent developments in stochastic optimization have allowed the α-Rényi objective to be optimized fairly easily; see Li and Turner [52] and Dieng, Tran, Ranganath, *et al.* [60].

5.1.2 Large sample properties

Despite often state-of-the-art empirical results, variational methods still present a number of unanswered theoretical questions. This is particularly true for α-Rényi divergence minimization which has empirically demonstrated very promising results for a number of applications [52], [60]. In recent work, Zhang and Gao [61] have shown conditions under which α-Rényi variational methods are consistent when α is less than one. Their results followed from a proof for the regular Kullback-Leibler variational algorithm, and thus only apply to situations when a *lower-bound* is optimized. As we mentioned before, the setting with α

greater than 1 is qualitatively different from both Kullback-Leibler and Rényi divergence with $\alpha < 1$. This setting, which is also of considerable practical interest, is the focus of this chapter and we address the question of asymptotic consistency of the approximate posterior distribution obtained by minimizing the Rényi divergence.

Asymptotic consistency [109] is a basic frequentist requirement of any statistical method, guaranteeing that the 'true' parameter is recovered as the number of observations tends to infinity. Table 5.1 summarizes the current known results on consistency of VI and EP, and highlights the gap that this chapter is intended to fill. We also want clarify that in this chapter, we are not analyzing the actual EP algorithm [126], but our analysis is a step towards understanding the global minimizer of the EP objective.

Table 5.1. Known results on the asymptotic consistency of variational methods.

Methods	Existing works
KL-VB	Wang and Blei [27],Zhang and Gao [61]
α-Rényi ($\alpha < 1$)	Zhang and Gao [61]
α-Rényi ($\alpha > 1$)	This chapter
1-Rényi ($\alpha \to 1$, global EP)	This chapter

As we will see, filling these gaps will require new developments. This follows from two complicating factors: 1) Rényi divergence with $\alpha > 1$ *upper-bounds* the log-likelihood, and 2) this requires new analytical approaches involving expectations with respect to the intractable $\boldsymbol{\pi}(\theta|\tilde{X}_n)$. We thus emphasize that the results in this chapter are not a consequence of recent analysis in Wang and Blei [27] and Zhang and Gao [61] for the KL-VB, and our proofs differ substantially from these results.

We establish our main result in Theorem 5.3.1 under mild regularity conditions. First, in Assumption 5.2.1 we assume that the prior distribution places positive mass in the neighborhood of the true parameter θ_0 and that it is uniformly bounded. The former condition is a reasonable assumption to make - clearly, if the prior does not place any mass in the neighborhood of the true parameter (assuming one exists) then neither will the posterior. The uniform boundedness condition on the other hand is attendant to a loss of generality. In particular, we cannot assume certain heavy-tailed priors (such as Pareto) which might be important for some engineering applications. Second, we also make the mild assumption that

the likelihood function is locally asymptotically normal (LAN) in Assumption 5.2.2. This is a standard assumption that holds for a variety of statistical/stochastic models. However, while the LAN assumption will be critical for establishing the asymptotic consistency results, it is unclear if it is necessary as well. We observe that Wang and Blei [27] make a similar assumption in analyzing the consistency of KL-VB. We note that any model P_θ that is twice differentiable in the parameter θ satisfies the LAN condition [109]. The properties of the variational family are critical to the consistency result. Assumption 5.2.3 is a mild condition that insists on there existing Dirac delta distributions in an open neighborhood of the true parameter θ_0. While it may appear that this condition is hard to verify, if the variational family consists of Gaussian distributions, for instance, then Dirac delta distributions are present at all points in the parameter space. Consequently, we assert that Assumption 5.2.3 is easy to satisfy in practice. Next, we assume that the variational family contains 'good sequences', that are constructed so as to converge at the same rate as the true posterior (in sequence with the sample size) and the first moment of an element in the sequence is precisely the maximum likelihood estimator of the parameter (at a given sample size). We also require the tails of the good sequence to bound the tails of the true posterior. We provide examples that verify the existence of good sequences in commonly used variational families, such as the mean-field family.

The proof of Theorem 5.3.1 is a consequence of a series of auxiliary results. First, in Lemma 5.3.1 we characterize α-Rényi minimizers and show that the sequence must have a Dirac delta distribution at the true parameter θ_0 in the large sample limit. Then, in Lemma 5.3.2 we argue that any convex combination of a Dirac delta distribution at the true parameter θ_0 with any other distribution can not achieve zero α-Rényi divergence in the limit. Next, we show in Proposition 5.3.1 that the α-Rényi divergence between the true posterior and the closest variational approximator is bounded above in the large sample limit. We demonstrate this by showing that a 'good sequence' of distributions (see Assumption 5.2.4) has asymptotically bounded α-Rényi divergence, implying that the minimizers do as well. Note that this does not yet prove that the minimizing sequence converges to a Dirac delta distribution at θ_0.

The next stage of the analysis is concerned with demonstrating that the minimizing sequence does indeed converge to a Dirac delta distribution concentrated at the true parameter. We demonstrate this fact as a consequence of Proposition 5.3.1, Lemma 5.3.1, and Lemma 5.3.2. In essence, Theorem 5.3.1 shows that, α-Rényi minimizing distributions are arbitrarily close to a good sequence, in the sense of Rényi divergence with the posterior in the large sample limit.

In our next result in Theorem 5.3.2, under additional regularity conditions, we further characterize the rate of convergence of the $\alpha-$Rényi minimizers. We demonstrate that the $\alpha-$Rényi minimizing sequence cannot concentrate to a point in the parameter space at a faster rate than the true posterior concentrates at the true parameter θ_0. Consequently, the tail mass in the α-Rényi minimizer could dominate that of the true posterior. This is in contrast with KL-VB, where the evidence lower bound (ELBO) maximizer typically under-estimates the variance of the true posterior.

Here is a brief roadmap of the chapter. In Section 5.2, we formally introduce the α-Rényi methodology, and rigorously state the necessary regularity assumptions. We present our main result in Section 5.3, presenting only the proofs of the primary results. In Section 5.4 we also recover the consistency of $1-$Rényi, approximate posteriors, the global minimizer of EP objective as a consequence of the results in Section 5.3. In Section 5.5, we generalize the notion of good sequence to the models with local latent parameters and under some additional regularity conditions we prove asymptotic consistency of the α-Rényi approximate posterior over global latent parameters. All proofs of auxiliary and technical results are delayed to the Appendix.

5.2 Variational Approximation using $\alpha-$Rényi Divergence

We assume that the data-generating distribution is parametrized by $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \geq 1$ and is absolutely continuous with respect to the Lebesgue measure, so that the likelihood function $p(\cdot|\theta)$ is well-defined. We place a prior $\boldsymbol{\pi}(\theta)$ on the unknown θ, and denote $\boldsymbol{\pi}(\theta|\tilde{X}_n) \propto p(\theta, \tilde{X}_n)$ as the posterior distribution, where $\tilde{X}_n = \{\xi_1, \ldots, \xi_n\}$ are the n

independent and identically distributed (i.i.d.) observed samples generated from the 'true' measure $P_{\theta_0}(\equiv P_0)$ in the likelihood family.

In this chapter we will study the $\alpha-$Rényi-approximate posterior q_n^* that minimizes the $\alpha-$Rényi divergence between $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ and $\tilde{q}(\cdot)$ in some set $\mathcal{Q}$ $\forall \alpha > 1$; that is,

$$q_r^*(\theta|\tilde{X}_n) := \text{argmin}_{\tilde{q}\in\mathcal{Q}} \left\{ D_\alpha \left(\boldsymbol{\pi}(\theta|\tilde{X}_n) \| \tilde{q}(\theta) \right) := \frac{1}{\alpha - 1} \log \int_\Theta \tilde{q}(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\tilde{q}(\theta)} \right)^\alpha d\theta \right\}. \quad (5.5)$$

Recall that

Definition 5.2.1 (Dominating distribution)**.** *The distribution Q dominates the distribution P ($P \ll Q$), when P is absolutely continuous with respect to Q; that is, $supp(P) \subseteq supp(Q)$.*

Clearly, the $\alpha-$Rényi divergence in (5.5) is infinite for any distribution $q(\theta) \in \mathcal{Q}$ that does not dominate the true posterior distribution [59]. Intuitively, this is the reason why the α-Rényi approximation can better capture the spread of the posterior distribution.

Our goal is to study the statistical properties of the $\alpha-$Rényi-approximate posterior as defined in (5.5). In particular, we show that under certain regularity conditions on the likelihood, the prior and the variational family the $\alpha-$Rényi-approximate posterior is consistent or converges weakly to a Dirac delta distribution at the true parameter θ_0 as the number of observations $n \to \infty$.

5.2.1 Asymptotic Notations

We first define asymptotic notations that frequently appear in our proofs and assumptions. We write $a_n \sim b_n$ when the sequence $\{a_n\}$ can be approximated by a sequence $\{b_n\}$ for large n, so that the ratio $\frac{a_n}{b_n}$ approaches 1 as $n \to \infty$, $a_n = O(b_n)$ as $n \to \infty$, when there exists a positive number M and $n_0 \geq 1$, such that $a_n \leq M b_n$ $\forall n \geq n_0$, and $a_n \lesssim b_n$ when the sequence $\{a_n\}$ is bounded above by a sequence $\{b_n\}$ for large n.

5.2.2 Assumptions and Definitions

First, we assume the following restrictions on permissible priors.

Assumption 5.2.1 (Prior Density)**.**

(1) The prior density function $\boldsymbol{\pi}(\theta)$ *is continuous with non-zero measure in the neighborhood of the true parameter* θ_0*, and*

(2) there exists a constant $M_p > 0$ *such that* $\boldsymbol{\pi}(\theta) \leq M_p \ \forall \theta \in \Theta$ *and* $\mathbb{E}_{\boldsymbol{\pi}(\theta)}[|\theta|] < \infty$*.*

Assumption 5.2.1(1) is typical in Bayesian consistency analysis - quite obviously, if the prior does not place any mass on the true parameter then the (true) posterior will not either. Indeed, it is well known [110], [111] that for any prior that satisfies Assumption 5.2.1(1), under very mild assumptions,

$$\boldsymbol{\pi}(U|\tilde{X}_n) = \int_U \boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta \Rightarrow 1 \quad P_0 - a.s. \text{ as } n \to \infty, \tag{5.6}$$

where P_0 represents the true data-generating distribution, U is some neighborhood of the true parameter θ_0 and $\Rightarrow$ represents weak convergence of measures. Assumption 5.2.1(2), on the other hand, is a mild technical condition which is satisfied by a large class of prior distributions, for instance, most of the exponential-family distributions. For simplicity, we write $q_n(\theta) \Rightarrow q(\theta)$ to represent weak convergence of the distributions corresponding to the densities $\{q_n\}$ and q.

We define a generic probabilistic order term, $o_{P_\theta}(1)$ with respect to measure P_θ as follows

Definition 5.2.2. *A sequence of random variables* $\{\xi_n\}$ *is of probabilistic order* $o_{P_\theta}(1)$ *when*

$$\lim_{n\to\infty} P_\theta(|\xi_n| > \delta) = 0, \ \textit{for any } \delta > 0 \ .$$

Next, we assume the likelihood function satisfies the following asymptotic normality property (see [109] as well),

Assumption 5.2.2 (Local Asymptotic Normality)**.** *Fix* $\theta_0 \in \Theta$*. The sequence of log-likelihood functions* $\{\log P_n(\theta) = \sum_{\mathrm{i}=1}^n \log p(\xi_\mathrm{i}|\theta)\}$ *satisfies a* local asymptotic normality (LAN) *condition, if there exists a sequence of matrices* $\{r_n\}$*, a matrix* $I(\theta_0)$ *and a sequence*

of random vectors $\{\Delta_{n,\theta_0}\}$ *weakly converging to* $\mathcal{N}(0, I(\theta_0)^{-1})$ *as* $n \to \infty$*, such that for every compact set* $K \subset \mathbb{R}^d$

$$\sup_{h \in K} \left| \log P_n(\theta_0 + r_n^{-1} h) - \log P_n(\theta_0) - h^T I(\theta_0) \Delta_{n,\theta_0} + \frac{1}{2} h^T I(\theta_0) h \right| \xrightarrow{P_0^n} 0 \text{ as } n \to \infty \ .$$

The LAN condition is standard, and holds for a wide variety of models. The assumption affords significant flexibility in the analysis by allowing the likelihood to be asymptotically approximated by a scaled Gaussian centered around θ_0 [109]. We observe that [27] makes a similar assumption in their consistency analysis of the variational lower bound. All statistical models P_θ, which are differentiable in quadratic mean with respect to parameter θ, satisfy the LAN condition with $r_n = \sqrt{n}I$, where I is an identity matrix [109, Chapter-7]. Also, all models P_θ which are twice continuously differentiable in θ are also differentiable in quadratic mean and thus satisfy LAN condition, for instance most of the exponential family model satisfy LAN condition.

Now, let δ_θ represent the Dirac delta distribution function, or singularity, concentrated at the parameter θ.

Definition 5.2.3 (Degenerate distribution)**.** *A sequence of distributions* $\{q_n(\theta)\}$ *converges weakly to* δ_θ *that is,* $q_n(\theta) \Rightarrow \delta_\theta$ *for some* $\theta \in \Theta$*, if and only if* $\forall \eta > 0$

$$\lim_{n \to \infty} \int_{\{|\theta - \theta| > \eta\}} q_n(\theta) d\theta = 0.$$

We use the term 'non-degenerate' for a sequence of distributions that does not converge in distribution to a Dirac delta distribution. We also use the term 'non-singular' to refer to a distribution that does not contain any singular components (i.e., it is absolutely continuous with respect to the Lebesgue measure). And, conversely, if a distribution contains both singularities and absolutely continuous components we term it a 'singular distribution'. More formally,

Definition 5.2.4 (Singular distributions). *Let $d(\theta)$ be a distribution with support Θ and for any $\mathrm{i} \in \{1, \ldots, K\}$ and $K < \infty$ denote $\delta_{\theta_\mathrm{i}}$, as the Dirac delta distributions at θ_i for any $\theta_\mathrm{i} \in \Theta$, then we define singular distribution $q(\theta)$;*

$$q(\theta) := wd(\theta) + \sum_{\mathrm{i}=1}^{K} w^{\mathrm{i}} \delta_{\theta_\mathrm{i}},$$

where $w, \{w^{\mathrm{i}}\}_{\mathrm{i}=1}^{K} \in [0, 1)$ and $w + \sum_{\mathrm{i}=1}^{K} w^{\mathrm{i}} = 1$ with at least one of the weights $\{w^{\mathrm{i}}\}_{\mathrm{i}=1}^{K}$ strictly positive.

Finally, we come to the conditions on the variational family $\mathcal{Q}$. We first assume that

Assumption 5.2.3 (Variational Family). *The variational family $\mathcal{Q}$ must contain all Dirac delta distributions in some open neighborhood of $\theta_0 \in \Theta$.*

Since we know that the posterior converges weakly to a Dirac delta distribution function, this assumption is a necessary condition to ensure that the variational approximator exists in the limit. Next, we define the rate of convergence of a sequence of distributions to a Dirac delta distribution as follows.

Definition 5.2.5 (Rate of convergence). *A sequence of distributions $\{q_n(\theta)\}$ converges weakly to δ_{θ_1}, $\forall \theta_1 \in \Theta$ at the rate of γ_n if*

(1) the sequence of means $\{\check{\theta}_n := \int \theta q_n(\theta) d\theta\}$ converges to θ_1 as $n \to \infty$, and

(2) the variance of $\{q_n(\theta)\}$ satisfies

$$E_{q_n(\theta)}[|\theta - \check{\theta}_n|^2] = O\left(\frac{1}{\gamma_n^2}\right).$$

A crucial assumption, on which rests the proof of our main result, is the existence of what we call a 'good sequence' in $\mathcal{Q}$.

Assumption 5.2.4 (Good sequence). *For any $\bar{M} > 0$, the variational family $\mathcal{Q}$ contains a sequence of distributions $\{\bar{q}_n(\theta)\}$ with the following properties:*

(1) there exists $n_1 \geq 1$ *such that* $\int_\Theta \theta \bar{q}_n(\theta) d\theta = \hat{\theta}_n$*, where* $\hat{\theta}_n$ *is the maximum likelihood estimate, for each* $n \geq n_1$*,*

(2) there exists $n_{\bar{M}} \geq 1$ *such that the rate of convergence is* $\gamma_n = \sqrt{n}$ *, that is* $E_{\bar{q}_n(\theta)}[|\theta - \hat{\theta}_n|^2] \leq \frac{\bar{M}}{\gamma_n^2}$ *for each* $n \geq n_{\bar{M}}$*,*

(3) there exist a compact ball $K \subset \Theta$ *containing the true parameter* θ_0 *and* $n_2 \geq 1$*, such that the sequence of Radon-Nikodym derivatives of the Bayes posterior density with respect to the sequence* $\{\bar{q}_n\}$ *exists and is bounded above by a finite positive constant* M_r *outside of* K *for all* $n \geq n_2$ *; that is,*

$$\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\bar{q}_n(\theta)} \leq M_r, \ \forall \theta \in \Theta \backslash K \ \text{and} \ \forall n \geq n_2, \quad P_0 - a.s.$$

(4) there exists $n_3 \geq 1$ *such that the good sequence* $\{\bar{q}_n(\theta)\}$ *is log-concave in* θ *for all* $n \geq n_3$*.*

We term such a sequence of distributions as 'good sequences'.

The first two parts of the assumption hold so long as the variational family $\mathcal{Q}$ contains an open neighborhood of distributions around δ_{θ_0}. The third part essentially requires that for $n \geq n_2$, the tails of $\{\bar{q}_n(\theta)\}$ must decay no faster than the tails of the posterior distribution. Since, the good sequence converges weakly to δ_{θ_0}, this assumption is a mild technical condition. The last assumption implies that the good sequence is, for large sample sizes, a maximum entropy distribution under some deviation constraints on the entropy maximization problem [127]. Note that this does not imply that the good sequence is necessarily Gaussian (which is the maximum entropy distribution specifically under standard deviation constraints).

We note that this assumption is on the family $\mathcal{Q}$, and not on the minimizer of the Rényi divergence. We demonstrate the existence of good sequences for some example models.

Example 5.2.1. *Consider a model whose likelihood is an* m*-dimensional multivariate Gaussian likelihood with unknown mean vector* $\boldsymbol{\mu}$ *and known covariance matrix* $\boldsymbol{\Sigma}$*. Using an* m*-*

dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu_0}$ *and covariance matrix* $\boldsymbol{\Sigma}$ *as conjugate prior, the posterior distribution is*

$$\boldsymbol{\pi}(\boldsymbol{\mu}|\tilde{X}_n) = \sqrt{\frac{(n+1)^m}{(2\boldsymbol{\pi})^m det\,(\boldsymbol{\Sigma})}} \mathrm{e}^{-\frac{n+1}{2}\left(\boldsymbol{\mu}-\frac{\sum_{\mathrm{i}=1}^{n}\xi_{\mathrm{i}}+\boldsymbol{\mu_0}}{n+1}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\frac{\sum_{\mathrm{i}=1}^{n}\xi_{\mathrm{i}}+\boldsymbol{\mu_0}}{n+1}\right)},$$

where exponents 'T*' and '*-1*' denote transpose and inverse. Next, consider the mean-field variational family, that is the product of* m *1-dimensional normal distributions. Consider a sequence in the variational family with mean* $\{\mu_{q_n}^{\mathrm{j}}, \mathrm{j} \in \{1,2,\ldots,m\}\}$ *and variance* $\left\{\frac{\sigma_{\mathrm{j}}^2}{\gamma_n^2}, \mathrm{j} \in \{1,2,\ldots,m\}\right\}$*:*

$$q_n(\boldsymbol{\mu}) = \prod_{\mathrm{j}=1}^{m} \sqrt{\frac{\gamma_n^2}{2\boldsymbol{\pi}\sigma_{\mathrm{j}}^2}} \mathrm{e}^{-\frac{\gamma_n^2}{2\sigma_{\mathrm{j}}^2}\left(\mu_{\mathrm{j}}-\mu_{q_n}^{\mathrm{j}}\right)^2} = \sqrt{\frac{\gamma_n^{2m}}{(2\boldsymbol{\pi})^m det(\mathbf{I}_\sigma)}} \mathrm{e}^{-\frac{\gamma_n^2}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_{q_n})^T \mathbf{I}_\sigma^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_{q_n})},$$

where $\boldsymbol{\mu}_{q_n} = \{\mu_{q_n}^1, \mu_{q_n}^2, \ldots, \mu_{q_n}^m\}$ *and* $\mathbf{I}_\sigma$ *is an* $m \times m$ *diagonal matrix with diagonal elements* $\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2\}$*. Notice that* γ_n *is the rate at which the sequence* $\{q_n(\boldsymbol{\mu})\}$ *converges weakly. It is straightforward to observe that the variational family contains sequences that satisfy properties (1) and (2) in Assumption 5.2.4, that is*

$$\gamma_n = \sqrt{n} \text{ and } \boldsymbol{\mu_{q_n}} = \frac{\sum_{\mathrm{i}=1}^{n}\xi_{\mathrm{i}} + \boldsymbol{\mu_0}}{n+1}.$$

For brevity, denote $\tilde{\boldsymbol{\mu}}_n := \boldsymbol{\mu} - \boldsymbol{\mu}_{q_n} = \boldsymbol{\mu} - \frac{\sum_{\mathrm{i}=1}^{n}\xi_{\mathrm{i}}+\boldsymbol{\mu_0}}{n+1}$ *. To verify property (3) in Assumption 5.2.4 consider the ratio,*

$$\frac{\boldsymbol{\pi}(\boldsymbol{\mu}|\tilde{X}_n)}{q_n(\boldsymbol{\mu})} = \frac{\sqrt{\frac{(n+1)^m}{(2\boldsymbol{\pi})^m det(\boldsymbol{\Sigma})}} \mathrm{e}^{-\frac{n+1}{2}\tilde{\boldsymbol{\mu}}_n^T \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}_n}}{\sqrt{\frac{\gamma_n^{2m}}{(2\boldsymbol{\pi})^m det(\mathbf{I}_\sigma)}} \mathrm{e}^{-\frac{\gamma_n^2}{2}\tilde{\boldsymbol{\mu}}_n^T \mathbf{I}_\sigma^{-1}\tilde{\boldsymbol{\mu}}_n}}.$$

Using the fact that $\gamma_n^2 = n < n+1$*, the ratio above can be bounded above by*

$$\frac{\boldsymbol{\pi}(\boldsymbol{\mu}|\tilde{X}_n)}{q_n(\boldsymbol{\mu})} \leq \sqrt{\frac{2^m det(\mathbf{I}_\sigma)}{det\,(\boldsymbol{\Sigma})}} \frac{\mathrm{e}^{-\frac{n+1}{2}\tilde{\boldsymbol{\mu}}_n^T \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\mu}}_n}}{\mathrm{e}^{-\frac{n+1}{2}\tilde{\boldsymbol{\mu}}_n^T \mathbf{I}_\sigma^{-1}\tilde{\boldsymbol{\mu}}_n}} = \sqrt{\frac{2^m det(\mathbf{I}_\sigma)}{det\,(\boldsymbol{\Sigma})}} \mathrm{e}^{-\frac{n+1}{2}\tilde{\boldsymbol{\mu}}_n^T\left(\boldsymbol{\Sigma}^{-1}-\mathbf{I}_\sigma^{-1}\right)\tilde{\boldsymbol{\mu}}_n}.$$

Observe that if the matrix $(\mathbf{\Sigma}^{-1} - \mathbf{I}_{\sigma}^{-1})$ *is positive definite then the ratio above is bounded by* $\sqrt{\frac{2^m det(\mathbf{I}_\sigma)}{det(\mathbf{\Sigma})}}$ *and if* $\mathcal{Q}$ *is large enough it will contain distributions that satisfy this condition. To fix the idea, consider the univariate case, where the positive definiteness implies that the variance of the good sequence is greater than the variance of the posterior for all large enough 'n'. That is, the tails of the good sequence decay slower than the tails of the posterior.*

Example 5.2.2. *Consider a model whose likelihood is a univariate Normal distribution with unknown mean* μ *and known variance* σ*. Using a univariate normal distribution with the mean* μ_0 *and the variance* σ *as prior, the posterior distribution is*

$$\pi(\mu|\tilde{X}_n) = \sqrt{\frac{n+1}{2\pi\sigma^2}} \mathrm{e}^{-\frac{(n+1)}{2\sigma^2}\left(\mu - \frac{\mu_0 + \sum_{i=1}^{n} \xi_i}{n+1}\right)^2}. \tag{5.7}$$

Next, suppose the variational family $\mathcal{Q}$ *is the set of all Laplace distributions. Consider a sequence* $\{q_n(\mu)\}$ *in* $\mathcal{Q}$ *with the location and the scale parameter* k_n *and* b_n *respectively, that is*

$$q_n(\mu) = \frac{1}{2b_n} \mathrm{e}^{-\frac{|\mu - k_n|}{b_n}}.$$

To satisfy properties (1) and (2) in Assumption 5.2.4, we can choose $k_n = \frac{\mu_0 + \sum_{i=1}^{n} \xi_i}{n+1}$ *and* $b_n = \sqrt{\frac{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}{2n}}$, $\forall \alpha > 1$*. For brevity denote* $\tilde{\mu}_n = \mu - \frac{\mu_0 + \sum_{i=1}^{n} \xi_i}{n+1}$*. To verify property (3) in Assumption 5.2.4 consider the ratio,*

$$\frac{\pi(\mu|\tilde{X}_n)}{q_n(\mu)} = \frac{\sqrt{\frac{n+1}{2\pi\sigma^2}} \mathrm{e}^{-\frac{(n+1)}{2\sigma^2}\tilde{\mu}_n^2}}{\frac{1}{2}\sqrt{\frac{2n}{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}} \mathrm{e}^{-\frac{\sqrt{2n}|\tilde{\mu}_n|}{\sqrt{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}}}} \leq \sqrt{\frac{2}{\alpha^{\frac{1}{\alpha-1}}}} \frac{\mathrm{e}^{-\frac{(n+1)}{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}\tilde{\mu}_n^2}}{\mathrm{e}^{-\left|\frac{\sqrt{2(n+1)}|\tilde{\mu}_n|}{\sqrt{\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}}\right|}} \leq \sqrt{\frac{2}{\alpha^{\frac{1}{\alpha-1}}}} \mathrm{e}^{1/2},$$

where the last inequality follows due to the fact that $\mathrm{e}^{-(\frac{x^2}{2} - |x|)} < \mathrm{e}^{1/2}$.

For the same posterior, we can also choose $\mathcal{Q}$ *to be the set of all Logistic distributions. Consider a sequence* $\{q_n(\mu)\}$ *in this variational family with the mean and the scale parameter* m_n *and* s_n *respectively; that is*

$$q_n(\mu) = \frac{1}{s_n} \left(\mathrm{e}^{\frac{\mu - m_n}{2s_n}} + \mathrm{e}^{-\frac{\mu - m_n}{2s_n}}\right)^{-2}.$$

To satisfy properties (1) and (2) in Assumption 5.2.4, we can choose $m_n = \frac{\mu_0+\sum_{i=1}^n \xi_i}{n+1}$ *and* $s_n = \sqrt{\frac{2\pi\alpha^{\frac{1}{\alpha-1}}\sigma^2}{n+1}}$, $\forall \alpha > 1$. *For brevity denote* $\tilde{\mu}_n = \mu - \frac{\mu_0+\sum_{i=1}^n \xi_i}{n+1}$. *To verify property (3) in Assumption 5.2.4 observe that,*

$$\frac{\pi(\lambda|\tilde{X}_n)}{q_n(\lambda)} = \frac{\sqrt{\frac{n+1}{2\pi\sigma^2}}\mathrm{e}^{-\frac{(n+1)}{2\sigma^2}\left(\mu-\frac{\mu_0+\sum_{i=1}^n \xi_i}{n+1}\right)^2}}{\frac{1}{s_n}\left(\mathrm{e}^{\frac{\mu-m_n}{2s_n}}+\mathrm{e}^{-\frac{\mu-m_n}{2s_n}}\right)^{-2}} = \frac{1}{\sqrt{\alpha^{\frac{1}{\alpha-1}}}}\mathrm{e}^{-\left(\frac{\tilde{\mu}_n}{s_n}\right)^2}\left(\mathrm{e}^{\left(\frac{\tilde{\mu}_n}{2s_n}\right)}+\mathrm{e}^{-\left(\frac{\tilde{\mu}_n}{2s_n}\right)}\right) \leq \frac{1}{\sqrt{\alpha^{\frac{1}{\alpha-1}}}}2\mathrm{e}^{1/16},$$

where the last inequality follows due to the fact that $\mathrm{e}^{-x^2}\left(\mathrm{e}^{x/2}+\mathrm{e}^{-x/2}\right) < 2\mathrm{e}^{1/16}$.

Example 5.2.3. *Finally, consider a univariate exponential likelihood model with the unknown rate parameter* λ. *For some prior distribution* $\pi(\lambda)$, *the posterior distribution is*

$$\pi(\lambda|\tilde{X}_n) = \frac{\pi(\lambda)\lambda^n\mathrm{e}^{-\lambda\sum_{i=1}^n \xi_i}}{\int \pi(\lambda)\lambda^n\mathrm{e}^{-\lambda\sum_{i=1}^n \xi_i}d\lambda}.$$

Choose $\mathcal{Q}$ *to be the set of Gamma distributions. Consider a sequence* $\{q_n(\mu)\}$ *in the variational family with the shape and the rate parameter* k_n *and* β_n *respectively, that is*

$$q_n(\lambda) = \frac{\beta_n^{k_n}}{\Gamma(k_n)}\lambda^{k_n-1}\mathrm{e}^{-\lambda\beta_n},$$

where $\Gamma(\cdot)$ *is the* $\Gamma-$ *function. To satisfy properties (1) and (2) in Assumption 5.2.4, we can choose* $k_n = n+1$ *and* $\beta_n = \sum_{i=1}^n \xi_i$. *To verify property (3) in Assumption 5.2.4 consider the ratio,*

$$\frac{\pi(\lambda|\tilde{X}_n)}{q_n(\lambda)} = \frac{\pi(\lambda)\lambda^n\mathrm{e}^{-\lambda\sum_{i=1}^n \xi_i}}{\frac{\beta_n^{k_n}}{\Gamma(k_n)}\lambda^{k_n-1}\mathrm{e}^{-\lambda\beta_n}\int \pi(\lambda)\lambda^n\mathrm{e}^{-\lambda\sum_{i=1}^n \xi_i}d\lambda} = \frac{\pi(\lambda)\Gamma(n+1)}{\left(\sum_{i=1}^n \xi_i\right)^{n+1}\int \pi(\lambda)\lambda^n\mathrm{e}^{-\lambda\sum_{i=1}^n \xi_i}d\lambda}.$$

Now, observe that $\frac{\left(\sum_{i=1}^n \xi_i\right)^{n+1}}{\Gamma(n+1)}\lambda^n\mathrm{e}^{-\lambda\sum_{i=1}^n \xi_i}$ *is the density of Gamma distribution with the mean* $\frac{n+1}{\sum_{i=1}^n \xi_i}$ *and the variance* $\frac{1}{n+1}\left(\frac{n+1}{\sum_{i=1}^n \xi_i}\right)^2$. *Since, we assumed in Assumption 5.2.1(2) that*

$\boldsymbol{\pi}(\lambda)$ *is bounded from above by* M_p*, therefore for large* n*,* $\frac{\left(\sum_{i=1}^{n}\xi_i\right)^{n+1}}{\Gamma(n+1)}\int \boldsymbol{\pi}(\lambda)\lambda^n \mathrm{e}^{-\lambda\sum_{i=1}^{n}\xi_i}d\lambda \sim \boldsymbol{\pi}\left(\frac{n+1}{\sum_{i=1}^{n}\xi_i}\right)$*. Hence, it follows that for large enough* n

$$\frac{\boldsymbol{\pi}(\lambda|\tilde{X}_n)}{q_n(\lambda)} \leq \frac{M_p}{\boldsymbol{\pi}(\lambda_0)},$$

where $\frac{\sum_{i=1}^{n}\xi_i}{n+1} \to \frac{1}{\lambda_0}$ *as* $n \to \infty$*.*

5.3 Consistency of $\alpha-$Rényi Approximate Posterior

Recall that the $\alpha-$Rényi-approximate posterior q_n^* is defined as

$$q_r^*(\theta|\tilde{X}_n) := \text{argmin}_{\tilde{q}\in\mathcal{Q}}\left\{D_\alpha\left(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{q}(\theta)\right) := \frac{1}{\alpha-1}\log\int_\Theta \tilde{q}(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\tilde{q}(\theta)}\right)^\alpha d\theta\right\}. \quad (5.8)$$

We now show that under the assumptions in the previous section, the $\alpha-$Rényi approximators are asymptotically consistent as the sample size increases in the sense that $q_n^* \Rightarrow \delta_{\theta_0}$ $P_0 - a.s.$ as $n \to \infty$. To illustrate the ideas clearly, we present our analysis assuming a univariate parameter space, and that the model P_θ is twice differentiable in parameter θ, and therefore satisfies the LAN condition with $r_n = \sqrt{n}$ [109]. The LAN condition together with the existence of a sequence of test functions [109, Theorem 10.1] also implies that the posterior distribution converges weakly to δ_{θ_0} at the rate of $\sqrt{n}$. The analysis can be easily adapted to multivariate parameter spaces.

We will first establish some structural properties of the minimizing sequence of distributions. We show that for any sequence of distributions converging weakly to a non-singular distribution the $\alpha-$Rényi divergence is unbounded in the limit.

Lemma 5.3.1. *Under Assumptions 5.2.1, 5.2.2, 5.2.3, and 5.2.4, the* $\alpha-$*Rényi divergence between the true posterior and the sequence of distribution* $\{q_n(\theta)\} \subset \mathcal{Q}$ *can only be finite in the limit if* $q_n(\theta)$ *converges weakly to a singular distribution* $q(\theta)$*, with a Dirac delta distribution at the true parameter* θ_0*.*

The result above implies that the $\alpha-$Rényi approximate posterior must have a Dirac delta distribution component at θ_0 in the limit; that is, it should converge in distribution to δ_{θ_0} or a convex combination of δ_{θ_0} with singular or non-singular distributions as $n \to \infty$. Next, we consider a sequence $\{q_n(\theta)\} \subset \mathcal{Q}$ that converges weakly to a convex combination of δ_{θ_0} and singular or non-singular distributions $q_{\mathrm{i}}(\theta)$, $\mathrm{i} \in \{1, 2, \ldots\}$ such that for weights $\{w^{\mathrm{i}} \in (0,1) : \sum_{\mathrm{i}=1}^{\infty} w^{\mathrm{i}} = 1\}$,

$$q_n(\theta) \Rightarrow w^{\mathrm{j}} \delta_{\theta_0} + \sum_{\mathrm{i}=1, \mathrm{i} \neq \mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta). \tag{5.9}$$

In the following result, we show that the $\alpha-$Rényi divergence between the true posterior and the sequence $\{q_n(\theta)\}$ is bounded below by a positive number.

Lemma 5.3.2. *Under Assumption 5.2.1, the $\alpha-$Rényi divergence between the true posterior and sequence $\{q_n(\theta) \in \mathcal{Q}\}$ is bounded away from zero; that is*

$$\liminf_{n \to \infty} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n) \| q_n(\theta)) \geq \eta > 0 \quad P_0 - a.s.$$

We also show in Lemma 5.6.5 in the appendix that if in (5.9) the components $\{q_{\mathrm{i}}(\theta)\ \mathrm{i} \in \{1, 2, \ldots\}\}$ are singular then

$$\liminf_{n \to \infty} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n) \| q_n(\theta)) \geq 2(1 - w^{\mathrm{j}})^2 > 0 \quad P_0 - a.s,$$

where w^{j} is the weight of δ_{θ_0}.

A consistent sequence asymptotically achieves zero $\alpha-$Rényi divergence. To show its existence, we first provide an asymptotic upper-bound on the minimal $\alpha-$Rényi divergence in the next proposition. This, coupled with the previous two structural results, will allow us to prove the consistency of the minimizing sequence.

Proposition 5.3.1. *For a given $\alpha > 1$ and under Assumptions 5.2.1, 5.2.2, 5.2.3, and 5.2.4, and for any good sequence $\bar{q}_n(\theta)$ there exist $n_0 \geq 1$ and $\bar{M} > 0$ such thatfor all $n \geq n_0$, the minimal $\alpha-$Rényi divergence satisfies*

$$\min_{q\in\mathcal{Q}} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q(\theta)) \leq D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta)) \leq B = \frac{1}{2}\log\left(\frac{\bar{\mathrm{e}}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}}\right) + o_{P_0^n}(1), \tag{5.10}$$

where $I(\theta_0)$ is defined in Assumption 5.2.2 and $\bar{\mathrm{e}}$ is the Euler's constant.

Now Proposition 5.3.1, Lemma 5.3.1, and Lemma 5.3.2 allow us to prove our main result that the $\alpha-$Rényi approximate posterior converges weakly to δ_{θ_0}.

Theorem 5.3.1. *Under Assumptions 5.2.1, 5.2.2, 5.2.3, and 5.2.4, the $\alpha-$Rényi approximate posterior $q_r^*(\theta|\tilde{X}_n)$ converges weakly to a Dirac delta distribution at the true parameter θ_0; that is,*

$$q_n^* \Rightarrow \delta_{\theta_0} \quad \textit{in-}P_0^n \textit{ probability as } n \to \infty.$$

Proof. First, we argue that there always exists a sequence $\{\tilde{q}_n(\theta)\} \subset \mathcal{Q}$ such that for every $\eta > 0$

$$\lim_{n\to\infty} P_0^n\left(D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{q}_n(\theta)) \leq \eta\right) = 1.$$

We demonstrate the existence of $\tilde{q}_n(\theta)$ by construction. Recall from Proposition 5.3.1(2) that there exist $0 < \bar{M} < \infty$ and $n_0 \geq 1$, such that for all $n \geq n_0$

$$D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta)) \leq \frac{1}{2}\log\frac{\bar{\mathrm{e}}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}} + o_{P_0^n}(1),$$

where $\bar{q}_n(\theta)$ is the good sequence as defined in Assumption 5.2.4 and $\bar{\mathrm{e}}$ is the Euler's constant. Now using the definition of $o_{P_0^n}(1)$, for every $\eta > 0$, it follows from the inequality above that

$$\lim_{n\to\infty} P_0^n\left(D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta)) - \frac{1}{2}\log\frac{\bar{\mathrm{e}}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}} > \eta\right) \leq \lim_{n\to\infty} P_0^n\left(o_{P_0^n}(1) > \eta\right) = 0. \tag{5.11}$$

Now a specific good sequence can be chosen by fixing $\bar{M} = \tilde{M} := \frac{\alpha^{\frac{1}{\alpha-1}}}{\bar{\mathrm{e}}I(\theta_0)}$, implying that

$$\lim_{n\to\infty} P_0^n\left(D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{q}_n(\theta)) > \eta\right) = 0. \tag{5.12}$$

The above result implies that there exist a sequence in family $\mathcal{Q}$ such that $D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{q}_n(\theta)) \to 0$ in P_0^n -probability.

Next, we will show that the minimizing sequence must converge to a Dirac delta distribution in probability. The previous result shows that the minimizing sequence must have zero α-Rényi divergence in the limit. Lemma 5.3.1 shows that the minimizing sequence must have a delta at θ_0, since otherwise the α-Rényi divergence is unbounded. Similarly, Lemma 5.3.2 shows that it cannot be a mixture of such a delta with other components, since otherwise the α-Rényi divergence is bounded away from zero.

Therefore, it follows that the $\alpha-$Rényi approximate posterior $q_r^*(\theta|\tilde{X}_n)$ must converge weakly to a Dirac delta distribution at the true parameter θ_0, in P_0^n- probability, thereby completing the proof. □

Note that the choice of $\bar{M}$ in the proof essentially determines the variance of the good sequence. As noted before, the asymptotic log-concavity of the good sequence implies that it is eventually an entropy maximizing sequence of distributions [127]. It does not necessarily follow that the sequence is Gaussian, however. If such a choice can be made (i.e., the variational family contains Gaussian distributions) then the choice of good sequence amounts to matching the entropy of a Gaussian distribution with variance $\frac{\alpha^{\frac{1}{\alpha-1}}}{\bar{e}I(\theta_0)}$.

We further characterize the rate of convergence of the $\alpha-$Rényi approximate posterior under additional regularity conditions. In particular, we establish an upper bound on the rate of convergence of the possible candidate $\alpha-$Rényi approximators when the variational family is sub-Gaussian. Additionally, we require that the posterior distribution satisfies the Bernstein-von Mises Theorem, that is for any compact set K containing θ_0

$$\int_K \boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta = \int_K \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1})d\theta + o_{P_0^n}(1). \tag{5.13}$$

According to Theorem 10.1 in [109], the Bernstein-von Mises Theorem holds under Assumption 5.2.1, 5.2.2, and the following additional assumption on the existence of consistent test functions:

Assumption 5.3.1 (Consistent Tests)**.** *For every $\epsilon > 0$ there exists a sequence of tests $\phi_n(\tilde{X}_n)$ such that* i) $\lim_{n\to\infty} \mathbb{E}_{P_0^n}(\phi_n(\tilde{X}_n)) = 0$*, and* $\lim_{n\to\infty} \sup_{\|\theta-\theta_0\|\geq\epsilon} \mathbb{E}_{P_0^n}(1-\phi_n(\tilde{X}_n)) = 0$.

A further modeling assumption is to choose a sub-Gaussian variational family $\mathcal{Q}$ that limits the variance. We choose a sub-Gaussian sequence of distributions $\{q_n(\theta)\} \subset \mathcal{Q}$, that is for some positive constant B and any $t \in \mathbb{R}$,

$$\mathbb{E}_{q_n(\theta)}[\mathrm{e}^{t\theta}] \leq \mathrm{e}^{\tilde{\theta}_n t + \frac{B}{2\gamma_n^2}t^2}, \tag{5.14}$$

where $\tilde{\theta}_n$ is the mean of $q_n(\theta)$ and γ_n is the rate (see Definition 5.2.5) at which $q_n(\theta)$ converges weakly to a Dirac delta distribution as $n \to \infty$.

Lemma 5.3.3. *Consider a sequence of sub-Gaussian distributions $\{q_n(\theta)\} \subset \mathcal{Q}$, with parameters B and t, that converges weakly to some Dirac delta distribution faster than the posterior converges weakly to δ_{θ_0} (that is, $\gamma_n > \sqrt{n}$), and suppose the true posterior distribution satisfies the Bernstein-von Mises Theorem (5.13). Then, there exists an $n_0 \geq 1$ such that the $\alpha-$Rényi divergence $D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q_n(\theta))$ is infinite for all $n > n_0$.*

We use the above result to show that, when the variational family $\mathcal{Q}$ is sub-Gaussian, then the $\alpha-$Rényi appropriate posterior cannot converge at a rate γ_n faster than $\sqrt{n}$, that is the rate at which the posterior converges weakly to δ_{θ_0}.

Theorem 5.3.2. *Under Assumptions 5.2.1, 5.2.2, 5.2.3, 5.2.4, and 5.3.1, and $\mathcal{Q}$ is a family of sub-Gaussian distribution, then the rate of convergence, γ_n, of $\alpha-$Rényi approximate posterior is bounded above by $\sqrt{n}$, that is $\gamma_n \leq \sqrt{n}$.*

Proof. Since we choose the variational family to be sub-Gaussian, the $\alpha-$Rényi approximate posterior must be one of the sequences satisfying (5.14) and as a consequence of Theorem 5.3.1, $\tilde{\theta}_n$ must converge to θ_0 as $n \to \infty$. On the other hand, using Lemma 5.3.3, it follows that the rate of convergence γ_n of $\alpha-$Rényi approximate posterior must be bounded above by $\sqrt{n}$, that is $\gamma_n \leq \sqrt{n}$. □

5.4 Consistency of $\alpha-$ Rényi Approximate Posterior as α converges to 1

Our results on the consistency of α-Rényi variational approximators in Section 5.3 can be a step forward in understanding the consistency of posterior approximations obtained using expectation propagation (EP) [54], [55]. Observe that for any $n \geq 1$, as $\alpha \to 1$,

$$D_\alpha \left(\boldsymbol{\pi}(\theta|\tilde{X}_n) \| \tilde{q}(\theta) \right) \to \mathrm{KL} \left(\boldsymbol{\pi}(\theta|\tilde{X}_n) \| \tilde{q}(\theta) \right), \tag{5.15}$$

where the limit is the EP objective using KL divergence. We define the 1-Rényi-approximate posterior s_n^* as the distribution in the variational family $\mathcal{Q}$ that minimizes the KL divergence between $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ and $\tilde{s}(\theta)$, where $\tilde{s}(\theta)$ is an element of $\mathcal{Q}$:

$$s_n^*(\theta) := \mathrm{argmin}_{\tilde{s} \in \mathcal{Q}} \left\{ \mathrm{KL} \left(\boldsymbol{\pi}(\theta|\tilde{X}_n) \| \tilde{s}(\theta) \right) := \int_\Theta \boldsymbol{\pi}(\theta|\tilde{X}_n) \log \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\tilde{s}(\theta)} \right) d\theta \right\}. \tag{5.16}$$

We note that the EP algorithm [54] is a message-passing algorithm that optimizes an approximations to this objective [126]. Nevertheless, understanding this idealized objective is an important step towards understanding the actual EP algorithm. Furthermore, ideas from [52] can be used to construct alternate algorithms that directly minimize equation (5.16). We thus focus on this objective, and show that under the assumptions in Section 5.2, the 1-Rényi-approximate posterior is asymptotically consistent as the sample size increases, in the sense that $s_n^* \Rightarrow \delta_{\theta_0}$, $P_0 - a.s.$ as $n \to \infty$. The proofs in this section are corollaries of the results in the previous section.

Recall that the KL divergence lower-bounds the $\alpha-$Rényi divergence when $\alpha > 1$; that is

$$\mathrm{KL} \left(p(\theta) \| q(\theta) \right) \leq D_\alpha \left(p(\theta) \| q(\theta) \right). \tag{5.17}$$

This is a direct consequence of Jensen's inequality. Analogous to Proposition 5.3.1, we first show that the minimal KL divergence between the true Bayesian posterior and the variational family $\mathcal{Q}$ is asymptotically bounded.

Proposition 5.4.1. *For a given $\alpha > 1$, and under Assumptions 5.2.1, 5.2.2, 5.2.3, 5.2.4, and for any good sequence $\bar{q}_n(\theta)$ there exist $n_0 \geq 1$ and $\bar{M} > 0$ such that the minimal* KL *divergence satisfies*

$$\min_{\tilde{s}\in\mathcal{Q}} \mathrm{KL}\left(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{s}(\theta)\right) < B = \frac{1}{2}\log\left(\frac{\bar{\mathrm{e}}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}}\right) + o_{P_0^n}(1). \tag{5.18}$$

where $I(\theta_0)$ is defined in Assumption 5.2.2 and $\bar{\mathrm{e}}$ *is the Euler's constant.*

Proof. The result follows immediately from Proposition 5.3.1 and (5.17), since for any $\tilde{s}(\theta) \in \mathcal{Q}$ and $\alpha > 1$,

$$\mathrm{KL}\left(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{s}(\theta)\right) \leq D_\alpha\left(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{s}(\theta)\right).$$

□

Next, we demonstrate that any sequence of distributions $\{s_n(\theta)\} \subset \mathcal{Q}$ that converges weakly to a distribution $s(\theta) \in \mathcal{Q}$ with positive probability outside the true parameter θ_0 cannot achieve zero KL divergence in the limit. Observe that this result is weaker than Lemma 5.3.1, and does not show that the KL divergence is necessarily infinite in the limit. This loses some structural insight.

Lemma 5.4.1. *There exists an $\eta > 0$ in the extended real line such that the* KL *divergence between the true posterior and sequence $\{s_n(\theta)\}$ is bounded away from zero; that is,*

$$\liminf_{n\to\infty} \mathrm{KL}(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|s_n(\theta)) \geq \eta > 0 \quad P_0 - a.s.$$

Now using Proposition 5.4.1 and Lemma 5.4.1 we show that the 1-Rényi-approximate posterior converges weakly to the δ_{θ_0}.

Theorem 5.4.1. *Under Assumptions 5.2.1, 5.2.2, 5.2.3, and 5.2.4, the 1-Rényi-approximate posterior $s_n^*(\theta)$ satisfies*

$$s_n^* \Rightarrow \delta_{\theta_0} \quad \textit{in-}P_0^n \textit{ probability as } n \to \infty.$$

Proof. Recall (5.12) from the proof of Theorem 5.3.1 that there exists a good sequence $\tilde{q}_n(\theta)$, such that

$$D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{q}_n(\theta)) \to 0 \text{ in-}P_0^n \text{ probability as } n \to \infty.$$

Since the KL divergence is always non-negative, using (5.17) it follows that

$$\mathrm{KL}(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\tilde{q}_n(\theta)) \to 0 \text{ in-}P_0^n \text{ probability as } n \to \infty.$$

Consequently, the sequence of 1-Rényi-approximate posteriors must also achieve zero KL divergence from the true posterior in the large sample limit with high probability. Finally, as demonstrated in Lemma 5.4.1, any other sequence of distribution that converges weakly to a distribution, that has positive probability at any point other that θ_0 cannot achieve zero KL divergence. Therefore, it follows that the 1-Rényi-approximate posterior $s_n^*(\theta)$ must converge weakly to a Dirac delta distribution at the true parameter θ_0, in-P_0^n probability as $n \to \infty$, thereby completing the proof. □

5.5 Models with Local Latent Parameters

We generalize the model we have worked with so far to include a collection of n independent local latent variables $z_{1:n} := \{z_1, z_2, \ldots, z_n\} \in \mathcal{Z}^n$, one for each observation ξ_i. We assume these are distributed as $\boldsymbol{\pi}(z_\mathrm{i}|\theta)$ for each i, with the observations distributed as $p(\xi_\mathrm{i}|z_\mathrm{i}, \theta)$. Recall that θ is the global latent variable with prior distribution $\boldsymbol{\pi}(\theta)$. Denote by z_0 and θ_0 the true local and global latent parameters respectively. In this section, P_0 denotes the true model P_{θ_0,z_0}. The posterior distribution over θ and $z_{1:n}$ is defined as

$$\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n) := \frac{\boldsymbol{\pi}(\theta)\prod_{\mathrm{i}=1}^n \boldsymbol{\pi}(z_\mathrm{i}|\theta)p(\xi_\mathrm{i}|z_\mathrm{i},\theta)}{\int\int \boldsymbol{\pi}(\theta)\prod_{\mathrm{i}=1}^n \boldsymbol{\pi}(z_\mathrm{i}|\theta)p(\xi_\mathrm{i}|z_\mathrm{i},\theta)d\theta dz_{1:n}}.$$

We denote the denominator above as $P(\tilde{X}_n)$, the model *evidence*, and the numerator as $p(\theta, \tilde{X}_n, z_{1:n})$. Since computing $P(\tilde{X}_n)$ is difficult, an approximate posterior can be obtained by minimizing the following objective over a carefully chosen variational family $\mathcal{Q}$:

$$D_\alpha\left(\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)\|q(\theta, z_{1:n})\right) := \frac{1}{\alpha-1}\log\int_{\Theta\times\mathcal{Z}^n} q(\theta, z_{1:n})\left(\frac{\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)}{q(\theta, z_{1:n})}\right)^\alpha d\theta dz_{1:n}, \text{ where } \alpha > 1.$$

This objective can be derived as an upper-bound to the model evidence similar to equation (5.4). It is common to assume that the variational family $\mathcal{Q}$ can be factorized into subsets $\mathcal{Q}^n$ (local) and $\bar{\mathcal{Q}}$ (global), and define the Rényi approximate posterior over the global latent parameter θ as

$$q_r^*(\theta|\tilde{X}_n) := \text{argmin}_{q(\theta)\in\bar{\mathcal{Q}}} \min_{q(z_{1:n})\in\mathcal{Q}^n} \log\int_{\Theta\times\mathcal{Z}^n} q(\theta)q(z_{1:n})\left(\frac{p(\theta, z_{1:n}, \tilde{X}_n)}{q(\theta)q(z_{1:n})}\right)^\alpha d\theta dz_{1:n}. \quad (5.19)$$

Notice that the objective above does not require computing the model evidence $P(\tilde{X}_n)$. In this section, we aim to show that $q_r^*(\theta|\tilde{X}_n)$ converges weakly to the Dirac delta distribution at θ_0. To show this we require the following additional assumptions:

First, we define the profile likelihood at $\theta = \theta_0 + n^{-1/2}h_n$ for any bounded and stochastic $h_n = o_{P_0^n}(1)$ as $p(\tilde{X}_n|\theta_0 + n^{-1/2}h_n, z_{1:n}^p)$, where $z_{1:n}^p = \text{argmax}_{z_{1:n}}\, p(\tilde{X}_n|\theta_0 + n^{-1/2}h_n, z_{1:n})$ is the maximum profile likelihood estimate of $z_{1:n}$ at $\theta = \theta_0 + n^{-1/2}h_n$. Denote $d_H(z_{1:n}, z_{1:n}^p) := H(P_{\theta_0,z_{1:n}}, P_{\theta_0,z_{1:n}^p})$ as the Helinger distance between models $P_{\theta_0,z_{1:n}}$ and $P_{\theta_0,z_{1:n}^p}$ Furthermore, for any $\rho > 0$ and for all bounded and stochastic $h_n = o_{P_0^n}(1)$, we define $D(\theta_0 + n^{-1/2}h_n, \rho) = \{z_{1:n} : d_H(z_{1:n}, z_{1:n}^p) < \rho\}$ as the Hellinger ball of radius ρ around $z_{1:n}^p$.

Next we impose regularity conditions on the conditioned latent posterior $p(z_{1:n}|\tilde{X}_n, \theta_0)$. Following Wang and Blei [27, Proposition 10] and motivated by Bickel, Kleijn, *et al.* [128, Theorem 4.2], we assume that

Assumption 5.5.1 (Conditioned latent posterior)**.** *The conditioned latent posterior* $p(z_{1:n}|\tilde{X}_n, \theta_0)$ *satisfies*

1. *The conditioned latent posterior is consistent under $n^{-1/2}$-perturbation at some rate ρ_n with $\rho_n \downarrow 0$ and $n\rho_n^2 \to \infty$; that is for all bounded, stochastic $h_n = O_{P_0^n}(1)$, $p(z_{1:n}|\tilde{X}_n, \theta_0)$ converges as*

$$\int_{D^c(\theta_0 + n^{-1/2}h_n, \rho_n)} p(z_{1:n}|\tilde{X}_n, \theta = \theta_0 + n^{-1/2}h_n) dz_{1:n} = O_{P_0^n}(1).$$

2. *The sequence $\{\rho_n\}$ as defined above should also satisfy the following conditions for all bounded and stochastic $h_n = O_{P_0^n}(1)$:*

$$\text{(i)} \quad \sup_{z_{1:n} \in \{z_{1:n} : d_H(z_{1:n}, z_{1:n}^p) < \rho_n\}} \mathbb{E}_{P_{\theta_0, z_{1:n}}} \left[\frac{p(\tilde{X}_n|z_{1:n}, \theta_0 + n^{-1/2}h_n)}{p(\tilde{X}_n|z_{1:n}, \theta_0)} \right] = O(1),$$

$$\text{(ii)} \quad d_H(z_0, z_{1:n}^p) = o(\rho_n).$$

The first condition ensures that conditioned latent posterior converges slower than the true posterior and the second condition is an additional regularity condition on the expected likelihood ratio. Bickel, Kleijn, *et al.* [128, Lemma 4.3] identifies mild differentiablity conditions on the likelihood ratio that imply condition 2(i) above. Also, Theorem 3.1 in Bickel, Kleijn, *et al.* [128] provide the regularity conditions under which the the conditioned latent posterior satisfies the first condition above.

The next assumption, adapted from Bickel, Kleijn, *et al.* [128], is an extension of LAN condition in Assumption 5.2.2 to models with both global and local latent parameters.

Assumption 5.5.2 (Stochastic LAN (s-LAN))**.** *Fix $\theta_0 \in \Theta$ and recall that $z_{1:n}^p$ is the profile likelihood maximizer. The sequence of log-likelihood functions $\{P_{\theta_0, z_{1:n}^p}^n := p(\tilde{X}_n|\theta_0, z_{1:n}^p)\}$ satisfies* stochastic local asymptotic normality (s-LAN) *condition if there exists a matrix $I(\theta_0, z_0)$ and a sequence of random vectors $\{\Delta_{n,(\theta_0,z_0)}\} \in L_2(P_{\theta_0,z_{1:n}}^n)$ such that for every bounded and stochastic sequence $\{h_n\}$, that is $h_n = O_{P_0^n}(1)$, we have*

$$\log \frac{P_{\theta_0 + n^{-1/2}h_n, z_{1:n}^p}^n}{P_{\theta_0, z_{1:n}^p}^n} = h_n^T I(\theta_0, z_0) \Delta_{n,(\theta_0,z_0)} - \frac{1}{2} h_n^T I(\theta_0, z_0) h_n + o_{P_0^n}(1),$$

where $P_0 = P_{\theta_0, z_0}$.

Stochastic LAN is slightly stronger than the usual LAN property. In most of the examples, the ordinary LAN property often extends to stochastic LAN without significant difficulties [128]. Also, Theorem 1 in Murphy and Vaart [129] identifies conditions under which the above LAN assumption is satisfied by models with both global and local latent variables. It must be noted that if $\hat{\theta}_n$ is an asymptotically efficient estimator of θ_0, then according to Lemma 25.25 in [109] $\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) = \Delta_{n,(\theta_0,z_0)} + o_{P_0^n}(1)$.

Next we state a modified version of Assumption 5.2.4(3) for the models that contain local latent variables:

Assumption 5.5.3 (Good Sequence-Local)**.** *For any $\bar{M} > 0$, the variational family $\bar{\mathcal{Q}}$ contains a sequence of distributions $\{\bar{q}_n(\theta)\}$ with the following properties:*

(1) there exists $n_1 \geq 1$ such that $\int_\Theta \theta \bar{q}_n(\theta) d\theta = \hat{\theta}_n$, where $\hat{\theta}_n$ is the maximum likelihood estimate, for each $n \geq n_1$,

(2) there exists $n_{\bar{M}} \geq 1$ such that the rate of convergence is $\gamma_n = \sqrt{n}$, that is $E_{\bar{q}_n(\theta)}[|\theta - \hat{\theta}_n|^2] \leq \frac{\bar{M}}{\gamma_n^2}$ for each $n \geq n_{\bar{M}}$,

(3) there exist a compact ball $K \subset \Theta$ containing the true parameter θ_0 and $n_2 \geq 1$, such that the sequence of Radon-Nikodym derivatives of the Bayes posterior density with respect to the sequence $\{\bar{q}_n\}$ exists and is bounded above by a finite positive constant M_r outside of K for all $n \geq n_2$; that is,

$$\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n, z^0_{1:n})}{\bar{q}_n(\theta)} \leq M_r, \ \forall \theta \in \Theta \backslash K \text{ and } \forall n \geq n_2, \quad P_0 - a.s,$$

where $z^0_{1:n}$ is the first n components of the true local latent parameter z_0.

(4) there exists $n_3 \geq 1$ such that the good sequence $\{\bar{q}_n(\theta)\}$ is log-concave in θ for all $n \geq n_3$.

Example 5.5.1 (Bayesian mixture model)**.** *Consider a mixture of uncorrelated L uni-variate Gaussians, each with mean $\mu_{\mathrm{i}}, \mathrm{i} \in \{1, 2, \ldots, L\}$ and unit variance. Each observation ξ_{i} is assumed to be generated using the following model:*

$$\begin{aligned}
&\mu_l \sim \boldsymbol{\pi}(\mu_l), \forall l \in \{1, 2, \ldots, L\} \\
&z_{\mathrm{i}} \sim \textit{Categorical}\left(\frac{1}{L}, \frac{1}{L}, \ldots, \frac{1}{L}\right), \forall \mathrm{i} \in \{1, 2, \ldots, n\} \\
&\xi_{\mathrm{i}} \sim \mathcal{N}(z_{\mathrm{i}}^T \boldsymbol{\mu}, 1) \forall \mathrm{i} \in \{1, 2, \ldots, n\}
\end{aligned}$$

Notice that $\boldsymbol{\mu}$ is the global and $z_{1:n}$ are the local latent parameters. Now observe that

$$\begin{aligned}
\boldsymbol{\pi}(\boldsymbol{\mu}|\tilde{X}_n, z^0_{1:n}) = \frac{\prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{i}=1}^n p(z^0_{\mathrm{i}}, \xi_{\mathrm{i}}|\boldsymbol{\mu})}{\int \prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{i}=1}^n p(z^0_{\mathrm{i}}, \xi_{\mathrm{i}}|\boldsymbol{\mu}) d\boldsymbol{\mu}} &= \frac{\prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{i}=1}^n p(\xi_{\mathrm{i}}|\boldsymbol{\mu}, z^0_{\mathrm{i}})}{\int \prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{i}=1}^n p(\xi_{\mathrm{i}}|\boldsymbol{\mu}, z^0_{\mathrm{i}}) d\boldsymbol{\mu}} \\
&= \frac{\prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{i}=1}^n \mathcal{N}(\xi_{\mathrm{i}}|\boldsymbol{\mu}^T z^0_{\mathrm{i}}, 1)}{\int \prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{i}=1}^n \mathcal{N}(\xi_{\mathrm{i}}|\boldsymbol{\mu}^T z^0_{\mathrm{i}}, 1) d\boldsymbol{\mu}} \quad (5.20) \\
&= \frac{\prod_{l=1}^L \left[\boldsymbol{\pi}(\mu_l) \prod_{\mathrm{j}=1}^{n_l} \mathcal{N}(\xi^l_{\mathrm{j}}|\mu_l, 1)\right]}{\int \prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{j}=1}^{n_l} \mathcal{N}(\xi^l_{\mathrm{j}}|\mu_l, 1) d\boldsymbol{\mu}}, \quad (5.21)
\end{aligned}$$

where ξ^l_{j} is the j^{th} observation in the l^{th} cluster and $n_l = \sum_{\mathrm{i}=1}^n z^0_{\mathrm{i},l}$ is the total number of observations in the l^{th} cluster . In practice, $\boldsymbol{\pi}(\mu_l) \sim \mathcal{N}(\mu_l|m, \sigma^2)\}$ is assumed to be Gaussian (conjugate) with known mean (m) and variance (σ^2) hyper-parameters, hence the distribution in (5.21) can be computed analytically, that is

$$\boldsymbol{\pi}(\mu|\tilde{X}_n, z^0_{1:n}) = \frac{\prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{i}=1}^n p(z^0_{\mathrm{i}}, \xi_{\mathrm{i}}|\mu)}{\int \prod_{l=1}^L \boldsymbol{\pi}(\mu_l) \prod_{\mathrm{i}=1}^n p(z^0_{\mathrm{i}}, \xi_{\mathrm{i}}|\mu) d\mu} = \prod_{l=1}^L \mathcal{N}\left(\mu_l \middle| \frac{1}{\frac{1}{\sigma^2} + n_l}\left(\frac{m}{\sigma^2} + \sum_{\mathrm{j}=1}^{n_l} \xi^l_{\mathrm{j}}\right), \left(\frac{1}{\sigma^2} + n_l\right)^{-1}\right).$$

In practice $\bar{\mathcal{Q}}$ is chosen to be a mean-field approximate family, in particular it is a product of L uni-variate Gaussians. Now consider the following sequence of distributions in $\bar{\mathcal{Q}}$, that is

$$q_n(\mu) = \prod_{l=1}^L \mathcal{N}\left(\mu_l|m_{n,l}, \sigma^2_{n,l}\right).$$

Clearly, by choosing $m_{n,l} = \frac{1}{\frac{1}{\sigma^2}+n_l}\left(\frac{m}{\sigma^2} + \sum_{\mathrm{j}=1}^{n_l} \xi^l_{\mathrm{j}}\right)$ and $\sigma^2_{n,l} = \left(\frac{1}{\sigma^2} + n_l\right)^{-1}$, the ratio $\frac{\boldsymbol{\pi}(\mu|\tilde{X}_n, z^0_{1:n})}{\bar{q}_n(\theta)}$ is bounded by 1.

The s-LAN assumption for finite mixtures model follows from the finiteness of the support of local latent variables [129], [130].

In the next result we show that a consistent sequence asymptotically achieves zero $\alpha-$Rényi divergence. To show its existence, we first provide an asymptotic upper-bound on the minimum of the LHS in (5.25) in the next proposition. This will allow us to prove the consistency of the minimizing sequence.

Proposition 5.5.1. *For a given $\alpha > 1$ and under Assumptions 5.2.1, 5.2.3 (for $\bar{\mathcal{Q}}$), 5.5.1, 5.5.2, 5.5.3, and for any good sequence there exist $n_0 \geq 1$ and $\bar{M} > 0$ such that for all $n \geq n_0$, the minimal $\alpha-$Rényi divergence satisfies*

$$\begin{aligned}\min_{q\in\bar{\mathcal{Q}}}\min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)\|q(\theta)q(z_{1:n})) &\leq \min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)\|\bar{q}_n(\theta)q(z_{1:n})) \\ &\leq B = \frac{1}{2}\log\left(\frac{\bar{\mathrm{e}}\bar{M}I(\theta_0, z_0)}{\alpha^{\frac{1}{\alpha-1}}}\right) + o_{P_0^n}(1) \qquad (5.22)\end{aligned}$$

where $\bar{\mathrm{e}}$ is the Euler's constant and $I(\theta_0, z_0)$ is as defined in Assumption 5.5.2.

Since the term on the RHS above in (5.22) is non-negative for all $n \geq n_0$, implying that $\bar{M} \geq \frac{\alpha^{\frac{1}{\alpha-1}}}{\bar{\mathrm{e}}I(\theta_0,z_0)}$ for all $n \geq n_0$. Therefore, a specific good sequence can be chosen by fixing $\tilde{M} = \frac{\alpha^{\frac{1}{\alpha-1}}}{\bar{\mathrm{e}}I(\theta_0,z_0)}$, implying that $\limsup_{n\to\infty} \min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)\|\tilde{q}_n(\theta)q(z_{1:n})) = 0\ \forall n \geq n_0$. Now analogous to the parametric case we are only left to show that the global Rényi approximator necessarily converges to a Dirac delta distribution concentrated at the true global parameter θ_0 to achieve zero Rényi divergence.

Now notice that for any $n \geq 1$,

$$\begin{aligned}&\min_{q(z_{1:n})\in\mathcal{Q}^n} \log\int_\Theta q(\theta)\left(\frac{\boldsymbol{\pi}(\theta)}{q(\theta)}\right)^\alpha \int_{\mathcal{Z}^n} q(z_{1:n})\left(\frac{p(z_{1:n}, \tilde{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n}d\theta \\ &\quad\geq \log\int_\Theta q(\theta)\left(\frac{\boldsymbol{\pi}(\theta)}{q(\theta)}\right)^\alpha \min_{q(z_{1:n})\in\mathcal{Q}^n}\int_{\mathcal{Z}^n} q(z_{1:n})\left(\frac{p(z_{1:n}, \tilde{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n}d\theta \\ &\quad= \log\int_\Theta q(\theta)\left(\frac{\boldsymbol{\pi}(\theta)M(\tilde{X}_n|\theta)}{q(\theta)}\right)^\alpha d\theta, \qquad (5.23)\end{aligned}$$

where $M(\tilde{X}_n|\theta)$ is the variational likelihood define as

$$M(\tilde{X}_n|\theta) := \left[\min_{q(z_{1:n})\in\mathcal{Q}^n} \int_{\mathcal{Z}^n} q(z_{1:n}) \left(\frac{p(z_{1:n}, \tilde{X}_n|\theta)}{q(z_{1:n})}\right)^{\alpha} dz_{1:n}\right]^{1/\alpha}. \tag{5.24}$$

Observe that subtracting the $\log P(\tilde{X}_n)^{\alpha}$ from either side of (5.23) yields:

$$\min_{q(z_{1:n})\in\mathcal{Q}^n} D_{\alpha}(\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)\|q(\theta)q(z_{1:n})) \geq D_{\alpha}(\boldsymbol{\pi}^*(\theta|\tilde{X}_n)\|q(\theta)), \tag{5.25}$$

where the ideal posterior $\boldsymbol{\pi}^*(\theta|\tilde{X}_n)$ is defined as

$$\boldsymbol{\pi}^*(\theta|\tilde{X}_n) := \frac{\boldsymbol{\pi}(\theta)M(\tilde{X}_n|\theta)}{\int \boldsymbol{\pi}(\theta)M(\tilde{X}_n|\theta)d\theta}. \tag{5.26}$$

In the subsequent lemma we show that under certain regularity conditions $M(\tilde{X}_n|\theta)$ satisfies the LAN condition with the similar expansion as of the true likelihood model for a given local latent parameter z_0. The proof parallels that of Wang and Blei [27, Proposition 10].

Lemma 5.5.1. *Fix $\theta \in \Theta$. Under Assumptions 5.5.1 and 5.5.2, the sequence of variational log-likelihood functions $\{M_n(\theta) := \log M(\tilde{X}_n|\theta)$ satisfies* s-LAN *condition, that is there exists a matrix $I(\theta_0, z_0)$ and a sequence of random vectors $\{\Delta_{n,(\theta_0,z_0)}\}$ as defined in Assumption 5.5.2, such that for every bounded and stochastic sequence $\{h_n\}$, that is $h_n = O_{P_0^n}(1)$, we have*

$$\log \frac{M_n(\theta_0 + n^{-1/2}h_n)}{M_n(\theta_0)} = h_n^T I(\theta_0, z_0)\Delta_{n,(\theta_0,z_0)} - \frac{1}{2}h_n^T I(\theta_0, z_0)h_n + o_{P_0^n}(1).$$

Next, we will show that the minimizing sequence must converge to a Dirac delta distribution at θ_0 using the results in Proposition 5.5.1 and Lemma 5.5.1.

Theorem 5.5.1. *For a given $\alpha > 1$ and under Assumptions 5.2.1, 5.2.3 (for $\bar{\mathcal{Q}}$) , 5.5.1, and 5.5.3, the $\alpha-$Rényi approximate posterior $q_r^*(\theta|\tilde{X}_n)$ over global latent parameters θ as*

defined in (5.19) *converges weakly to a Dirac delta distribution at the true parameter* θ_0*; that is,*

$$q_r^*(\theta|\tilde{X}_n) \Rightarrow \delta_{\theta_0} \quad \textit{in } P_0^n - \textit{probability as } n \to \infty.$$

Proof. Using the result in Proposition 5.5.1 and following similar steps as used in Theorem 5.3.1, we can show that the minimizing sequence must have zero α-Rényi divergence in the limit with high probability. Recall the inequality in (5.25)

$$\min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)\|q(\theta)q(z_{1:n})) \geq D_\alpha(\boldsymbol{\pi}^*(\theta|\tilde{X}_n)\|q(\theta)). \tag{5.27}$$

Also note that $q_r^*(\theta|\tilde{X}_n)$ is the minimizer of the LHS in the equation above. Since the variational likelihood satisfies the LAN condition due to Lemma 5.5.1, under the consistent testability assumption, the ideal posterior $\boldsymbol{\pi}^*(\theta|\tilde{X}_n)$ also degenerates to a Dirac delta distribution at the true parameter θ_0 [131].

Now recall Lemma 5.3.1 and 5.3.2. Following the arguments in Lemma 5.3.1, and using the inequality in (5.27) we can argue that any sequence of distributions in $\bar{\mathcal{Q}}$ that minimizes the LHS in (5.27) must converge weakly to a Dirac delta distribution at the true parameter θ_0 in the large sample limit, since otherwise the objective in the LHS of (5.27) is unbounded. In addition, using Lemma 5.3.2 and the inequality in (5.27) we can also show that any sequence of distribution in $\bar{\mathcal{Q}}$ that converges weakly to a convex combination of a Dirac delta distribution at θ_0 with any other distribution can not achieve zero $\alpha-$Rényi divergence in the limit. This completes the proof. □

5.6 Proofs

We begin with the following standard lemma.

Lemma 5.6.1. *[Laplace Approximation of integrals] Consider an integral of the form*

$$I = \int_a^b h(y)\mathrm{e}^{-ng(y)}dy,$$

where $g(y)$ is a smooth function which has a local minimum at $y^* \in (a,b)$ and $h(y)$ is a smooth function. Then

$$I \sim h(y^*)\mathrm{e}^{-ng(y^*)}\sqrt{\frac{2\pi}{ng(y^*)}} \text{ as } n \to \infty.$$

Proof. Readers are directed to Wong [132, Chapter-2] for the proof. □

Now we prove a technical lemma that bounds the differential entropy of the good sequence.

Lemma 5.6.2. *For a good sequence $\bar{q}_n(\theta)$, there exist an $n_M \geq 1$ and $\bar{M} > 0$, such that for all $n \geq n_M$*

$$-\int \bar{q}_n(\mu)\log \bar{q}_n(\mu) \leq \frac{1}{2}\log\left(2\pi\bar{\mathrm{e}}\frac{\bar{M}}{n}\right),$$

where $\bar{\mathrm{e}}$ is the Euler's constant.

Proof. Recall from Assumption 5.2.4 that the $\bar{q}_n(\theta)$ converges weakly to δ_{θ_0} at the rate of $\sqrt{n}$. It follows from the Definition 5.2.5 for rate of convergence that,

$$E_{\bar{q}_n(\theta)}[|\theta - \hat{\theta}_n|^2] = O\left(\frac{1}{n}\right).$$

There exist an $n_M \geq 1$ and $\bar{M} > 0$, such that for all $n \geq n_M$

$$\mathbb{E}_{\bar{q}_n(\theta)}[(\theta - \hat{\theta}_n)^2] \leq \frac{\bar{M}}{n}.$$

Using the fact that, the differential entropy of random variable with a given variance is bounded by the differential entropy of the Gausian distribution of the same variance [133, Theorem 9.6.5]), it follows that the differential entropy of $\bar{q}_n(\mu)$ is bounded by $\frac{1}{2}\log(2\pi\bar{\mathrm{e}}\frac{\bar{M}}{n})$, where $\bar{\mathrm{e}}$ is the Euler's constant. □

Next, we prove the following result on the prior distributions. This result will be useful in proving Lemma 5.6.4 and 5.3.1.

Lemma 5.6.3. *Given a prior distribution $\boldsymbol{\pi}(\theta)$ with $\mathbb{E}_{\boldsymbol{\pi}(\theta)}[|\theta|] < \infty$, for any $\beta > 0$, there exists a sequence of compact sets $\{K_n\} \subset \Theta$ such that*

$$\int_{\Theta \setminus K_n} \boldsymbol{\pi}(\gamma) d\gamma = O(n^{-\beta}).$$

Proof. Fix $\theta_1 \in \Theta$. Define a sequence of compact sets

$$K_n = \{\theta \in \Theta : |\theta - \theta_1| \leq n^{\beta}\} \forall \beta > 0.$$

Clearly, as n increases K_n approaches Θ. Now, using the Markov's inequality followed by the triangular inequality,

$$\begin{aligned} \int_{\Theta \setminus K_n} \boldsymbol{\pi}(\gamma) d\gamma = \int_{\{\gamma \in \Theta : |\gamma - \theta_1| > n^{\beta}\}} \boldsymbol{\pi}(\gamma) d\gamma &\leq n^{-\beta} \mathbb{E}_{\boldsymbol{\pi}(\theta)}[|\gamma - \theta_1|] \\ &\leq n^{-\beta} \left(\mathbb{E}_{\boldsymbol{\pi}(\theta)}[|\gamma|] + |\theta_1| \right). \end{aligned} \tag{5.28}$$

Since, $\mathbb{E}_{\boldsymbol{\pi}(\gamma)}[|\gamma|] < \infty$, it follows that $\forall \beta > 0$, $\int_{\Theta \setminus K_n} \boldsymbol{\pi}(\gamma) d\gamma = O(n^{-\beta})$. □

The next result approximates the normalizing sequence of the posterior distribution using the lemma above and the LAN condition.

Lemma 5.6.4. *There exists a sequence of compact balls $\{K_n \subset \Theta\}$, such that $\theta_0 \in K_n$ and under Assumptions 5.2.1 and 5.2.2, the normalizing sequence of the posterior distribution*

$$\begin{aligned} &\int_{\Theta} \prod_{\mathrm{i}=1}^{n} \frac{p(\xi_{\mathrm{i}}|\gamma)}{p(\xi_{\mathrm{i}}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma \\ &= \sqrt{\frac{2\boldsymbol{\pi}}{nI(\theta_0)}} \mathrm{e}^{\left(\frac{1}{2} nI(\theta_0)((\hat{\theta}_n - \theta_0)^2)\right)} \left(\mathrm{e}^{o_{P_0^n}(1)} \int_{K_n} \boldsymbol{\pi}(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\gamma + o(1) \right). \end{aligned} \tag{5.29}$$

Proof. Let $\{K_n \subset \Theta\}$ be a sequence of compact balls such that $\theta_0 \in K_n$, where θ_0 is any point in Θ where prior distribution $\boldsymbol{\pi}(\theta)$ places positive density. Using Lemma 5.6.3, we can

always find a sequence of sets $\{K_n\}$ for a prior distribution, such that $\theta_0 \in K_n$ and for any positive constant $\beta > \frac{3}{2}$,

$$\int_{\Theta \setminus K_n} \boldsymbol{\pi}(\gamma) d\gamma = O(n^{-\beta}). \tag{5.30}$$

Observe that

$$\int_{\Theta} \prod_{\mathrm{i}=1}^{n} \frac{p(\xi_{\mathrm{i}}|\gamma)}{p(\xi_{\mathrm{i}}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma = \left(\int_{K_n} \prod_{\mathrm{i}=1}^{n} \frac{p(\xi_{\mathrm{i}}|\gamma)}{p(\xi_{\mathrm{i}}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma + \int_{\Theta \setminus K_n} \boldsymbol{\pi}(\gamma) \prod_{\mathrm{i}=1}^{n} \frac{p(\xi_{\mathrm{i}}|\gamma)}{p(\xi_{\mathrm{i}}|\theta_0)} d\gamma \right). \tag{5.31}$$

Consider the first term in (5.31); following similar steps as in (5.49) and (5.50) and using Assumption 5.2.2, we have

$$\begin{aligned}
&\int_{K_n} \prod_{\mathrm{i}=1}^{n} \frac{p(\xi_{\mathrm{i}}|\gamma)}{p(\xi_{\mathrm{i}}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma \\
&\quad = \mathrm{e}^{o_{P_0^n}(1)} \exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2 \right) \right) \int_{K_n} \boldsymbol{\pi}(\gamma) \exp\left(-\frac{1}{2} n I(\theta_0) \left((\gamma - \hat{\theta}_n)^2 \right) \right) d\gamma \\
&\quad = \mathrm{e}^{o_{P_0^n}(1)} \exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2 \right) \right) \sqrt{\frac{2\pi}{n I(\theta_0)}} \int_{K_n} \boldsymbol{\pi}(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (n I(\theta_0))^{-1}) d\gamma,
\end{aligned} \tag{5.32}$$

where the last equality follows from the definition of Gaussian density, $\mathcal{N}(\cdot; \hat{\theta}_n, (n I(\theta_0))^{-1})$.

Substituting (5.32) into (5.31), we obtain

$$\begin{aligned}
&\int_{\Theta} \prod_{\mathrm{i}=1}^{n} \frac{p(\xi_{\mathrm{i}}|\gamma)}{p(\xi_{\mathrm{i}}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma \\
&= \exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2 \right) \right) \sqrt{\frac{2\pi}{n I(\theta_0)}} \left(\mathrm{e}^{o_{P_0^n}(1)} \int_{K_n} \boldsymbol{\pi}(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (n I(\theta_0))^{-1}) d\gamma \right. \\
&\quad \left. + \exp\left(-\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2 \right) \right) \sqrt{\frac{n I(\theta_0)}{2\pi}} \int_{\Theta \setminus K_n} \boldsymbol{\pi}(\gamma) \prod_{\mathrm{i}=1}^{n} \frac{p(\xi_{\mathrm{i}}|\gamma)}{p(\xi_{\mathrm{i}}|\theta_0)} d\gamma \right).
\end{aligned} \tag{5.33}$$

Next, using the Markov's inequality and then Fubini's Theorem, for arbitrary $\delta > 0$, we have

$$\begin{aligned} P_0^n\left(\sqrt{\frac{nI(\theta_0)}{2\pi}}\int_{\Theta\setminus K_n}\prod_{i=1}^n\frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\pi(\gamma)d\gamma > \delta\right) &\leq \sqrt{\frac{nI(\theta_0)}{\delta^2 2\pi}}\mathbb{E}_{P_0^n}\left[\int_{\Theta\setminus K_n}\prod_{i=1}^n\frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\pi(\gamma)d\gamma\right] \\ &= \sqrt{\frac{nI(\theta_0)}{\delta^2 2\pi}}\int_{\Theta\setminus K_n}\mathbb{E}_{P_0^n}\left[\prod_{i=1}^n\frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\right]\pi(\gamma)d\gamma \\ &= \sqrt{\frac{nI(\theta_0)}{\delta^2 2\pi}}\int_{\Theta\setminus K_n}\pi(\gamma)d\gamma, \end{aligned} \tag{5.34}$$

since $\mathbb{E}_{P_0^n}\left[\prod_{i=1}^n\frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\right] = 1$.

Hence, using (5.30) for $\beta > 3/2$, it is straightforward to observe that

$$P_0^n\left(\sqrt{\frac{nI(\theta_0)}{2\pi}}\int_{\Theta\setminus K_n}\prod_{i=1}^n\frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\pi(\gamma)d\gamma > \delta\right) \leq \sqrt{\frac{I(\theta_0)}{\delta^2 2\pi}}\frac{1}{n^{\beta-1/2}}.$$

Since the upper bound above is summable, using First Borel-Cantelli Theorem it follows that

$$\sqrt{\frac{nI(\theta_0)}{2\pi}}\int_{\Theta\setminus K_n}\prod_{i=1}^n\frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\pi(\gamma)d\gamma = o(1)\ P_0 - \text{a.s.}. \tag{5.35}$$

Since, $\exp\left(-\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right) \leq 1$, it follows from substituting (5.35) into (5.33) that

$$\begin{aligned} &\int_\Theta\prod_{i=1}^n\frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\pi(\gamma)d\gamma \\ &= \exp\left(\frac{1}{2}nI(\theta_0)\left((\hat{\theta}_n-\theta_0)^2\right)\right)\sqrt{\frac{2\pi}{nI(\theta_0)}}\left(e^{o_{P_0^n}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma + o(1)\right). \end{aligned}$$

□

Next we prove Lemma 5.3.1, showing that the $\alpha-$Rényi divergence between the posterior and any non-degenerate distribution diverges in the large sample limit.

Proof of Lemma 5.3.1. Let $K_n \subset \Theta$ be a sequence of compact sets such that $\theta_0 \in K_n$, where θ_0 is any point in Θ where prior distribution $\pi(\theta)$ places positive density. Using Lemma 5.6.3,

we can always find a sequence of sets $\{K_n\}$ for a prior distribution, such that $\theta_0 \in K_n$ and for any positive constant $\beta > \frac{1}{2}$,

$$\int_{\Theta \setminus K_n} \boldsymbol{\pi}(\gamma) d\gamma = O(n^{-\beta}). \tag{5.36}$$

Now, observe that

$$\begin{aligned}
&\frac{\alpha - 1}{\alpha} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n) \| q_n(\theta)) \\
&\quad = \frac{1}{\alpha} \log \left(\int_{K_n} q_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q_n(\theta)} \right)^\alpha d\theta + \int_{\Theta \setminus K_n} q_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q_n(\theta)} \right)^\alpha d\theta \right) \\
&\quad \geq \frac{1}{\alpha} \log \left(\int_{K_n} q_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q_n(\theta)} \right)^\alpha d\theta \right),
\end{aligned} \tag{5.37}$$

where the last inequality follows from the fact that the integrand is always positive.

Next, we approximate the ratio in the integrand on the right hand side of the above equation using the LAN condition in Assumption 5.2.2. Let $\Delta_{n,\theta_0} := \sqrt{n}(\hat{\theta}_n - \theta_0)$, such that $\hat{\theta}_n \to \theta_0$, $P_0 - a.s.$ and Δ_{n,θ_0} converges in distribution to $\mathcal{N}(0, I(\theta_0)^{-1})$. Re-parameterizing the expression with $\theta = \theta_0 + n^{-1/2}h$, we have

$$\begin{aligned}
\int_{K_n} q_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q_n(\theta)} \right)^\alpha d\theta &= n^{-1/2} \int_{K_n} q_n(\theta_0 + n^{-1/2}h) \left(\frac{\boldsymbol{\pi}(\theta_0 + n^{-1/2}h) \prod_{i=1}^n \frac{p(\xi_i|(\theta_0 + n^{-1/2}h))}{p(\xi_i|\theta_0)}}{q_n(\theta_0 + n^{-1/2}h) \int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma} \right)^\alpha dh \\
&= n^{-1/2} \int_{K_n} q_n(\theta_0 + n^{-1/2}h) \left(\frac{\boldsymbol{\pi}(\theta_0 + n^{-1/2}h) \prod_{i=1}^n \frac{p(\xi_i|(\theta_0 + n^{-1/2}h))}{p(\xi_i|\theta_0)}}{q_n(\theta_0 + n^{-1/2}h) \int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma} \right)^\alpha dh \qquad (5.38) \\
&= n^{-1/2} \int_{K_n} q_n(\theta_0 + n^{-1/2}h) \left(\boldsymbol{\pi}(\theta_0 + n^{-1/2}h) \frac{\exp(hI(\theta_0)\Delta_{n,\theta_0} - \frac{1}{2}h^2 I(\theta_0) + o_{P_0^n}(1))}{q_n(\theta_0 + n^{-1/2}h) \int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma} \right)^\alpha dh. \qquad (5.39)
\end{aligned}$$

Resubstituting $h = \sqrt{n}(\theta - \theta_0)$ in the expression above and reverting to the previous parametrization,

$$= \int_{K_n} q_n(\theta) \left(\pi(\theta) \frac{\exp\left(\sqrt{n}(\theta-\theta_0) I(\theta_0) \Delta_{n,\theta_0} - \frac{1}{2} n (\theta-\theta_0)^2 I(\theta_0) + o_{P_0^n}(1)\right)}{q_n(\theta) \int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)} \pi(\gamma) d\gamma} \right)^\alpha d\theta$$

$$= \int_{K_n} q_n(\theta) \left(\pi(\theta) \frac{e^{o_{P_0^n}(1)} \exp\left(-\frac{1}{2} n I(\theta_0) \left((\theta-\theta_0)^2 - 2(\theta-\theta_0)(\hat{\theta}_n - \theta_0)\right)\right)}{q_n(\theta) \int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)} \pi(\gamma) d\gamma} \right)^\alpha d\theta.$$

Now completing the square by dividing and multiplying the numerator by $\exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2\right)\right)$ we obtain

$$= \int_{K_n} q_n(\theta) \left(\pi(\theta) \frac{e^{o_{P_0^n}(1)} \exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2\right)\right) \exp\left(-\frac{1}{2} n I(\theta_0) \left((\theta - \hat{\theta}_n)^2\right)\right)}{q_n(\theta) \int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)} \pi(\gamma) d\gamma} \right)^\alpha d\theta$$

$$= \int_{K_n} q_n(\theta) \left(\pi(\theta) \frac{e^{o_{P_0^n}(1)} \exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2\right)\right) \sqrt{\frac{2\pi}{nI(\theta_0)}} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1})}{q_n(\theta) \int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)} \pi(\gamma) d\gamma} \right)^\alpha d\theta, \tag{5.40}$$

where, in the last equality we used the definition of Gaussian density, $\mathcal{N}(\cdot; \hat{\theta}_n, (nI(\theta_0))^{-1})$.

Next, we approximate the integral in the denominator of (5.50). Using Lemma 5.6.4 it follows that, there exist a sequence of compact balls $\{K_n \subset \Theta\}$, such that $\theta_0 \in K_n$ and

$$\int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)} \pi(\gamma) d\gamma$$
$$= \sqrt{\frac{2\pi}{nI(\theta_0)}} e^{\left(\frac{1}{2} nI(\theta_0)\left((\hat{\theta}_n - \theta_0)^2\right)\right)} \left(e^{o_{P_0^n}(1)} \int_{K_n} \pi(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\gamma + o(1) \right). \tag{5.41}$$

Substituting (5.41) into (5.40) and simplifying, we obtain

$$\int_{K_n} q_n(\theta) \left(\frac{\pi(\theta|\tilde{X}_n)}{q_n(\theta)} \right)^\alpha d\theta$$
$$= \int_{K_n} q_n(\theta)^{1-\alpha} \left(\frac{e^{o_{P_0^n}(1)} \pi(\theta) \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1})}{\left(e^{o_{P_0^n}(1)} \int_{K_n} \pi(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\gamma + o(1) \right)} \right)^\alpha d\theta. \tag{5.42}$$

Observe that:

$$\left(\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})\right)^{\alpha}=\left(\sqrt{\frac{nI(\theta_0)}{2\pi}}\right)^{\alpha}\left(\sqrt{\frac{2\pi}{n\alpha I(\theta_0)}}\right)\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1}).$$

Substituting this into the right hand side of (5.42)

$$\begin{aligned}
&\frac{1}{\alpha}\log\int_{K_n}q_n(\theta)^{1-\alpha}\left(\frac{\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(\mathrm{e}^{o_{P_0^n}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)}\right)^{\alpha}d\theta\\
&=-\log\left(\mathrm{e}^{o_{P_0^n}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)+\frac{\alpha-1}{2\alpha}\log n-\frac{\log\alpha}{2\alpha}\\
&\quad+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{1}{\alpha}\log\int_{K_n}q_n(\theta)^{1-\alpha}\pi(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})d\theta. \qquad (5.43)
\end{aligned}$$

From the Laplace approximation (Lemma 5.6.1) and the continuity of the logarithm, we have

$$-\log\left(\mathrm{e}^{o_{P_0^n}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)\sim-\log\left(\mathrm{e}^{o_{P_0^n}(1)}\pi(\hat{\theta}_n)\right).$$

Next, using the Laplace approximation on the last term in (5.43)

$$\frac{1}{\alpha}\log\int_{K_n}q_n(\theta)^{1-\alpha}\pi(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})d\theta\sim\frac{\alpha-1}{\alpha}\log\frac{1}{q_n(\hat{\theta}_n)}+\log\pi(\hat{\theta}_n).$$

Substituting the above two approximations into (5.43), we have

$$\begin{aligned}
&\frac{1}{\alpha}\log\int_{K_n}q_n(\theta)^{1-\alpha}\left(\frac{\pi(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(\mathrm{e}^{o_{P_0^n}(1)}\int_{K_n}\pi(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)}\right)^{\alpha}d\theta\\
&\sim-\log\left(\mathrm{e}^{o_{P_0^n}(1)}\pi(\hat{\theta}_n)\right)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}\\
&\quad+\frac{\alpha-1}{2\alpha}\log n-\frac{\alpha-1}{\alpha}\log q_n(\hat{\theta}_n)+\log\pi(\hat{\theta}_n)\\
&\sim-\log\left(\pi(\hat{\theta}_n)\right)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{\alpha-1}{2\alpha}\log n-\frac{\alpha-1}{\alpha}\log q(\hat{\theta}_n)+\log\pi(\hat{\theta}_n)+o_{P_0^n}(1)\\
&=-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{\alpha-1}{2\alpha}\log n-\frac{\alpha-1}{\alpha}\log q(\hat{\theta}_n)+o_{P_0^n}(1), \qquad (5.44)
\end{aligned}$$

where the penultimate approximation follows from the fact that

$$q_n(\hat{\theta}_n) \sim q(\hat{\theta}_n).$$

Note that $\hat{\theta}_n \to \theta_0,\ P_0 - a.s.$ Therefore, if $q(\theta_0) = 0$, then the right hand side in (5.44) will diverge as $n \to \infty$ because $\frac{\alpha-1}{2\alpha} \log n$ also diverges as $n \to \infty$. Also observe that, for any $q(\theta)$ that places finite mass on θ_0, the $\alpha-$Rényi divergence diverges as $n \to \infty$. Hence, $\alpha-$Rényi approximate posterior must converge weakly to a distribution that has a Dirac delta distribution at the true parameter θ_0. □

Next, we show that the $\alpha-$Rényi divergence between the true posterior and the sequence $\{q_n(\theta)\} \in \mathcal{Q}$ as defined in (5.9) is bounded below by a positive number.

Proof of Lemma 5.3.2. Van Erven and Harremos [59, Theorem 19] shows that, for any $\alpha > 0$, the $\alpha-$Rényi divergence $D_\alpha(p(\theta)\|q(\theta))$ is a lower semi-continuous function of the pair $(p(\theta), q(\theta))$ in the weak topology on the space of probability measures. Recall from (5.6) that the true posterior distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ converges weakly to $\delta_{\theta_0}\ P_0 - a.s.$ Using this fact it follows that

$$\liminf_{n\to\infty} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q_n(\theta)) \geq D_\alpha\left(\delta_{\theta_0}\middle\|w^{\mathrm{j}}\delta_{\theta_0} + \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right) \quad P_0 - a.s.$$

Next, using Pinsker's inequality [133] for $\alpha > 1$, we have

$$\begin{aligned} D_\alpha\left(\delta_{\theta_0}\middle\|w^{\mathrm{j}}\delta_{\theta_0} + \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right) &\geq \frac{1}{2}\left(\int_\Theta \left|\delta_{\theta_0} - w^{\mathrm{j}}\delta_{\theta_0} - \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right| d\theta\right)^2 \\ &= \frac{1}{2}\left(\int_\Theta \left|(1-w^{\mathrm{j}})\delta_{\theta_0} - \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right| d\theta\right)^2. \end{aligned}$$

Now dividing the integral over ball of radius ϵ centered at θ_0, $B(\theta_0, \epsilon)$ and its complement, we obtain

$$\begin{aligned}
&\liminf_{n\to\infty} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q_n(\theta)) \\
&\quad \geq \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)}\left|(1-w^{\mathrm{j}})\delta_{\theta_0} - \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right| d\theta + \int_{B(\theta_0,\epsilon)^C}\left|(1-w^{\mathrm{j}})\delta_{\theta_0} - \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right| d\theta\right)^2 \\
&\quad \geq \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C}\left|(1-w^{\mathrm{j}})\delta_{\theta_0} - \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right| d\theta\right)^2 \\
&\quad = \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C}\left|- \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right| d\theta\right)^2 \quad P_0 - a.s. \qquad (5.45)
\end{aligned}$$

Since, $w^{\mathrm{i}} \in (0,1)$, observe that for any $\epsilon > 0$, there exists $\eta(\epsilon) > 0$, such that

$$\frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C}\left|- \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}} q_{\mathrm{i}}(\theta)\right| d\theta\right)^2 \geq \eta(\epsilon).$$

Therefore, it follows that

$$\liminf_{n\to\infty} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q_n(\theta)) \geq \eta(\epsilon) > 0 \quad P_0 - a.s.$$

□

In the following result, we show that if $q_{\mathrm{i}}(\theta), \mathrm{i} \in \{1, 2, \ldots\}$ in the definition of $\{q_n(\theta)\}$ in (5.9) are Dirac delta distributions then

$$\liminf_{n\to\infty} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q_n(\theta)) \geq 2(1-w^{\mathrm{j}})^2 > 0 \quad P_0 - a.s,$$

where w^{j} is the weight of δ_{θ_0}. Consider a sequence $\{q_n(\theta)\}$, that converges weakly to a convex combination of $\delta_{\theta_{\mathrm{i}}}, \mathrm{i} \in \{1, 2, \ldots\}$ such that for weights $\{w^{\mathrm{i}} \in (0,1) : \sum_{\mathrm{i}=1}^{\infty} w^{\mathrm{i}} = 1\}$,

$$q_n(\theta) \Rightarrow \sum_{\mathrm{i}=1}^{\infty} w^{\mathrm{i}} \delta_{\theta_{\mathrm{i}}}, \qquad (5.46)$$

where for any $\mathrm{j} \in \{1, 2, \ldots\}$, $\theta_{\mathrm{j}} = \theta_0$ and for all $\mathrm{i} \in \{1, 2, \ldots\} \backslash \{\mathrm{j}\}$, $\theta_{\mathrm{j}} \neq \theta_0$.

Lemma 5.6.5. *The $\alpha-$Rényi divergence between the true posterior and sequence $\{q_n(\theta)\}$ is bounded below by a positive number $2(1-w^{\mathrm{j}})^2$; that is,*

$$\liminf_{n\to\infty} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q_n(\theta)) \geq 2(1-w^{\mathrm{j}})^2 > 0 \quad P_0 - a.s,$$

where w^{j} is the weight of δ_{θ_0} in the definition of sequence $\{q_n(\theta)\}$.

Proof. Van Erven and Harremos [59, Theorem 19] shows that, for any $\alpha > 0$, the $\alpha-$Rényi divergence $D_\alpha(p(\theta)\|q(\theta))$ is a lower semi-continuous function of the pair $(p(\theta), q(\theta))$ in the weak topology on the space of probability measures. Recall from (5.6) that the true posterior distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ converges weakly to δ_{θ_0}, $P_0 - a.s.$ Using this fact it follows that

$$\liminf_{n\to\infty} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q_n(\theta)) \geq D_\alpha\left(\delta_{\theta_0}\middle\|\sum_{\mathrm{i}=1}^{\infty} w_{\mathrm{i}}\delta_{\theta_{\mathrm{i}}}\right) \quad P_0 - a.s.$$

Next, using Pinsker's inequality [133] for $\alpha > 1$, we have

$$\begin{aligned}
D_\alpha\left(\delta_{\theta_0}\middle\|\sum_{\mathrm{i}=1}^{\infty} w^{\mathrm{i}}\delta_{\theta_{\mathrm{i}}}\right) &\geq \frac{1}{2}\left(\int_\Theta \left|\delta_{\theta_0} - \sum_{\mathrm{i}=1}^{\infty} w^{\mathrm{i}}\delta_{\theta_{\mathrm{i}}}\right| d\theta\right)^2 \\
&= \frac{1}{2}\left(\int_\Theta \left|(1-w^{\mathrm{j}})\delta_{\theta_0} - \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}}\delta_{\theta_{\mathrm{i}}}\right| d\theta\right)^2 \\
&= \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)} (1-w^{\mathrm{j}})|\delta_{\theta_0}|d\theta + \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}}\int_{B(\theta_{\mathrm{i}},\epsilon)} |-\delta_{\theta_{\mathrm{i}}}|d\theta\right)^2 \\
&= \frac{1}{2}\left((1-w^{\mathrm{j}}) + \sum_{\mathrm{i}=1,\mathrm{i}\neq\mathrm{j}}^{\infty} w^{\mathrm{i}}\right)^2 = 2(1-w^{\mathrm{j}})^2, \qquad (5.47)
\end{aligned}$$

where $B(\theta_{\mathrm{i}}, \epsilon)$ is the ball of radius ϵ centered at θ_{i}. Note that, there always exist an $\epsilon > 0$, such that $\bigcap_{\mathrm{i}=1}^{\infty} B(\theta_{\mathrm{i}}, \epsilon) = \phi$. Since, by the definition of sequence $\{q_n(\theta)\}$, $w^{\mathrm{j}} \in (0, 1)$, therefore $2(1-w^{\mathrm{j}})^2 > 0$ and the lemma follows. □

Now we show that any sequence of distributions $\{s_n(\theta)\} \subset \mathcal{Q}$ that converges weakly to a distribution $s(\theta) \in \mathcal{Q}$, that has positive density at any point other than the true parameter θ_0, cannot achieve zero KL divergence in the limit.

Proof of Proposition 5.3.1. Observe that for any good sequence $\{\bar{q}_n(\theta)\}$

$$\min_{q\in\mathcal{Q}} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|q(\theta)) \leq D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta)).$$

Therefore, for the second part, it suffices to show that

$$D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta)) < B + o_{P_0^n}(1).$$

The subsequent arguments in the proof are for any $n \geq \max(n_1, n_2, n_3, n_M)$, where n_1, n_2, and n_3 are defined in Assumption 5.2.4. First observe that, for any compact ball K containing the true parameter θ_0,

$$\begin{aligned}
&\frac{\alpha-1}{\alpha} D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta)) \\
&\qquad = \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta + \int_{\Theta\setminus K} \bar{q}_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta\right). \qquad (5.48)
\end{aligned}$$

First, we approximate the first integral on the right hand side using the LAN condition in Assumption 5.2.2. Let $\Delta_{n,\theta_0} := \sqrt{n}(\hat{\theta}_n - \theta_0)$, where $\hat{\theta}_n \to \theta_0$, $P_0 - a.s.$ and Δ_{n,θ_0} converges in distribution to $\mathcal{N}(0, I(\theta_0)^{-1})$. Re-parameterizing the expression with $\theta = \theta_0 + n^{-1/2}h$, we have

$$\begin{aligned}
&\int_K \bar{q}_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta = n^{-1/2}\int_K \bar{q}_n(\theta_0+n^{-1/2}h)\left(\frac{\boldsymbol{\pi}(\theta_0+n^{-1/2}h)\prod_{i=1}^n \frac{p(\xi_i|(\theta_0+n^{-1/2}h))}{p(\xi_i|\theta_0)}}{\bar{q}_n(\theta_0+n^{-1/2}h)\int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\boldsymbol{\pi}(\gamma)d\gamma}\right)^\alpha dh \\
&= n^{-1/2}\int_K \bar{q}_n(\theta_0+n^{-1/2}h)\left(\frac{\boldsymbol{\pi}(\theta_0+n^{-1/2}h)\prod_{i=1}^n \frac{p(\xi_i|(\theta_0+n^{-1/2}h))}{p(\xi_i|\theta_0)}}{\bar{q}_n(\theta_0+n^{-1/2}h)\int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\boldsymbol{\pi}(\gamma)d\gamma}\right)^\alpha dh \\
&= n^{-1/2}\int_K \bar{q}_n(\theta_0+n^{-1/2}h)\left(\boldsymbol{\pi}(\theta_0+n^{-1/2}h)\frac{\exp(hI(\theta_0)\Delta_{n,\theta_0} - \frac{1}{2}h^2 I(\theta_0) + o_{P_0^n}(1))}{\bar{q}_n(\theta_0+n^{-1/2}h)\int_\Theta \prod_{i=1}^n \frac{p(\xi_i|\gamma)}{p(\xi_i|\theta_0)}\boldsymbol{\pi}(\gamma)d\gamma}\right)^\alpha dh. \qquad (5.49)
\end{aligned}$$

Resubstituting $h = \sqrt{n}(\theta - \theta_0)$ in the expression above and reverting to the previous parametrization,

$$= \int_K \bar{q}_n(\theta) \left(\boldsymbol{\pi}(\theta) \frac{\exp\left(\sqrt{n}(\theta-\theta_0) I(\theta_0) \Delta_{n,\theta_0} - \frac{1}{2} n (\theta-\theta_0)^2 I(\theta_0) + o_{P_0^n}(1)\right)}{\bar{q}_n(\theta) \int_\Theta \prod_{\mathrm{i}=1}^n \frac{p(\xi_\mathrm{i}|\gamma)}{p(\xi_\mathrm{i}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma} \right)^\alpha d\theta$$

$$= \int_K \bar{q}_n(\theta) \left(\boldsymbol{\pi}(\theta) \frac{\mathrm{e}^{o_{P_0^n}(1)} \exp\left(-\frac{1}{2} n I(\theta_0) \left((\theta-\theta_0)^2 - 2(\theta-\theta_0)(\hat{\theta}_n - \theta_0) \right)\right)}{\bar{q}_n(\theta) \int_\Theta \prod_{\mathrm{i}=1}^n \frac{p(\xi_\mathrm{i}|\gamma)}{p(\xi_\mathrm{i}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma} \right)^\alpha d\theta.$$

Now completing the square by dividing and multiplying the numerator by $\exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2 \right)\right)$ we obtain

$$= \int_K \bar{q}_n(\theta) \left(\boldsymbol{\pi}(\theta) \frac{\mathrm{e}^{o_{P_0^n}(1)} \exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2 \right)\right) \exp\left(-\frac{1}{2} n I(\theta_0) \left((\theta - \hat{\theta}_n)^2 \right)\right)}{\bar{q}_n(\theta) \int_\Theta \prod_{\mathrm{i}=1}^n \frac{p(\xi_\mathrm{i}|\gamma)}{p(\xi_\mathrm{i}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma} \right)^\alpha d\theta$$

$$= \int_K \bar{q}_n(\theta) \left(\boldsymbol{\pi}(\theta) \frac{\mathrm{e}^{o_{P_0^n}(1)} \exp\left(\frac{1}{2} n I(\theta_0) \left((\hat{\theta}_n - \theta_0)^2 \right)\right) \sqrt{\frac{2\pi}{nI(\theta_0)}} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1})}{\bar{q}_n(\theta) \int_\Theta \prod_{\mathrm{i}=1}^n \frac{p(\xi_\mathrm{i}|\gamma)}{p(\xi_\mathrm{i}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma} \right)^\alpha d\theta, \tag{5.50}$$

where, in the last equality we used the definition of Gaussian density, $\mathcal{N}(\cdot; \hat{\theta}_n, (nI(\theta_0))^{-1})$.

Next, we approximate the integral in the denominator of (5.50). Using Lemma 5.6.4 (in the appendix) it follows that, there exist a sequence of compact balls $\{K_n \subset \Theta\}$, such that $\theta_0 \in K_n$ and

$$\int_\Theta \prod_{\mathrm{i}=1}^n \frac{p(\xi_\mathrm{i}|\gamma)}{p(\xi_\mathrm{i}|\theta_0)} \boldsymbol{\pi}(\gamma) d\gamma$$
$$= \sqrt{\frac{2\pi}{nI(\theta_0)}} \mathrm{e}^{\left(\frac{1}{2} nI(\theta_0)((\hat{\theta}_n - \theta_0)^2)\right)} \left(\mathrm{e}^{o_{P_0^n}(1)} \int_{K_n} \boldsymbol{\pi}(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\gamma + o(1) \right). \tag{5.51}$$

Now, substituting (5.51) into (5.50), we obtain

$$\int_K \bar{q}_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\bar{q}_n(\theta)} \right)^\alpha d\theta = \int_K \bar{q}_n(\theta)^{1-\alpha} \left(\frac{\mathrm{e}^{o_{P_0^n}(1)} \boldsymbol{\pi}(\theta) \mathcal{N}(\theta; \hat{\theta}_n, \frac{1}{nI(\theta_0)})}{\left(\mathrm{e}^{o_{P_0^n}(1)} \int_{K_n} \boldsymbol{\pi}(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, \frac{1}{nI(\theta_0)}) d\gamma + o(1) \right)} \right)^\alpha d\theta. \tag{5.52}$$

Now, recall the definition of compact ball K, n_1 and n_2 from Assumption 5.2.4 and fix $n \geq n_0$, where $n_0 = \max(n_1, n_2)$. Note that n_2 is chosen, such that for all $n \geq n_2$, the bound in Assumption 5.2.4(3) holds on the set $\Theta \backslash K$. Next, consider the second term inside the logarithm function on the right hand side of (5.48). Using Assumption 5.2.4(3), we obtain

$$\int_{\Theta \backslash K} \bar{q}_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\bar{q}_n(\theta)} \right)^{\alpha} d\theta \leq M_r^{\alpha} \int_{\Theta \backslash K} \bar{q}_n(\theta) d\theta \quad P_0 - a.s. \tag{5.53}$$

Recall that the good sequence $\{\bar{q}_n(\cdot)\}$ exists $P_0 - a.s$ with mean $\hat{\theta}_n$, for all $n \geq n_1$ and therefore it converges weakly to δ_{θ_0} (as assumed in Assumption 5.2.4(2)). Combined with the fact that compact set K contains the true parameter θ_0, it follows that the second term in (5.48) is of $o(1)$, $P_0 - a.s.$ Therefore, the second term inside the logarithm function on the right hand side of (5.48) is $o(1)$:

$$\int_{\Theta \backslash K} \bar{q}_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{\bar{q}_n(\theta)} \right)^{\alpha} d\theta = o(1) \; P_0 - \text{a.s.} \tag{5.54}$$

Substituting (5.52) and (5.54) into (5.48), we have

$$\begin{aligned}
&\frac{\alpha - 1}{\alpha} D_{\alpha}(\boldsymbol{\pi}(\theta|\tilde{X}_n) \| \bar{q}_n(\theta)) \\
&\quad = \frac{1}{\alpha} \log \left(\int_K \bar{q}_n(\theta)^{1-\alpha} \left(\frac{\mathrm{e}^{o_{P_0^n}(1)} \boldsymbol{\pi}(\theta) \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1})}{\left(\mathrm{e}^{o_{P_0^n}(1)} \int_{K_n} \boldsymbol{\pi}(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\gamma + o(1) \right)} \right)^{\alpha} d\theta + o(1) \right) \\
&\quad = \frac{1}{\alpha} \log \left(\mathrm{e}^{o_{P_0^n}(1)} \int_K \bar{q}_n(\theta)^{1-\alpha} \left(\frac{\boldsymbol{\pi}(\theta) \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1})}{\left(\mathrm{e}^{o_{P_0^n}(1)} \int_{K_n} \boldsymbol{\pi}(\gamma) \mathcal{N}(\gamma; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\gamma + o(1) \right)} \right)^{\alpha} d\theta + o(1) \right). \qquad (\star\star)
\end{aligned}$$

Now observe that,

$$
\begin{aligned}
(\star\star) &\sim \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)^{1-\alpha}\left(\frac{\boldsymbol{\pi}(\theta)\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})}{\left(\mathrm{e}^{o_{P_0^n}(1)}\int_{K_n}\boldsymbol{\pi}(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)}\right)^{\alpha}d\theta\right)\\
&= \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)^{1-\alpha}\boldsymbol{\pi}(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})^{\alpha}d\theta\right)\\
&\quad-\log\left(\mathrm{e}^{o_{P_0^n}(1)}\int_{K_n}\boldsymbol{\pi}(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma+o(1)\right)\\
&\sim \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)^{1-\alpha}\boldsymbol{\pi}(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})^{\alpha}d\theta\right)\\
&\quad-\log\left(\int_{K_n}\boldsymbol{\pi}(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma\right)+o_{P_0^n}(1). \qquad (5.55)
\end{aligned}
$$

Note that $\left(\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})\right)^{\alpha}=\left(\sqrt{\frac{nI(\theta_0)}{2\pi}}\right)^{\alpha}\left(\sqrt{\frac{2\pi}{n\alpha I(\theta_0)}}\right)\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})$.
Substituting this into (5.55), for large enough n, we have

$$
\begin{aligned}
&\frac{\alpha-1}{\alpha}D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta))\\
&\quad\sim\frac{\alpha-1}{2\alpha}\log n-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\pi}+\frac{1}{\alpha}\log\int_K\bar{q}_n(\theta)^{1-\alpha}\boldsymbol{\pi}(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})d\theta\\
&\quad-\log\left(\int_{K_n}\boldsymbol{\pi}(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma\right). \qquad (5.56)
\end{aligned}
$$

From the Laplace approximation (Lemma 5.6.1) and the continuity of the logarithm, we have

$$
\frac{1}{\alpha}\log\int_K\bar{q}_n(\theta)^{1-\alpha}\boldsymbol{\pi}(\theta)^{\alpha}\mathcal{N}(\theta;\hat{\theta}_n,(n\alpha I(\theta_0))^{-1})d\theta\sim\frac{1-\alpha}{\alpha}\log\bar{q}_n(\hat{\theta}_n)+\log\boldsymbol{\pi}(\hat{\theta}_n).
$$

Next, using the Laplace approximation (Lemma 5.6.1) on the last term in (5.56) yields

$$
-\log\left(\int_{K_n}\boldsymbol{\pi}(\gamma)\mathcal{N}(\gamma;\hat{\theta}_n,(nI(\theta_0))^{-1})d\gamma\right)\sim-\log\left(\boldsymbol{\pi}(\hat{\theta}_n)\right).
$$

Substituting the above two approximations into (5.56), for large enough n, we obtain

$$\begin{aligned}
&\frac{\alpha-1}{\alpha}D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta)) \\
&\quad\sim\frac{1-\alpha}{\alpha}\log\bar{q}_n(\hat{\theta}_n)+\log\boldsymbol{\pi}(\hat{\theta}_n)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\boldsymbol{\pi}}+\frac{\alpha-1}{2\alpha}\log n-\log\boldsymbol{\pi}(\hat{\theta}_n)+o_{P_0^n}(1) \\
&\quad=\frac{1-\alpha}{\alpha}\log\bar{q}_n(\hat{\theta}_n)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\boldsymbol{\pi}}+\frac{\alpha-1}{2\alpha}\log n+o_{P_0^n}(1). \qquad (5.57)
\end{aligned}$$

Now, recall Assumption 5.2.4(4) which, combined with the monotonicity of logarithm function, implies that $\log\bar{q}_n(\cdot)$ is concave for all $n\geq n_3$. Using Jensen's inequality,

$$\log\bar{q}_n(\hat{\theta}_n)=\log\bar{q}_n\left(\int\theta\bar{q}_n(\theta)d\theta\right)\geq\int\bar{q}_n(\theta)\log\bar{q}_n(\theta)d\theta.$$

Since $\alpha>1$,

$$\frac{1-\alpha}{\alpha}\log\bar{q}_n(\hat{\theta}_n)\leq-\frac{\alpha-1}{\alpha}\int\bar{q}_n(\theta)\log\bar{q}_n(\theta)d\theta.$$

Now using Lemma 5.6.2 (in the appendix), there exists $n_M\geq 1$ and $0<\bar{M}<\infty$, such that for all $n\geq n_M$

$$-\frac{\alpha-1}{\alpha}\int\bar{q}_n(\theta)\log\bar{q}_n(\theta)d\theta\leq\frac{\alpha-1}{2\alpha}\log\left(2\boldsymbol{\pi}\bar{\mathrm{e}}\frac{\bar{M}}{n}\right)=\frac{\alpha-1}{2\alpha}\log(2\boldsymbol{\pi}\bar{\mathrm{e}}\bar{M})-\frac{\alpha-1}{2\alpha}\log n, \qquad (5.58)$$

where $\bar{\mathrm{e}}$ is the Euler's constant. Substituting (5.58) into the right hand side of (5.57), we have for all $n\geq n_0$, where $n_0=\max(n_0,n_3,n_M)$,

$$\begin{aligned}
&\frac{1-\alpha}{\alpha}\log\bar{q}_n(\hat{\theta}_n)-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\boldsymbol{\pi}}+\frac{\alpha-1}{2\alpha}\log n. \\
&\quad\leq\frac{\alpha-1}{2\alpha}\log(2\boldsymbol{\pi}\bar{\mathrm{e}}\bar{M})-\frac{\alpha-1}{2\alpha}\log n-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\boldsymbol{\pi}}+\frac{\alpha-1}{2\alpha}\log n \\
&\quad=\frac{\alpha-1}{2\alpha}\log(2\boldsymbol{\pi}\bar{\mathrm{e}}\bar{M})-\frac{\log\alpha}{2\alpha}+\frac{\alpha-1}{2\alpha}\log\frac{I(\theta_0)}{2\boldsymbol{\pi}} \\
&\quad=\frac{\alpha-1}{\alpha}\frac{1}{2}\log\frac{\bar{\mathrm{e}}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}}. \qquad (5.59)
\end{aligned}$$

Observe that the left hand side in (5.57) is always non-negative, implying the right hand side must be too for large n. Therefore, the following inequality must hold for all $n \geq n_0$:

$$\frac{\bar{\mathrm{e}}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}} \geq 1.$$

Consequently, substituting (5.59) into (5.57), we have

$$D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|\bar{q}_n(\theta)) \leq \frac{1}{2}\log\frac{\bar{\mathrm{e}}\bar{M}I(\theta_0)}{\alpha^{\frac{1}{\alpha-1}}} + o_{P_0^n}(1)\ \forall n \geq n_0, \tag{5.60}$$

and the result follows.

□

Proof of Lemma 5.4.1. Posner [112, Theorem 1] shows that, the KL divergence $\mathrm{KL}(p(\theta)\|s(\theta))$ is a lower semi-continuous function of the pair $(p(\theta), s(\theta))$ in the weak topology on the space of probability measures. Recall from (5.6) that the true posterior distribution $\boldsymbol{\pi}(\theta|\tilde{X}_n)$ converges weakly to δ_{θ_0}, $P_0 - a.s.$ Using this fact it follows that

$$\liminf_{n\to\infty} \mathrm{KL}(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|s_n(\theta)) \geq \mathrm{KL}\left(\delta_{\theta_0}\|s(\theta)\right) \quad P_0 - a.s.$$

Next, using Pinsker's inequality [133] for $\alpha > 1$, we have

$$\mathrm{KL}\left(\delta_{\theta_0}\|s(\theta)\right) \geq \frac{1}{2}\left(\int_\Theta |\delta_{\theta_0} - s(\theta)|\, d\theta\right)^2.$$

Now, fixing $\epsilon > 0$ such that $s(\theta)$ has positive density in the complement of the ball of radius ϵ centered at θ_0, $B(\theta_0, \epsilon)^C$, we have

$$\begin{aligned}
\liminf_{n\to\infty} \mathrm{KL}(\boldsymbol{\pi}(\theta|\tilde{X}_n)\|s_n(\theta)) &\geq \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)} |\delta_{\theta_0} - s(\theta)|\, d\theta + \int_{B(\theta_0,\epsilon)^C} |\delta_{\theta_0} - s(\theta)|\, d\theta\right)^2 \\
&\geq \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C} |\delta_{\theta_0} - s(\theta)|\, d\theta\right)^2 \\
&= \frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C} |-s(\theta)|\, d\theta\right)^2 \quad P_0 - a.s.
\end{aligned} \tag{5.61}$$

Since $s(\theta)$ has positive density in the set $B(\theta_0, \epsilon)^C$, there exists $\eta(\epsilon) > 0$, such that

$$\frac{1}{2}\left(\int_{B(\theta_0,\epsilon)^C} |-s(\theta)|\, d\theta\right)^2 \geq \eta(\epsilon),$$

completing the proof. □

Next, we state an important inequality, that is a direct consequence of Hölder's inequality. We use the following result in the proof of Theorem 5.3.3.

Lemma 5.6.6. *For any set $K \subset \Theta$ and $\alpha > 1$ and any sequence of distributions $\{q_n(\theta)\} \subset \mathcal{Q}$, the following inequality holds true*

$$\int_\Theta q_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q_n(\theta)}\right)^\alpha d\theta \geq \frac{\left(\int_K \boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta\right)^\alpha}{\left(\int_K q_n(\theta)d\theta\right)^{\alpha-1}}. \tag{5.62}$$

Proof. Fix a set $K \subset \Theta$. Since $\alpha > 1$, using Hölder's inequality for $f(\theta) = \frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q_n(\theta)^{1-\frac{1}{\alpha}}}$ and $g(\theta) = q_n(\theta)^{1-\frac{1}{\alpha}}$,

$$\begin{aligned}\int_K \boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta &= \int_K f(\theta)g(\theta)d\theta \\ &\leq \left(\int_K \frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)^\alpha}{q_n(\theta)^{\alpha-1}}d\theta\right)^{\frac{1}{\alpha}}\left(\int_K q_n(\theta)d\theta\right)^{1-\frac{1}{\alpha}}.\end{aligned}$$

It is straightforward to observe from the above equation that,

$$\int_K \frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)^\alpha}{q_n(\theta)^{\alpha-1}}d\theta \geq \frac{\left(\int_K \boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta\right)^\alpha}{\left(\int_K q_n(\theta)d\theta\right)^{\alpha-1}}.$$

Also note that, for any set K, the following inequality holds true,

$$\int_\Theta q_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q_n(\theta)}\right)^\alpha d\theta \geq \int_K \frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)^\alpha}{q_n(\theta)^{\alpha-1}}d\theta \geq \frac{\left(\int_K \boldsymbol{\pi}(\theta|\tilde{X}_n)d\theta\right)^\alpha}{\left(\int_K q_n(\theta)d\theta\right)^{\alpha-1}}, \tag{5.63}$$

and the result follows immediately. □

Proof of Lemma 5.3.3. First, we fix $n \geq 1$ and let M_r be a sequence such that $M_r \to \infty$ as $r \to \infty$. Recall that $\hat{\theta}_n$ is the maximum likelihood estimate and denote $\tilde{\theta}_n = \mathbb{E}_{q_n(\theta)}[\theta]$. Define a set

$$K_r := \{\theta \in \Theta : |\theta - \hat{\theta}_n| > M_r\} \bigcup \{\theta \in \Theta : |\theta - \tilde{\theta}_n| > M_r\}.$$

Now, using Lemma 5.6.6 with $K = K_r$, we have

$$\int_\Theta q_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n)}{q_n(\theta)} \right)^\alpha d\theta \geq \frac{\left(\int_{K_r} \boldsymbol{\pi}(\theta|\tilde{X}_n) d\theta \right)^\alpha}{\left(\int_{K_r} q_n(\theta) d\theta \right)^{\alpha - 1}}. \tag{5.64}$$

Note that the left hand side in the above equation does not depend on r and when $r \to \infty$ both the numerator and denominator on the right hand side converges to zero individually. For the ratio to diverge, however, we require the denominator to converge much faster than the numerator. To be more precise, observe that for a given n, since $\alpha - 1 < \alpha$ the tails of $q_n(\theta)$ must decay significantly faster than the tails of the true posterior for the right hand side in (5.64) to diverge as $r \to \infty$.

We next show that there exists an $n_0 \geq 1$ such that for all $n \geq n_0$, the right hand side in (5.64) diverges as $r \to \infty$. Since the posterior distribution satisfies the Bernstein-von Mises Theorem [109], we have

$$\int_{K_r} \boldsymbol{\pi}(\theta|\tilde{X}_n) d\theta = \int_{K_r} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta + o_{P_0^n}(1).$$

Observe that the numerator on the right hand side of (5.64) satisfies,

$$\begin{aligned}
\left(\int_{K_r} \boldsymbol{\pi}(\theta|\tilde{X}_n) d\theta \right)^\alpha &= \left(\int_{K_r} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta + o_{P_0^n}(1) \right)^\alpha \\
&\geq \left(\int_{\{|\theta - \hat{\theta}_n| > M_r\}} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta + o_{P_0^n}(1) \right)^\alpha \\
&= \left(\int_{\{\theta - \hat{\theta}_n > M_r\}} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta + \int_{\{\theta - \hat{\theta}_n \leq -M_r\}} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta + o_{P_0^n}(1) \right)^\alpha \\
&\geq \left(\int_{\{\theta - \hat{\theta}_n > M_r\}} \mathcal{N}(\theta; \hat{\theta}_n, (nI(\theta_0))^{-1}) d\theta + o_{P_0^n}(1) \right)^\alpha .
\end{aligned} \tag{5.65}$$

Now, using the lower bound on the Gaussian tail distributions from [134]

$$\begin{aligned}\left(\int_{K_r}\pi(\theta|\tilde{X}_n)d\theta\right)^{\alpha} &= \left(\int_{K_r}\mathcal{N}(\theta;\hat{\theta}_n,(nI(\theta_0))^{-1})d\theta + o_{P_0^n}(1)\right)^{\alpha}\\ &\geq \left(\frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sqrt{nI(\theta_0)}M_r} - \frac{1}{(\sqrt{nI(\theta_0)}M_r)^3}\right)\mathrm{e}^{-\frac{nI(\theta_0)}{2}M_r^2} + o_{P_0^n}(1)\right)^{\alpha}\\ &\sim \left(\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{nI(\theta_0)}M_r}\mathrm{e}^{-\frac{nI(\theta_0)}{2}M_r^2} + o_{P_0^n}(1)\right)^{\alpha}, \end{aligned} \tag{5.66}$$

where the last approximation follows from the fact that, for large r,

$$\left(\frac{1}{\sqrt{nI(\theta_0)}M_r} - \frac{1}{(\sqrt{nI(\theta_0)}M_r)^3}\right) \sim \frac{1}{\sqrt{nI(\theta_0)}M_r}.$$

Next, consider the denominator on the right hand side of (5.64). Using the union bound

$$\left(\int_{K_r} q_n(\theta)d\theta\right)^{\alpha-1} \leq \left(\int_{\{|\theta-\tilde{\theta}_n|>M_r\}} q_n(\theta)d\theta + \int_{\{|\theta-\hat{\theta}_n|>M_r\}} q_n(\theta)d\theta\right)^{\alpha-1}. \tag{5.67}$$

Since, $\tilde{\theta}_n$ and $\hat{\theta}_n$ are finite for all $n \geq 1$, there exists an $\epsilon > 0$ such that for large n, $|\tilde{\theta}_n - \hat{\theta}_n| \leq \epsilon$. Applying the triangle inequality,

$$|\theta - \hat{\theta}_n| \leq |\theta - \tilde{\theta}_n| + |\tilde{\theta}_n - \hat{\theta}_n| \leq |\theta - \tilde{\theta}_n| + \epsilon.$$

Therefore, $\{|\theta - \hat{\theta}_n| > M_r\} \subseteq \{|\theta - \tilde{\theta}_n| > M_r - \epsilon\}$ and it follows from (5.67) that

$$\left(\int_{K_r} q_n(\theta)d\theta\right)^{\alpha-1} \leq \left(\int_{\{|\theta-\tilde{\theta}_n|>M_r\}} q_n(\theta)d\theta + \int_{\{|\theta-\tilde{\theta}_n|>M_r-\epsilon\}} q_n(\theta)d\theta\right)^{\alpha-1}.$$

Next, using the sub-Gaussian tail distribution bound from [135, Theorem 2.1], we have

$$\left(\int_{\{|\theta-\tilde{\theta}_n|>M_r\}} q_n(\theta)d\theta + \int_{\{|\theta-\tilde{\theta}_n|>M_r-\epsilon\}} q_n(\theta)d\theta\right)^{\alpha-1} \leq \left(2\mathrm{e}^{-\frac{\gamma_n^2 M_r^2}{2B}} + 2\mathrm{e}^{-\frac{\gamma_n^2(M_r-\epsilon)^2}{2B}}\right)^{\alpha-1}. \tag{5.68}$$

For large r, $M_r \sim M_r - \epsilon$, and it follows that

$$\left(\int_{\{|\theta-\tilde{\theta}_n|>M_r\}} q_n(\theta)d\theta + \int_{\{|\theta-\tilde{\theta}_n|>M_r-\epsilon\}} q_n(\theta)d\theta\right)^{\alpha-1} \lesssim \left(4\mathrm{e}^{-\frac{\gamma_n^2 M_r^2}{2B}}\right)^{\alpha-1}. \tag{5.69}$$

Substituting (5.66) and (5.69) into (5.64), we obtain

$$\int_\Theta q_n(\theta)\left(\frac{\pi(\theta|\tilde{X}_n)}{q_n(\theta)}\right)^\alpha d\theta \gtrsim \left(\frac{\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{nI(\theta_0)}M_r}\mathrm{e}^{-\frac{nI(\theta_0)}{2}M_r^2} + o_{P_0^n}(1)}{\left(4\mathrm{e}^{-\frac{\gamma_n^2 M_r^2}{2B}}\right)^{\frac{\alpha-1}{\alpha}}}\right)^\alpha,$$

for large r. Observe that

$$\frac{\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{nI(\theta_0)}M_r}\mathrm{e}^{-\frac{nI(\theta_0)}{2}M_r^2}}{\left(4\mathrm{e}^{-\frac{\gamma_n^2 M_r^2}{2B}}\right)^{\frac{\alpha-1}{\alpha}}} = \frac{1}{4^{\frac{\alpha-1}{\alpha}}\sqrt{2\pi}}\frac{1}{M_r}\left(\frac{1}{\sqrt{nI(\theta_0)}}\mathrm{e}^{M_r^2\left(\frac{\alpha-1}{\alpha}\frac{\gamma_n^2}{2B}-\frac{nI(\theta_0)}{2}\right)}\right). \tag{5.70}$$

Since $\gamma_n^2 > n$, choosing $n_0 = \min\left\{n : \left(\frac{\alpha-1}{\alpha}\frac{\gamma_n^2}{2B} - \frac{nI(\theta_0)}{2}\right) > 0\right\}$ implies that for all $n \geq n_0$, as $r \to \infty$, the left hand side in (5.70) diverges and the result follows.

□

Proof of Lemma 5.5.1. We prove the assertion of the Lemma for the class of local latent parameters z_{i} that have discrete and finite support. First observe that for $\alpha > 1$, using Jensen's inequality

$$M(\tilde{X}_n|\theta)^\alpha = \min_{q(z_{1:n})\in\mathcal{Q}^n}\int_{\mathcal{Z}^n} q(z_{1:n})\left(\frac{p(z_{1:n},\tilde{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n} \geq \left[\int_{\mathcal{Z}^n} p(z_{1:n},\tilde{X}_n|\theta)dz_{1:n}\right]^\alpha. \tag{5.71}$$

Now since family $\mathcal{Q}^n$ contains point masses, we choose a member of family $\mathcal{Q}^n$ which is a joint distribution of point masses at $z_{1:n}^p := \{z_1^p, z_2^p, \ldots, z_n^p\}$ to obtain

$$M(\tilde{X}_n|\theta)^\alpha = \min_{q(z_{1:n})\in\mathcal{Q}^n}\int_{\mathcal{Z}^n} q(z_{1:n})\left(\frac{p(z_{1:n},\tilde{X}_n|\theta)}{q(z_{1:n})}\right)^\alpha dz_{1:n} \leq \left[p(z_{1:n}^p,\tilde{X}_n|\theta)\right]^\alpha, \tag{5.72}$$

where $z_{1:n}^p$ is as defined in Assumption 5.5.1.

Since, $f(x) = x^\alpha$ is increasing for $\alpha > 1$ and $x > 0$, it follows from (5.71), (5.72), and monotonicity of the logarithm function that

$$\log \int_{\mathcal{Z}^n} p(z_{1:n}, \tilde{X}_n|\theta) dz_{1:n} \leq \log M(\tilde{X}_n|\theta) \leq \log p(z^p_{1:n}, \tilde{X}_n|\theta). \tag{5.73}$$

Now using Assumption 5.5.1 (1) and (2(ii)), that is $d_H(z_0, z^p_{1:n}) = o(\rho_n)$, it follows that at some rate ρ_n with $\rho_n \downarrow 0$ and $n\rho_n^2 \to \infty$; that is for all bounded, stochastic $h_n = O_{P_0^n}(1)$,

$$\begin{aligned}
&\int_{\{z_{1:n}: d_H(z_{1:n}, z_0) \geq \rho_n\}} p(z_{1:n}|\tilde{X}_n, \theta = \theta_0 + n^{-1/2}h_n) dz_{1:n} \\
&\qquad \leq \int_{\{z_{1:n}: d_H(z_{1:n}, z^p_{1:n}) + d_H(z_0, z^p_{1:n}) \geq \rho_n\}} p(z_{1:n}|\tilde{X}_n, \theta = \theta_0 + n^{-1/2}h_n) dz_{1:n} \\
&\qquad \leq \int_{\{z_{1:n}: d_H(z_{1:n}, z^p_{1:n}) \geq \rho_n(1-\epsilon)\}} p(z_{1:n}|\tilde{X}_n, \theta = \theta_0 + n^{-1/2}h_n) dz_{1:n} = O_{P_0^n}(1),
\end{aligned}$$

where the first inequality follows from using the fact that $d_H(z_{1:n}, z_0) \leq d_H(z_{1:n}, z^p_{1:n}) + d_H(z_0, z^p_{1:n})$, the second inequality uses the fact that $d_H(z_0, z^p_{1:n}) = o(\rho_n)$, that is for some $\epsilon \in (0, 1)$, $d_H(z_0, z^p_{1:n}) < \epsilon\rho_n$ for sufficiently large n, and the last inequality is due to Assumption 5.5.1 (1).

Therefore, it can be observed from the above result that the conditioned latent posterior $p(z_{1:n}|\tilde{X}_n, \theta_0)$ concentrates at z_0. Consequently, when the local latent parameters are discrete it follows that

$$\log \int_{\mathcal{Z}^n} p(z_{1:n}, \tilde{X}_n|\theta_0) dz_{1:n} = \log \int_{\mathcal{Z}^n} \frac{p(z_{1:n}|\tilde{X}_n, \theta_0)}{p(z_{1:n}|\tilde{X}_n, \theta_0)} p(z_{1:n}, \tilde{X}_n|\theta_0) dz_{1:n} = \log p(z_0, \tilde{X}_n|\theta_0) + o_{P_0^n}(1).$$

Now it follows that

$$\log M(\tilde{X}_n|\theta_0) = \log p(z_0, \tilde{X}_n|\theta_0) + o_{P_0^n}(1) = \log \int_{\mathcal{Z}^n} p(z_{1:n}, \tilde{X}_n|\theta_0) dz_{1:n} + o_{P_0^n}(1). \tag{5.74}$$

Subtracting $\log M(\tilde{X}_n|\theta_0)$ from (5.73) and using (5.74) yields

$$\log \frac{\int_{\mathcal{Z}^n} p(z_{1:n}, \tilde{X}_n|\theta) dz_{1:n}}{\int_{\mathcal{Z}^n} p(z_{1:n}, \tilde{X}_n|\theta_0) dz_{1:n}} + o_{P_0^n}(1) \leq \log \frac{M(\tilde{X}_n|\theta)}{M(\tilde{X}_n|\theta_0)} \leq \log \frac{p(z_0, \tilde{X}_n|\theta)}{p(z_0, \tilde{X}_n|\theta_0)} + o_{P_0^n}(1). \tag{5.75}$$

Now, substituting $\theta = \theta_0 + n^{-1/2}h_n$ for all bounded and stochastic $h_n = O_{P_0^n}(1)$, and using the result in Bickel, Kleijn, *et al.* [128, Theorem 4.2] under the conditions in Assumption 5.5.1 the RHS and LHS above have the same LAN expansion and the result follows. Notice that, by definition, the s-LAN condition in Assumption 5.2.2 is also true at $z_{1:n} = z^p_{1:n}$. Assumption 5.5.1 (2(ii)) implies $d_H(z_0, z^p_{1:n}) = o(\rho_n)$ with $\rho_n \downarrow 0$ and $n\rho_n^2 \to \infty$, so that

$$\log \frac{P^n_{\theta_0+n^{-1/2}h_n, z^p_{1:n}}}{P^n_{\theta_0, z^p_{1:n}}} = \log \frac{P^n_{\theta_0+n^{-1/2}h_n, z_0}}{P^n_{\theta_0, z_0}} + o(1).$$

Therefore, $\log \frac{p(z_0, \tilde{X}_n|\theta_0+n^{-1/2}h_n)}{p(z_0,\tilde{X}_n|\theta_0)} = \log \frac{p(\tilde{X}_n|z_0,\theta_0+n^{-1/2}h_n)}{p(\tilde{X}_n|z_0,\theta_0)} + \log \frac{p(z_0|\theta_0+n^{-1/2}h_n)}{p(z_0|\theta_0)} = \log \frac{P^n_{\theta_0+n^{-1/2}h_n,z_0}}{P^n_{\theta_0,z_0}} + o(1)$ also have the same expansion as given in the s-LAN condition in Assumption 5.2.2. □

Proof of Proposition 5.5.1. Observe that for any good sequence $\{\bar{q}_n(\theta)\}$ and $q(z_{1:n})$ as point masses (discrete distribution) at the truth $z^0_{1:n} := \{z^0_1, z^0_2, \ldots, z^0_n\}$, we have

$$\begin{aligned}
&\min_{q\in\mathcal{Q}} \min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)\|q(\theta)q(z_{1:n})) \\
&\qquad = \min_{q(\theta)\in\bar{\mathcal{Q}}, q(z_{1:n})\in\mathcal{Q}^n} \frac{1}{\alpha-1}\log \int_{\Theta\times\mathcal{Z}^n} q(\theta)q(z_{1:n}) \left(\frac{p(\theta, z_{1:n}, \tilde{X}_n)}{p(\tilde{X}_n)q(\theta)q(z_{1:n})}\right)^\alpha d\theta dz_{1:n} \\
&\qquad \leq \frac{1}{\alpha-1}\log \int_\Theta \bar{q}_n(\theta) \left(\frac{p(\theta, z^0_{1:n}, \tilde{X}_n)}{p(\tilde{X}_n)\bar{q}_n(\theta)}\right)^\alpha d\theta \\
&\qquad \leq \frac{1}{\alpha-1}\log \int_\Theta \bar{q}_n(\theta) \left(\frac{\boldsymbol{\pi}(\theta, z^0_{1:n}|\tilde{X}_n)}{\bar{q}_n(\theta)}\right)^\alpha d\theta. \qquad (5.76)
\end{aligned}$$

Also note that, using the definition of $\boldsymbol{\pi}(\theta, z^0_{1:n}|\tilde{X}_n)$, we have

$$\boldsymbol{\pi}(\theta, z^0_{1:n}|\tilde{X}_n) = \frac{\boldsymbol{\pi}(\theta)\boldsymbol{\pi}(z^0_{1:n}|\theta)p(\tilde{X}_n|\theta, z^0_{1:n})}{\int_{\Theta\times\mathcal{Z}^n} \boldsymbol{\pi}(\theta)\boldsymbol{\pi}(z_{1:n}|\theta)p(\tilde{X}_n|\theta, z_{1:n})d\theta dz_{1:n}} \leq \frac{\boldsymbol{\pi}(\theta)\boldsymbol{\pi}(z^0_{1:n}|\theta)p(\tilde{X}_n|\theta, z^0_{1:n})}{\int_\Theta \boldsymbol{\pi}(\theta)\boldsymbol{\pi}(z^0_{1:n}|\theta)p(\tilde{X}_n|\theta, z^0_{1:n})d\theta}, \qquad (5.77)$$

where the second inequality follows from the fact that $z_{1:n}$ is a discrete random variable. Therefore substituting (5.77) into (5.76) yields

$$\begin{aligned}\min_{q\in\mathcal{Q}}\min_{q(z_{1:n})\in\mathcal{Q}^n} D_\alpha(\boldsymbol{\pi}(\theta, z_{1:n}|\tilde{X}_n)\|q(\theta)q(z_{1:n})) &\leq \frac{1}{\alpha-1}\log\int_\Theta \bar{q}_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta)p(\tilde{X}_n, z^0_{1:n}|\theta)}{\bar{q}_n(\theta)\int_\Theta \boldsymbol{\pi}(\theta)p(\tilde{X}_n, z^0_{1:n}|\theta)d\theta}\right)^\alpha d\theta \\ &= \frac{1}{\alpha-1}\log\int_\Theta \bar{q}_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n, z^0_{1:n})}{\bar{q}_n(\theta)}\right)^\alpha d\theta \\ &=: D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n, z^0_{1:n})\|\bar{q}_n(\theta)). \end{aligned} \tag{5.78}$$

Therefore, for the second part, it suffices to show that

$$D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n, z^0_{1:n})\|\bar{q}_n(\theta)) < B + o_{P_0^n}(1).$$

The subsequent arguments in the proof are for any $n \geq \max(n_1, n_2, n_3, n_M)$, where n_1, n_2, and n_3 are defined in Assumption 5.2.4. First observe that, for any compact ball K containing the true parameter θ_0,

$$\begin{aligned}&\frac{\alpha-1}{\alpha}D_\alpha(\boldsymbol{\pi}(\theta|\tilde{X}_n, z^0_{1:n})\|\bar{q}_n(\theta)) \\ &\quad= \frac{1}{\alpha}\log\left(\int_K \bar{q}_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n, z^0_{1:n})}{\bar{q}_n(\theta)}\right)^\alpha d\theta + \int_{\Theta\setminus K}\bar{q}_n(\theta)\left(\frac{\boldsymbol{\pi}(\theta|\tilde{X}_n, z^0_{1:n})}{\bar{q}_n(\theta)}\right)^\alpha d\theta\right). \end{aligned} \tag{5.79}$$

First, we approximate the first integral on the right hand side using the LAN condition in Assumption 5.2.2. Let $\Delta_{n,(\theta_0,z_0)} := \sqrt{n}(\hat{\theta}_n - \theta_0)$, where $\hat{\theta}_n \to \theta_0$, $P_0 - a.s.$ and $\Delta_{n,(\theta_0,z_0)}$ converges in distribution to $\mathcal{N}(0, I(\theta_0, z_0)^{-1})$ [109, Lemma 25.23 and 25.25]. Now the proof follows similar steps as used in the proof of Proposition 5.3.1. □

6. CONCLUSION

Data-driven decision-making has received significant research interest in the recent literature, in particular since the nature of the interplay between data and optimal decision-making can be quite different from the standard machine learning setting. While much of the literature focuses on empirical methods, Bayesian approaches afford advantages, particularly when making decisions in the context of stochastic models. However, Bayesian methods also suffer from an issue of posterior intractability, which is hard to resolve in practice.

This thesis proposed computationally tractable Bayesian methodologies to approximate stochastic programs (SP) with deterministic and epistemically uncertain constraints. We first proposed a novel VI framework for risk-sensitive data-driven decision-making in Chapter 2, which we call risk-sensitive variational Bayes (RSVB) to approximate SP with deterministic constraints. Thereafter, we introduced the Bayesian joint chance-constrained stochastic program (BJCCP) for modeling decision-making problems with epistemically uncertain constraints and its VB approximation (VBJCCP) in Chapter 4. Broadly, such methodologies can be theoretically studied under two categories: 1) statistical (accuracy) and 2) computational (speed). Statistical properties such as asymptotic consistency and convergence rates provide theoretical guarantees on learning the truth, given an infinite (large) amount of data. On the other hand, evaluating the computational performance is more towards understanding their algorithmic efficiency both in terms of the number of data points used and computational time (or steps) required to optimize the risk/loss of quantifying the deviation from the truth or taking a sub-optimal decision. In this thesis, we mainly focus on establishing the statistical performance of the proposed methods. The work in this thesis can be extended to several directions as part of the future works.

First, an obvious set of open problems is to develop computational algorithms to solve the minimax optimization in RSVB and the chance-constrained optimization in VBJCCP efficiently and study its computational complexity for a given number of samples. It would also be interesting to establish their theoretical properties to understand trade-offs between statistical accuracy and computational complexity. Second, recall from Chapter 5 that using KL divergence in the VB framework tends to produce an 'overconfident' approximate poste-

rior that underestimates the tails of the posterior, therefore it would be useful to extend the theoretical results in the first three chapters of this thesis to divergence measures other than the KL divergence. For instance, the α-Rényi divergence [26], [136] has been demonstrated to provide better support coverage. Third, recall that our rate of convergence results in Chapters 2 and 4 only hold for a large enough sample size. A sample complexity result is significantly harder but can be immensely useful for applications where large datasets are hard to collect (healthcare, for instance).

Furthermore, also note that the KL optimization problem in the VB method could be a non-convex program either in measure-space or the parameterized case, and therefore obtaining a global solution is difficult. On the other hand, to the best of our knowledge, all the extant statistical inferential works establishing large sample properties of the VB optimizer implicitly assume that the global optimizer is computable. Since in practice finding global optima is difficult, it is an important problem to study the theoretical properties of the local VB optimizer. In a similar vein, this problem arises in the RSVB objective too, where we implicitly assumed that the inner optimization can be solved globally, which is not true in general. Studying the statistical performance of the RSVB approach, while relaxing this implicit assumption, would be an important analytical contribution.

Recently, [137] proposed likelihood-free variational inference, a VB algorithm to incorporate implicit probabilistic (likelihood) models which are defined using a simulation process. These implicit probabilistic models usually represent some real-world physical systems and are so rich in their structure to be represented by a tractable likelihood function. To the best of our knowledge, all the extant theoretical work on the analysis of VB methods assumes that the likelihood form is known *a priori*. Finally, extending the work in this thesis to understand the inferential properties of such likelihood-free variational Bayesian algorithms would be a significant contribution.

REFERENCES

[1] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[2] A. Prékopa, "Probabilistic programming," *Handbooks in operations research and management science*, vol. 10, pp. 267–351, 2003.

[3] R. T. Rockafellar, "Coherent approaches to risk in optimization under uncertainty," in *OR Tools and Applications: Glimpses of Future Technologies*, INFORMS, Sep. 2007, pp. 38–61. DOI: 10.1287/educ.1073.0032. [Online]. Available: https://doi.org/10.1287/educ.1073.0032.

[4] N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, Apr. 2003. DOI: 10.1287/msom.5.2.79.16071. [Online]. Available: https://doi.org/10.1287/msom.5.2.79.16071.

[5] D. Gross, J. F. Shortie, J. M. Thompson, and C. M. Harris, *Simple Markovian Queueing Models*. Wiley, Jul. 2008, pp. 49–115. DOI: 10.1002/9781118625651.ch2. [Online]. Available: https://doi.org/10.1002/9781118625651.ch2.

[6] Z. Aksin, M. Armony, and V. Mehrotra, "The modern call center: A multi-disciplinary perspective on operations management research," *Production and Operations Management*, vol. 16, no. 6, pp. 665–688, Jan. 2009. DOI: 10.1111/j.1937-5956.2007.tb00288.x. [Online]. Available: https://doi.org/10.1111/j.1937-5956.2007.tb00288.x.

[7] D. Bertsimas and A. Thiele, "A data-driven approach to newsvendor problems," *Working Paper, Massachusetts Institute of Technology*, 2005.

[8] R. Levi, G. Perakis, and J. Uichanco, "The data-driven newsvendor problem: New bounds and insights," *Oper. Res.*, vol. 63, no. 6, pp. 1294–1306, 2015.

[9] H. E. Scarf, "Some remarks on bayes solutions to the inventory problem," *Nav. Res. Logist.*, vol. 7, no. 4, pp. 591–596, 1960.

[10] S. Lacoste-Julien, F. Huszár, and Z. Ghahramani, "Approximate inference for the loss-calibrated bayesian," in *Int. Conf. Artif. Intell. Statist.*, 2011, pp. 416–424.

[11] T. Kuśmierczyk, J. Sakaya, and A. Klami, "Variational bayesian decision-making for continuous utilities," in *Advances in Neural Information Processing Systems*, 2019, pp. 6395–6405.

[12] T. L. Lai, H. Xing, and Z. Chen, "Mean–variance portfolio optimization when means and covariances are unknown," *The Annals of Applied Statistics*, vol. 5, no. 2A, Jun. 2011. DOI: 10.1214/10-aoas422. [Online]. Available: https://doi.org/10.1214/10-aoas422.

[13] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Oper. Res.*, vol. 58, no. 3, pp. 595–612, 2010.

[14] G. Bayraksan and D. K. Love, "Data-driven stochastic programming using phi-divergences," in *The Operations Research Revolution*, INFORMS, 2015, pp. 1–19.

[15] D. Wu, H. Zhu, and E. Zhou, "A bayesian risk approach to data-driven stochastic optimization: Formulations and asymptotics," *SIAM J. Optim*, vol. 28, no. 2, pp. 1588–1612, 2018.

[16] K. Chowdhary and P. Dupuis, "Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 47, no. 3, pp. 635–662, Mar. 2013. DOI: 10.1051/m2an/2012038. [Online]. Available: https://doi.org/10.1051/m2an/2012038.

[17] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, Mar. 2021. DOI: 10.1007/s10994-021-05946-3. [Online]. Available: https://doi.org/10.1007/s10994-021-05946-3.

[18] T. Aktekin and T. Ekin, "Stochastic call center staffing with uncertain arrival, service and abandonment rates: A bayesian perspective," *Naval Research Logistics (NRL)*, vol. 63, no. 6, pp. 460–478, Sep. 2016. DOI: 10.1002/nav.21716. [Online]. Available: https://doi.org/10.1002/nav.21716.

[19] M. Lu, J. G. Shanthikumar, and Z.-J. M. Shen, "Technical note–operational statistics: Properties and the risk-averse case," *Nav. Res. Logist.*, vol. 62, no. 3, pp. 206–214, 2015.

[20] A. M. Stuart, "Inverse problems: A bayesian perspective," *Acta Numerica*, vol. 19, pp. 451–559, May 2010. DOI: 10.1017/s0962492910000061. [Online]. Available: https://doi.org/10.1017/s0962492910000061.

[21] H. Wang, X. Zhang, and S. H. Ng, "A nonparametric bayesian approach for simulation optimization with input uncertainty," *arXiv preprint arXiv:2008.02154*, 2020.

[22] M. Pearce and J. Branke, "Bayesian simulation optimization with input uncertainty," in *2017 Winter Simulation Conference (WSC)*, 2017, pp. 2268–2278. DOI: 10.1109/WSC.2017.8247958.

[23] E. Song, "Sequential bayesian risk set inference for robust discrete optimization via simulation," *arXiv preprint arXiv:2101.07466*, 2021.

[24] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal, "Markov chain monte carlo in practice: A roundtable discussion," *The American Statistician*, vol. 52, no. 2, pp. 93–100, May 1998. DOI: 10.1080/00031305.1998.10480547. [Online]. Available: https://doi.org/10.1080/00031305.1998.10480547.

[25] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[26] R. E. Turner, P. Berkes, and M. Sahani, "Two problems with variational expectation maximisation for time-series models," *Bayesian Time-Series Models*, 2011.

[27] Y. Wang and D. M. Blei, "Frequentist consistency of variational bayes," *J. Amer. Stat. Assoc.*, vol. 0, no. 0, pp. 1–15, Jun. 2018. DOI: 10.1080/01621459.2018.1473776. eprint: https://doi.org/10.1080/01621459.2018.1473776. [Online]. Available: https://doi.org/10.1080/01621459.2018.1473776.

[28] F. Zhang and C. Gao, "Convergence rates of variational posterior distributions," *The Annals of Statistics*, vol. 48, no. 4, Aug. 2020. DOI: 10.1214/19-aos1883. eprint: 1712.02519. [Online]. Available: https://doi.org/10.1214/19-aos1883.

[29] Y. Yang, D. Pati, and A. Bhattacharya, "α-variational inference with statistical guarantees," *arXiv preprint arXiv:1710.03266*, 2017.

[30] P. Jaiswal, V. Rao, and H. Honnappa, "Asymptotic consistency of α-rényi-approximate posteriors," *Journal of Machine Learning Research*, vol. 21, no. 156, pp. 1–42, 2020.

[31] A. Wald, *Statistical decision functions.* eng, [2d ed.] Bronx, N.Y.: Chelsea Pub. Co, 1971, ISBN: 0828402434.

[32] L. Schwartz, "On bayes procedures," *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, vol. 4, no. 1, pp. 10–26, Mar. 1965, ISSN: 1432-2064. DOI: 10.1007/BF00535479. [Online]. Available: https://doi.org/10.1007/BF00535479.

[33] S. N. Cohen, "Data-driven nonlinear expectations for statistical uncertainty in decisions," *Electronic Journal of Statistics*, vol. 11, no. 1, Jan. 2017. DOI: 10.1214/17-ejs1278. [Online]. Available: https://doi.org/10.1214/17-ejs1278.

[34] J. Duchi and H. Namkoong, "Variance-based regularization with convex objectives," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 2450–2504, 2019.

[35] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *International Conference on Machine Learning*, PMLR, 2018, pp. 1929–1938.

[36] E. Zhou and D. Wu, "Simulation optimization under input model uncertainty," in *Advances in Modeling and Simulation*, Springer, 2017, pp. 219–247.

[37] S. Cakmak, D. Wu, and E. Zhou, "Solving bayesian risk optimization via nested stochastic gradient estimation," *IISE Transactions*, pp. 1–13, 2021.

[38] Y. Lin, Y. Ren, and E. Zhou, "A bayesian risk approach to mdps with parameter uncertainty," *arXiv preprint arXiv:2106.02558*, 2021.

[39] P. Jaiswal, H. Honnappa, and V. A. Rao, "Risk-sensitive variational bayes: Formulations and bounds," *arXiv preprint arXiv:1903.05220v3*, 2019.

[40] G. C. Calafiore and L. El Ghaoui, "On distributionally robust chance-constrained linear programs," *Journal of Optimization Theory and Applications*, vol. 130, no. 1, pp. 1–22, 2006.

[41] G. C. Calafiore and M. C. Campi, "The scenario approach to robust control design," *IEEE Transactions on automatic control*, vol. 51, no. 5, pp. 742–753, 2006.

[42] M. C. Campi and G. C. Calafiore, "Notes on the scenario design approach," *IEEE Transactions on Automatic Control*, vol. 54, no. 2, pp. 382–385, 2009.

[43] W. Xie, "On distributionally robust chance constrained programs with wasserstein distance," *Mathematical Programming*, pp. 1–41, 2019.

[44] R. Jiang and Y. Guan, "Data-driven chance constrained stochastic program," *Mathematical Programming*, vol. 158, no. 1-2, pp. 291–327, 2016.

[45] A. R. Hota, A. Cherukuri, and J. Lygeros, "Data-driven chance constrained optimization under wasserstein ambiguity sets," in *2019 American Control Conference (ACC)*, IEEE, 2019, pp. 1501–1506.

[46] J. Luedtke and S. Ahmed, "A sample approximation approach for optimization with probabilistic constraints," *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 674–699, 2008.

[47] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, "Sample average approximation method for chance constrained programming: Theory and applications," *Journal of optimization theory and applications*, vol. 142, no. 2, pp. 399–416, 2009.

[48] X. Geng and L. Xie, "Data-driven decision making in power systems with probabilistic guarantees: Theory and applications of chance-constrained optimization," *Annual reviews in control*, vol. 47, pp. 341–363, 2019.

[49] A. Cherukuri and A. R. Hota, "Consistency of distributionally robust risk-and chance-constrained optimization under wasserstein ambiguity sets," *IEEE Control Systems Letters*, vol. 5, no. 5, pp. 1729–1734, 2020.

[50] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[51] A. Pena-Ordieres, J. R. Luedtke, and A. Wachter, "Solving chance-constrained problems via a smooth sample-based nonlinear approximation," 2019.

[52] Y. Li and R. E. Turner, "Rényi divergence variational inference," in *Advances in Neural Information Processing Systems*, 2016, pp. 1073–1081.

[53] R. E. Turner and M. Sahani, *Two problems with variational expectation maximisation for time-series models*, D. Barber, T. Cemgil, and S. Chiappa, Eds. Cambridge University Press, 2011, ch. 5, pp. 109–130.

[54] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[55] T. P. Minka, "A family of algorithms for approximate bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[56] M. Wainwright, T. Jaakkola, and A. Willsky, "A new class of upper bounds on the log partition function," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2313–2335, 2005. DOI: 10.1109/TIT.2005.850091.

[57] M. R. Andersen, A. Vehtari, O. Winther, and L. K. Hansen, "Bayesian inference for spatio-temporal spike-and-slab priors," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 5076–5133, Jan. 2017, ISSN: 1532-4435.

[58] A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. M. Blei, "Variational inference via χ -upper bound minimization," *arXiv preprint arXiv:1611.00328*, 2016.

[59] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[60] A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. Blei, "Variational inference via χ upper bound minimization," in *Advances in Neural Information Processing Systems*, 2017, pp. 2732–2741.

[61] F. Zhang and C. Gao, "Convergence rates of variational posterior distributions," *arXiv preprint arXiv:1712.02519*, 2017.

[62] H. Scarf, "Bayes solutions of the statistical inventory problem," *Ann. Math. Statist.*, vol. 30, no. 2, pp. 490–508, 1959.

[63] L. Y. Chu, J. Shanthikumar, and Z.-J. M. Shen, "Solving operational statistics via a bayesian analysis," *Operations Research Letters*, vol. 36, no. 1, pp. 110–116, Jan. 2008. DOI: 10.1016/j.orl.2007.04.010. [Online]. Available: https://doi.org/10.1016/j.orl.2007.04.010.

[64] S. E. Chick, "Chapter 9 subjective probability and bayesian methodology," in *Simulation*, Elsevier, 2006, pp. 225–257. DOI: 10.1016/s0927-0507(06)13009-1. [Online]. Available: https://doi.org/10.1016/s0927-0507(06)13009-1.

[65] D. Bauder, T. Bodnar, N. Parolya, and W. Schmid, "Bayesian mean–variance analysis: Optimal portfolio selection under parameter uncertainty," *Quantitative Finance*, vol. 21, no. 2, pp. 221–242, May 2020. DOI: 10.1080/14697688.2020.1748214. [Online]. Available: https://doi.org/10.1080/14697688.2020.1748214.

[66] T. Bodnar, S. Mazur, and Y. Okhrin, "Bayesian estimation of the global minimum variance portfolio," *European Journal of Operational Research*, vol. 256, no. 1, pp. 292–307, Jan. 2017. DOI: 10.1016/j.ejor.2016.05.044. [Online]. Available: https://doi.org/10.1016/j.ejor.2016.05.044.

[67] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, Feb. 2017.

[68] D. Pati, A. Bhattacharya, and Y. Yang, "On statistical optimality of variational bayes," in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, A. Storkey and F. Perez-Cruz, Eds., ser. Proceedings of Machine Learning Research, vol. 84, PMLR, Apr. 2018, pp. 1579–1588. [Online]. Available: http://proceedings.mlr.press/v84/pati18a.html.

[69] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. iv," *Commun. Pure Appl. Math.*, vol. 36, no. 2, pp. 183–212, 1983.

[70] A. W. van der Vaart and J. H. van Zanten, "Rates of contraction of posterior distributions based on Gaussian process priors," *The Annals of Statistics*, vol. 36, no. 3, pp. 1435–1463, 2008. DOI: 10.1214/009053607000000613. [Online]. Available: https://doi.org/10.1214/009053607000000613.

[71] D. Bertsimas and N. Kallus, "From predictive to prescriptive analytics," *arXiv preprint arXiv:1402.5481*, 2014.

[72] H. E. Scarf, "Some remarks on bayes solutions to the inventory problem," *Nav. Res. Logist.*, vol. 7, no. 4, pp. 591–596, 1960.

[73] D. Bertsimas, N. Kallus, and A. Hussain, "Inventory management in the era of big data," *Prod. Oper. Manage.*, vol. 25, no. 12, pp. 2006–2009, 2016.

[74] G.-Y. Ban and C. Rudin, "The big data newsvendor: Practical insights from machine learning," *Oper. Res.*, vol. 67, no. 1, pp. 90–108, 2018, Working Paper.

[75] D. Bertsimas and C. McCord, "Optimization over continuous and multi-dimensional decisions with observational data," in *Adv. Neural Inf. Process. Syst.*, 2018, pp. 2966–2974.

[76] Y. Deng, J. Liu, and S. Sen, "Coalescing data and decision sciences for analytics," in *Recent Advances in Optimization and Modeling of Contemporary Problems*, INFORMS, 2018, pp. 20–49.

[77] A. N. Elmachtoub and P. Grigas, "Smart" predict, then optimize"," *arXiv preprint arXiv:1710.08005*, 2017.

[78] B. Wilder, B. Dilkina, and M. Tambe, "Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization," *arXiv preprint arXiv:1809.05504*, 2018.

[79] T. Homem-de-Mello and G. Bayraksan, "Monte carlo sampling-based methods for stochastic optimization," *Surv. Oper. Res. Manage. Sci.*, vol. 19, no. 1, pp. 56–85, 2014.

[80] L. H. Liyanage and J. G. Shanthikumar, "A practical inventory control policy using operational statistics," *Oper. Res. Lett.*, vol. 33, no. 4, pp. 341–348, 2005.

[81] L. Y. Chu, J. G. Shanthikumar, and Z.-J. M. Shen, "Solving operational statistics via a bayesian analysis," *Oper. Res. Lett.*, vol. 36, no. 1, pp. 110–116, 2008.

[82] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *J. Amer. Stat. Assoc.*, vol. 101, no. 473, pp. 138–156, 2006.

[83] B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin, "Learning structured prediction models: A large margin approach," in *Proceedings of the 22nd international conference on Machine learning*, ACM, 2005, pp. 896–903.

[84] P. Jaiswal, H. Honnappa, and V. A. Rao, "Asymptotic consistency of decision-theoretic variational bayesian methods," In Preparation, 2019.

[85] R. T. Rockafellar, "Coherent approaches to risk in optimization under uncertainty," in *OR Tools and Applications: Glimpses of Future Technologies*, Informs, 2007, pp. 38–61.

[86] H. Föllmer and T. Knispel, "Entropic risk measures: Coherence vs. convexity, model ambiguity and robust large deviations," *Stochastics Dyn.*, vol. 11, no. 02n03, pp. 333–351, 2011.

[87] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, i," *Commun. Pure Appl. Math.*, vol. 28, no. 1, pp. 1–47, 1975.

[88] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, ii," *Commun. Pure Appl. Math.*, vol. 28, no. 2, pp. 279–301, 1975.

[89] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time—iii," *Commun. Pure Appl. Math.*, vol. 29, no. 4, pp. 389–461, 1976.

[90] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000, vol. 3.

[91] S. Ghosal, J. K. Ghosh, and A. W. van der Vaart, "Convergence rates of posterior distributions," *Ann. Statist.*, vol. 28, no. 2, pp. 500–531, 2000, ISSN: 00905364. [Online]. Available: http://www.jstor.org/stable/2674039.

[92] G. Pflug, "Stochastic optimization and statistical inference," in *Stochastic Programming*, ser. Handbooks in Operations Research and Management Science, vol. 10, Elsevier, 2003, pp. 427–482. DOI: https://doi.org/10.1016/S0927-0507(03)10007-2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0927050703100072.

[93] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2019, vol. 49.

[94] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *Int. Stat. Rev.*, vol. 70, no. 3, pp. 419–435, 2002.

[95] A. Braides, *Gamma-Convergence for Beginners*. Oxford University Press, Jul. 2002. DOI: 10.1093/acprof:oso/9780198507840.001.0001. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780198507840.001.0001.

[96] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[97] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[98] M. Gil, F. Alajaji, and T. Linder, "Rényi divergence measures for commonly used univariate continuous distributions," *Information Sciences*, vol. 249, pp. 124–131, 2013.

[99] M. H. Quang, *Regularized divergences between covariance operators and gaussian measures on hilbert spaces*, 2019. arXiv: 1904.05352 [math.PR].

[100] R. M. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in graphical models*, M. I. Jordan, Ed., Dordrecht: Springer, 1998, pp. 355–368.

[101] D. M. Blei and M. I. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.

[102] P. Alquier and J. Ridgway, "Concentration of tempered posteriors and of their variational approximations," *The Annals of Statistics*, vol. 48, no. 3, Jun. 2020. DOI: 10.1214/19-aos1855. [Online]. Available: https://doi.org/10.1214/19-aos1855.

[103] B.-E. Chérief-Abdellatif and P. Alquier, "Consistency of variational bayes inference for estimation and model selection in mixtures," *Electron. J. Statist.*, vol. 12, no. 2, pp. 2995–3035, 2018. DOI: 10.1214/18-EJS1475. [Online]. Available: https://doi.org/10.1214/18-EJS1475.

[104] T. Campbell and X. Li, "Universal boosting variational inference," *arXiv preprint arXiv:1906.01235*, 2019.

[105] B.-E. Chérief-Abdellatif, "Generalization error bounds for deep variational inference," *arXiv preprint arXiv:1908.04847*, 2019.

[106] J. H. Huggins, T. Campbell, M. Kasprzak, and T. Broderick, "Practical bounds on the error of bayesian posterior approximations: A nonasymptotic approach," *arXiv preprint arXiv:1809.09505*, 2018.

[107] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[108] J. Zhu, N. Chen, and E. P. Xing, "Bayesian inference with posterior regularization and applications to infinite latent svms," *Journal of Machine Learning Research*, vol. 15, p. 1799, 2014.

[109] A. W. van der Vaart, *Asymptotic Statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511802256.

[110] L. Schwartz, "On bayes procedures," *Probab. Theory Related Fields*, vol. 4, no. 1, pp. 10–26, 1965.

[111] S. Ghosal, "A review of consistency and convergence of posterior distribution," in *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*, 1997.

[112] E. Posner, "Random coding strategies for minimum entropy," eng, *Information Theory, IEEE Transactions on*, vol. 21, no. 4, pp. 388–391, 1975, ISSN: 0018-9448.

[113] J. Dupacova and R. Wets, "Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems," *The Annals of Statistics*, vol. 16, no. 4, pp. 1517–1549, Dec. 1988. DOI: 10.1214/aos/1176351052. [Online]. Available: https://doi.org/10.1214/aos/1176351052.

[114] W. K. Newey, "Uniform convergence in probability and stochastic equicontinuity," *Econometrica*, vol. 59, no. 4, p. 1161, Jul. 1991. DOI: 10.2307/2938179. [Online]. Available: https://doi.org/10.2307/2938179.

[115] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 969–996, 2006.

[116] C. Bandi and D. Gupta, "Operating room staffing and scheduling," *Manufacturing & Service Operations Management*, vol. 22, no. 5, pp. 958–974, Sep. 2020. DOI: 10.1287/msom.2019.0781. [Online]. Available: https://doi.org/10.1287/msom.2019.0781.

[117] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, "Computational study of a chance constrained portfolio selection problem," *Journal of Optimization Theory and Applications*, vol. 142, no. 2, pp. 399–416, 2009.

[118] Q. P. Zheng, J. Wang, and A. L. Liu, "Stochastic optimization for unit commitment—a review," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1913–1924, 2015. DOI: 10.1109/TPWRS.2014.2355204.

[119] A. Prékopa, *Stochastic Programming*. Springer Netherlands, 1995. DOI: 10.1007/978-94-017-3087-7. [Online]. Available: https://doi.org/10.1007/978-94-017-3087-7.

[120] C. M. Lagoa, X. Li, and M. Sznaier, "Probabilistically constrained linear programs and risk-adjusted controller design," *SIAM Journal on Optimization*, vol. 15, no. 3, pp. 938–951, Jan. 2005. DOI: 10.1137/s1052623403430099. [Online]. Available: https://doi.org/10.1137/s1052623403430099.

[121] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine Learning*, vol. 50, no. 1/2, pp. 5–43, 2003. DOI: 10.1023/a:1020281327116. [Online]. Available: https://doi.org/10.1023/a:1020281327116.

[122] F. Zhang and C. Gao, "Convergence rates of variational posterior distributions," *Annals of Statistics*, 2019. [Online]. Available: https://www.e-publications.org/ims/submission/AOS/user/submissionFile/34901?confirm=c41c6a61.

[123] P. Jaiswal, H. Honnappa, and V. A. Rao, "Asymptotic consistency of loss-calibrated variational bayes," *Stat*, vol. 9, no. 1, Feb. 2020. DOI: 10.1002/sta4.258. [Online]. Available: https://doi.org/10.1002/sta4.258.

[124] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, no. 4, p. 327, Nov. 1995. DOI: 10.2307/2684568. [Online]. Available: https://doi.org/10.2307/2684568.

[125] R. B. Grosse, Z. Ghahramani, and R. P. Adams, "Sandwiching the marginal likelihood using bidirectional monte carlo," *arXiv preprint arXiv:1511.02543*, 2015.

[126] M. J. Wainwright, M. I. Jordan, *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[127] B. Grechuk, A. Molyboha, and M. Zabarankin, "Maximum entropy principle with general deviation measures," *Mathematics of Operations Research*, vol. 34, no. 2, pp. 445–467, 2009.

[128] P. J. Bickel, B. J. Kleijn, *et al.*, "The semiparametric bernstein–von mises theorem," *The Annals of Statistics*, vol. 40, no. 1, pp. 206–237, 2012.

[129] S. A. Murphy and A. W. van der Vaart, "On profile likelihood," *Journal of the American Statistical Association*, vol. 95, no. 450, pp. 449–465, Jun. 2000. DOI: 10.1080/01621459.2000.10474219. [Online]. Available: https://doi.org/10.1080/01621459.2000.10474219.

[130] S. Murphy and A. van der Vaart, "Likelihood inference in the errors-in-variables model," *Journal of Multivariate Analysis*, vol. 59, no. 1, pp. 81–108, Oct. 1996. DOI: 10.1006/jmva.1996.0055. [Online]. Available: https://doi.org/10.1006/jmva.1996.0055.

[131] B. Kleijn and A. van der Vaart, "The bernstein-von-mises theorem under misspecification," *Electronic Journal of Statistics*, vol. 6, no. 0, pp. 354–381, 2012. DOI: 10.1214/12-ejs675. [Online]. Available: https://doi.org/10.1214/12-ejs675.

[132] R. Wong, "Ii - classical procedures," in *Asymptotic Approximations of Integrals*, R. Wong, Ed., Academic Press, 1989, pp. 55–146, ISBN: 978-0-12-762535-5. DOI: https://doi.org/10.1016/B978-0-12-762535-5.50006-4. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780127625355500064.

[133] T. M. Cover, *Elements of information theory*, eng, 2nd ed.. Hoboken, N.J.: Wiley-Interscience, 2006, ISBN: 0471241954.

[134] W. Feller, *An introduction to probability theory and its applications*, ser. Wiley series in probability and mathematical statistics. Probability and mathematical statistics v. 1. Wiley, 1968, ISBN: 9780471257080. [Online]. Available: https://books.google.com/books?id=mfRQAAAAMAAJ.

[135] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[136] P. Jaiswal, V. A. Rao, and H. Honnappa, "Asymptotic consistency of α-rényi-approximate posteriors.," *J. Mach. Learn. Res.*, vol. 21, pp. 156–1, 2020.

[137] D. Tran, R. Ranganath, and D. Blei, "Hierarchical implicit models and likelihood-free variational inference," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 5523–5533. [Online]. Available: http://papers.nips.cc/paper/7136-hierarchical-implicit-models-and-likelihood-free-variational-inference.pdf.

VITA

EDUCATION

Purdue University, West Lafayette, IN

Ph.D., School of Industrial Engineering – Aug 2016-July 2021

Indian Institute of Technology, Patna, India

B. Tech., Department of Mechanical Engineering – Aug 2008- May 2012

EXPERIENCE

- **Graduate Research Assistant, Stochastic Systems Lab, Purdue University**, May 2017 - May 2020. PI: Prof. Harsha Honnappa
- **Givens Associate, Argonne National Laboratory, Lemont, IL**, May 2020 - Aug 2020. PI: Dr. Mohan Krishnamoorthy
- **Givens Associate, Argonne National Laboratory, Lemont, IL**, May 2019 - Aug 2019. PI: Dr. Jeffrey M. Larson
- **Graduate Teaching Assistant, Industrial Engineering, Purdue University**. Engineering Economics (Fall 2016), Industrial Applications of Statistics (Spring 2017), Operations Research-Stochastic Models (Fall 2020), Probability and Statistics in Engineering (Spring 2021).
- **Assistant Manager, Bharat Petroleum Corp. Ltd. (BPCL), India**, July 2012 - Jul 2016.
- **Research Assistant, CACM, University of Auckland, NZ**, May - July. 2011. PI: Prof. Debes Bhattacharyya

PROJECTS/ PUBLICATIONS

- **Variational Bayesian (VB) methods for risk-sensitive machine/system design and posterior inference**

- Developed a novel risk-sensitive VB method for approximating a posterior distribution and used it for risk-sensitive data-driven system design problem.
- Implemented the algorithm in R (Optim package) and Python (Scipy, ADAM) to test its empirical performance on a system design problem.
- Derived theoretical bounds on the statistical performance of the proposed algorithm.
- **Research output:**

  1. (**JMLR'20**) **Jaiswal, P.**, Rao, V.A.; and Honnappa, H. "Asymptotic Consistency of $\alpha-$Rényi-Approximate Posteriors", *Journal of Machine Learning Research, (156):1–42, 2020.*
  2. (**STAT'20**) **Jaiswal, P.**, Honnappa, H., and Rao, V.A. "Asymptotic Consistency of Loss-calibrated Variational Bayes", *Stat 9, no. 1 (2020): e258.*
  3. (**NeurIPS'20 Workshop**) **Jaiswal, P.**, Honnappa, H., and Rao, V.A. "On the Statistical Consistency of Risk-Sensitive Bayesian Decision-Making", *Under revision.*
     Shorter version published at NeurIPS 2019 workshop on Safety and Robustness in Decision Making.

- **Variational Bayesian (VB) methods for Chance-constrained system design problems.**
  - Extended VB methodology to solve chance-constrained system design problems
  - Implemented the algorithm in R and tested its empirical performance on designing a Queuing system
  - Demonstrated the efficacy of the approach by comparing it against Markov Chain Monte Carlo (MCMC) method used for posterior approximation through sampling.
  - Established its statistical convergence properties with high -probability feasibility guarantee.

- **Research Output:**

  1. **(AABI'20) Jaiswal, P.**, Honnappa, H., and Rao, V.A. "Bayesian Joint Chance Constrained Optimization: Approximations and Statistical Consistency", *Submitted.*
     Shorter version published in Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference, PMLR 118:1-12, 2020.
  2. **(WSC'20) Jaiswal, P.**, and Honnappa, H. ' Statistical Inference for Approximate Bayesian Optimal Design'.*In Proceedings of the 2020 Winter Simulation Conference, Piscataway, NJ, 2020. Institute of Electrical and Electronics Engineers, Inc.*

- **Stochastic optimization methods for high-dimensional model calibration/ tuning**

  - Developed a multistart algorithm for non-convex stochastic optimization to find all the local minima of a non-convex function in a compact space.
  - Established asymptotic guarantees that our algorithm identifies all the local minima with high probability in finitely many local optimization runs.
  - Implemented the algorithm in Python with an interface to use local optimization techniques developed in MATLAB like ASTRO-DF (SIMOPT library) and tested its performance on two benchmark nonconvex problems.
  - **Research Output:**

    1. **Jaiswal, P.** and Larson, J. "Multistart Algorithm for Identifying All Optima of a Nonconvex Stochastic Oracle". *To be submitted to Journal of Global Optimization.*

- **Developing model-driven Deep Learning frameworks for statistical inference**

  - Developed a Variational deep-learning framework to compute posterior distribution over intensity of a Cox-process (Doubly stochastic Poisson Process).

- Implemented the method in Python (PyTorch) and MATLAB (Deep Learning Toolbox) and compared its performance against a Finite Element Method (FEM).
- **Research output:**
  1. **Jaiswal, P.**, Honnappa, H., and Rao, V.A. "Variational Inference for Diffusion Modulated Cox Processes", *Working paper*
  2. (**WSC'20**) Wang R., **Jaiswal, P.**, and Honnappa, H. 'Estimating Stochastic Poisson Intensities Using Deep Latent Models'.*In Proceedings of the 2020 Winter Simulation Conference, Piscataway, NJ, 2020. Institute of Electrical and Electronics Engineers, Inc.*

- **Large deviations**
  - **Research output:**
    1. **Jaiswal, P.**, Honnappa, H., and Pasupathy, R. 'Optimal Allocations for Sample Average Approximation '.*In Proceedings of the 2018 Winter Simulation Conference, Piscataway, NJ, 2018. Institute of Electrical and Electronics Engineers, Inc.*
    2. Honnappa, H., Pasupathy, R ; and **Jaiswal, P.** "The Large Deviations of Gaussian Extrema", *Working paper.*

SKILLS

Python (Pacakges: PyTorch, SciPy, NumPy, scikit-learn) • Matlab (Toolboxes: Deep Learning, Optimization, SIMOPT) • R (Optim) • LaTeX/ MS-Office • C/C++ • Version Control (Git, Bitbucket) • High Performance Computing (Clusters: Purdue RCAC- Brown and Argonne National Lab- Powell)

ACHIEVEMENTS

- Awarded Sustainable Horizons Institute grant to attend the SIAM CSE21 conference and Broader Engagement (BE) program.
- Awarded PGSG Travel grant to attend INFORMS 2020 and NeurIPS 2020.

- Qualified 3 papers from Institute of Actuaries of India (IAI), 2013-14.
- Awarded MCM scholarship for 3 consecutive academic years (2009-12) at IIT, Patna.
- Training & Placement Cell, Student Head (Founder), IIT Patna, 2010-12.
- Selected for University of Auckland-IIT internship program - 2011.

ACADEMIC SERVICE

Reviewer: IISE Transaction, ICML, NeurIPS, AISTATS, and WSC