

# Extending Mean-Field Variational Inference via Entropic Regularization: Theory and Computation

**Bohan Wu**

*Department of Statistics  
Columbia University  
New York, NY 10027, USA*

BW2766@COLUMBIA.EDU

**David M. Blei**

*Department of Computer Science and Department of Statistics  
Columbia University  
New York, NY 10027, USA*

DAVID.BLEI@COLUMBIA.EDU

**Editor:** Pierre Alquier

## Abstract

Variational inference (VI) has emerged as a popular method for approximate inference for high-dimensional Bayesian models. In this paper, we propose a novel VI method that extends the naive mean field via entropic regularization, referred to as  $\Xi$ -variational inference ( $\Xi$ -VI).  $\Xi$ -VI has a close connection to the entropic optimal transport problem and benefits from the computationally efficient Sinkhorn algorithm. We show that  $\Xi$ -variational posteriors effectively recover the true posterior dependency, where the likelihood function is downweighted by a regularization parameter. We analyze the role of dimensionality of the parameter space on the accuracy of  $\Xi$ -variational approximation and the computational complexity of computing the approximate distribution, providing a rough characterization of the statistical-computational trade-off in  $\Xi$ -VI, where higher statistical accuracy requires greater computational effort. We also investigate the frequentist properties of  $\Xi$ -VI and establish results on consistency, asymptotic normality, high-dimensional asymptotics, and algorithmic stability. We provide sufficient criteria for our algorithm to achieve polynomial-time convergence. Finally, we show the inferential benefits of using  $\Xi$ -VI over mean-field VI and other competing methods, such as normalizing flow, on simulated and real datasets.

**Keywords:** Variational inference, optimal transport, mean-field approximation, statistical-computational tradeoff, high-dimensional Bayesian inference.

## 1 Introduction

Variational inference (VI) is a widely used method for approximate probabilistic inference. VI approximates a difficult-to-compute distribution by positing a family of simpler distributions and minimizing the KL divergence between the family and the target. In Bayesian modeling, the target is a posterior distribution of latent variables given observations  $p(\theta | \mathbf{x})$  and the variational family is of distributions of the latent variables  $q(\theta) \in \mathbb{Q}(\Theta)$ . VI approximates the posterior with

$$q^*(\theta) = \arg \min_{q \in \mathbb{Q}} D_{\text{KL}}(q(\theta) \| p(\theta | \mathbf{x})). \quad (1.1)$$

To set up variational inference, we need to select the family of distributions over which to optimize. In many applications, practitioners use the *mean-field* or *fully factorized* family. This is the family of product distributions, where each variable is independent and endowed with its own distributional factor. Consider a model with  $D$  latent variables  $\theta = \{\theta_1, \dots, \theta_D\}$ . The corresponding mean-field family is

$$q(\theta) = \prod_{i=1}^D q_i(\theta_i), \quad (1.2)$$

where each  $q_i(\theta_i)$  is the variational factor for  $\theta_i$ . Thanks to this simple family, the variational optimization is computationally efficient (to a local optimum). But this efficiency comes at a cost. Mean-field VI suffers in accuracy because it cannot capture posterior dependencies between the elements of  $\theta$  (Blei et al., 2017).

In this paper, we develop a new way of doing variational inference. The idea is to optimize over *all* distributions of the latent variables, i.e.,  $q \in \mathbb{P}(\Theta)$ , but to regularize the variational objective function to encourage simpler distributions that are “more like the mean-field.” At one end of the regularization path we effectively optimize over the mean-field family, providing traditional mean-field VI (MFVI). At the other end we optimize over all distributions, providing exact inference (but at prohibitive cost). Between these extremes, our method smoothly trades off efficiency and accuracy.

In detail, consider a probabilistic model  $p(\theta, \mathbf{x}) = p(\theta)p(\mathbf{x} | \theta)$  and the goal to approximate the posterior  $p(\theta | \mathbf{x})$ . Denote the prior  $\pi(\theta) := p(\theta)$  and the log likelihood  $\ell(\mathbf{x}; \theta) := \log p(\mathbf{x} | \theta)$ . We propose to approximate the posterior by optimizing an *expressivity-regularized variational objective* over the entire space of distributions  $q \in \mathbb{P}(\Theta)$ .

Take an arbitrary distribution  $q(\theta)$  with marginal distributions denoted  $q_j(\theta_j)$ . We define the *expressivity functional* as the KL divergence between  $q(\theta)$  and the product of its marginals:

$$\Xi(q) = D_{\text{KL}} \left( q(\theta) \parallel \prod_{i=1}^D q_i(\theta_i) \right). \quad (1.3)$$

Expressivity measures the dependence among the  $D$  latent variables under  $q(\theta)$ . In the language of information theory, it is the multivariate mutual information of  $\theta \sim q(\theta)$  (Cover and Thomas, 2006). Intuitively, it quantifies how “un-mean-field” the distribution  $q$  is. A larger  $\Xi(q)$  indicates that the distribution is further from a factorized distribution.

We define  *$\Xi$ -variational inference* ( $\Xi$ -VI) as an expressivity-regularized optimization problem:

$$q_\lambda^*(\theta) = \arg \max_{q \in \mathbb{P}(\Theta)} \underbrace{\mathbb{E}_q [\ell(\mathbf{x}; \theta)] - D_{\text{KL}}(q(\theta) \parallel \pi(\theta))}_{\text{ELBO}(q)} - \underbrace{\lambda \Xi(q)}_{\text{Expressivity penalty}}. \quad (1.4)$$

The first two terms comprise the *evidence lower bound* (ELBO) (Jordan et al., 1999; Blei et al., 2017), which is the usual objective function for variational inference. When optimized relative to the full set of distributions of  $\theta$ , maximizing the ELBO recovers the exact posterior (Zellner, 1988; Knoblauch et al., 2022). The third term, however, is a penalty

term. It encourages the optimal  $q$  to resemble a product distribution, i.e., a member of the mean-field family. By varying  $\lambda > 0$ , we interpolate between the exact posterior and its mean-field approximation.

We will study the theory and application of Eq. (1.4), which we call  $\Xi$ -VI (pronounced “ksee VI”). First we show that we can solve this optimization by iterating between (1) calculating approximate posterior marginals for each variable and (2) solving a problem of *entropic optimal transport* (EOT) with a multi-marginal Sinkhorn algorithm (Cuturi, 2013; Lin et al., 2022). We then develop *expressivity-corrected mean field*. It first approximates marginals using traditional VI (e.g., black-box VI (Ranganath et al., 2014) or expectation propagation (Minka, 2013)), and then optimizes Eq. (1.4) with the Sinkhorn algorithm to model dependencies in the variational approximation.

We prove that  $\Xi$ -VI gives frequentist guarantees including posterior consistency and a Bernstein-von Mises theorem. Further, we theoretically characterize how to choose the regularization parameter  $\lambda$  to balance accuracy and efficiency. Specifically, we characterize the regions of possible  $\lambda$  values where the resulting variational approximation is either mean-field or Bayes-optimal.

Empirically, we apply  $\Xi$ -VI correction to multivariate Gaussians, linear regression with a Laplace prior, and hierarchical Bayesian modeling. The results demonstrate the competitive performance of  $\Xi$ -VI over other variational inference methods, including mean-field and full-rank ADVI (Kucukelbir et al., 2017), normalizing flow (Rezende and Mohamed, 2015), and Stein variational gradient descent (Liu and Wang, 2016). To set the regularization strength  $\lambda$ , our empirical findings suggest that  $\lambda = D$  is a reasonable choice, marking a phase transition between the computationally efficient and statistically accurate regimes.

The rest of the paper is organized as follows. Section 2 introduces  $\Xi$ -VI and the  $\Xi$ -VI correction algorithm. Section 3 provides an empirical study. Section 4 establishes theoretical guarantees for the  $\Xi$ -variational posterior, including posterior consistency, Bernstein-von Mises theorem, high-dimensional bounds, finite-sample convergence, and algorithmic stability. Section 5 concludes the paper with a discussion of limitations and further research.

**Related Work.** This paper proposes  $\Xi$ -VI, a new way to relax the mean-field assumption in variational inference. With this new algorithm, we also add to two existing areas of VI research: statistical guarantees and computational guarantees.

Mean-field VI is efficient, but it also has limitations. It poorly approximates posteriors in settings such as multivariate Gaussian models (Blei et al., 2017), state-space models (Wang and Titterington, 2004), piecewise-constant models (Zhang and Gao, 2020), and spike covariance models (Ghorbani et al., 2019). To address these shortcomings, researchers have proposed a variety of solutions, including structural VI (Xing et al., 2012; Ranganath et al., 2016), copula-based methods (Tran et al., 2015, 2017), linear response corrections (Giordano et al., 2018; Raymond and Ricci-Tersenghi, 2017), TAP corrections (Opper and Saad, 2001; Fan et al., 2021; Celentano et al., 2023a,b), and variational boosting (Miller et al., 2017; Locatello et al., 2018). Our method makes a contribution to this landscape of research, providing a principled and theoretically supported approach to capture dependencies among latent variables and to manage the statistical-computation tradeoff.

Several lines of recent research examine the statistical properties of VI approximations. This work includes results on asymptotic normality (Hall et al., 2011b,a; Bickel et al.,

2013; Wang and Blei, 2019), posterior contraction rates (Zhang and Gao, 2020; Zhang and Zhou, 2020), finite-sample bounds (Alquier et al., 2016; Alquier and Ridgway, 2020; Yang et al., 2020), and performance in high-dimensional settings (Basak and Mukherjee, 2017; Ray et al., 2020; Ray and Szabó, 2022; Mukherjee and Sen, 2022; Mukherjee et al., 2023; Qiu, 2024). We contribute to this research by proving frequentist guarantees—posterior consistency and a Bernstein-von Mises theorem—for our proposed class of variational approximations.

Other research examines computational aspects of VI, including convergence rates for coordinate ascent methods (Mukherjee and Sarkar, 2018; Plummer et al., 2020; Zhang and Zhou, 2020; Xu and Campbell, 2022; Bhattacharya et al., 2025), black-box optimization (Kim et al., 2023), and the trade-off between statistical accuracy and computational complexity (Bhatia et al., 2022). Related work also analyzes VI through gradient flow techniques (Yao and Yang, 2022; Lambert et al., 2022; Diao et al., 2023; Jiang et al., 2025). Our paper contributes to these computational analyses by explicitly characterizing the trade-off of accuracy for computational simplicity. We also expand the interface between VI and optimal transport, in using entropic optimal transport methods (Cuturi, 2013; Lin et al., 2022) in VI optimization.

## 2 $\Xi$ -variational inference

Again, we consider a general probabilistic model

$$p(\theta, \mathbf{x}) = \pi(\theta) \exp\{\ell(\mathbf{x}; \theta)\}, \quad (2.1)$$

where  $\pi(\theta)$  is the prior of the unknown parameter and  $\ell(\mathbf{x}; \theta)$  is the log likelihood of the data under  $\theta$ . While we use the log-likelihood throughout our methods and experiments (Section 3), all proposed techniques and theoretical results extend readily to generalized posteriors in which  $\ell(\mathbf{x} | \theta)$  is a loss function (Bissiri et al., 2016; Knoblauch, 2019; Miller, 2021). We consider the prior  $\pi$  that is a product distribution of the form  $\pi(\theta) = \prod_{i=1}^d \pi_i(\theta_i)$ . Our goal is to approximate the posterior  $p(\theta | \mathbf{x})$ .

In this section, we formally define  $\Xi$ -VI and analyze its structure. We reformulate  $\Xi$ -VI as a nested optimization, separating the problem into an outer optimization over marginals and an inner optimization over their couplings, i.e., a representation of the dependency structure in the variational approximation. We focus on solving the inner optimization, showing how to correct mean-field (factorized) solutions using entropic optimal transport (EOT). We present a computationally practical algorithm and discuss its interpretation.

### 2.1 The $\Xi$ variational objective and its nested formulation

We aim to find the distribution  $q_\lambda^*$  that solves the problem in Eq. (1.4). In this problem,  $\lambda \geq 0$  is a user-defined regularization parameter, and the optimal  $q_\lambda^*(\theta)$  is called the  $\Xi$ -variational posterior. When it is not unique,  $q_\lambda^*$  is one of the optimizers of Eq. (1.4).

We make two observations about the  $\Xi$ -VI problem: (1) When  $\lambda = 0$ ,  $q_0^*$  is the exact posterior. When  $\lambda = \infty$ ,  $q_\infty^*$  is a mean-field variational posterior. (2) By the standard duality theory, the  $\Xi$ -VI problem is equivalent to optimizing the standard ELBO over a

neighborhood of the mean-field family:

$$q_\lambda^* = \arg \max_{q \in \mathbb{P}(\Theta): \Xi(q) \leq \delta} \text{ELBO}(q).$$

The  $\Xi$ -VI posterior is the distribution over the latent variables closest to the posterior, but within the neighborhood of expressivity.

We can rewrite  $\Xi$ -VI as a *nested minimization* problem. Let  $m_i(\theta_i)$  denote a marginal distribution of  $\theta_i$  and let  $\mathbb{M}(\Theta)$  denote the space of product distributions over  $\Theta$ ,

$$\mathbb{M}(\Theta) = \left\{ m(\theta) : m(\theta) = \prod_{i=1}^D m_i(\theta_i) \right\}. \quad (2.2)$$

Given a set of  $D$  marginals let  $\mathcal{C}(m_1, \dots, m_D)$  denote the set of  $D$ -dimensional joint distributions where  $m_j(\theta_j)$  is the  $j^{th}$  marginal,

$$\mathcal{C}(m_1, \dots, m_D) = \{q(\theta_1, \dots, \theta_D) : q_j(\theta_j) = m_j(\theta_j), j = 1, \dots, D\}. \quad (2.3)$$

The set  $\mathcal{C}(m_1, \dots, m_D)$  is called the set of *couplings* over the distributions  $\{m_1, \dots, m_D\}$ . As shorthand, we write  $\mathcal{C}(m)$  as the set of couplings over the marginal distributions of  $m(\theta)$ . Note the set  $\mathcal{C}(m)$  is convex and closed in the Wasserstein distance (Nutz, 2021), and we assume that there exists  $q \in \mathcal{C}(m)$  with finite (Boltzmann) entropy.

With these definitions in place, we write  $\Xi$ -VI as a double minimization problem,

$$\min_{m \in \mathbb{M}(\Theta)} \min_{q \in \mathcal{C}(m)} \mathbb{E}_q[-\ell(\mathbf{x}; \theta)] + \lambda D_{\text{KL}}(q \parallel m) + D_{\text{KL}}(q \parallel \pi). \quad (2.4)$$

The equation follows from expressing the minimization set  $\mathbb{P}(\Theta)$  as  $\{q \in \mathcal{C}(m), m \in \mathbb{M}(\Theta)\}$ , while the objective stays the same.

In Eq. (2.4), the *outer variational problem* minimizes the objective with respect to the space of marginal distributions. Given a set of marginals, the *inner variational problem* minimizes the objective over its set of couplings. Here we will focus on the inner variational problem. Given fixed marginal distributions—such as those produced by mean-field VI—the inner problem finds the optimal coupling that corrects these marginals, i.e., the  $\lambda$ -regularized optimal dependencies between the variables. We show this problem is solvable using entropic optimal transport tools. Our method improves a given mean-field solution to capture dependencies in the latent variables and better approximate the posterior.

## 2.2 Expressivity-corrected mean-field VI

We now derive an algorithm to correct mean-field variational inference using  $\Xi$ -VI. We fix the solution to the outer variational problem with a product distribution  $m(\theta)$ , obtained from mean-field VI or another approximate method. We then solve Eq. (2.4) for the optimal distribution  $q(\theta)$  that matches these marginals, i.e., by solving the inner variational problem with respect to the coupling  $q \in \mathcal{C}(m)$ . As we will see, we can optimize over the set of couplings using tools from entropic optimal transport (EOT) (Villani, 2009; Nutz, 2021).

**Solving the inner variational problem with the Sinkhorn algorithm.** We view the inner variational problem as an EOT problem (Nutz, 2021). Again, we fix  $m$  and optimize  $q$ . A simple calculation (in Section C) shows that

$$q_\lambda^*(\theta) = \arg \min_{q \in \mathcal{C}(m)} \mathbb{E}_q[-\ell(\mathbf{x}; \theta)] + (\lambda + 1) D_{\text{KL}}(q \parallel m). \quad (2.5)$$

Theorem 25 (Section C) shows the unique solution to Eq. (2.5) has the following form:

$$q_\lambda^*(\theta) = \exp \left( \sum_{i=1}^D \phi_i^*(\theta_i) + \frac{1}{\lambda + 1} \ell(\mathbf{x}; \theta) \right) m(\theta), \quad (2.6)$$

where each  $\phi_i^* : \Theta_i \rightarrow \mathbb{R}$  is a measurable function, called an *EOT potential*.

The set of EOT potentials  $\phi^* := (\phi_1^*, \dots, \phi_D^*)$  are identifiable up to an additive constant. So we identify the solution by imposing  $D - 1$  constraints that each one has mean-zero under the marginal,

$$\mathbb{E}_{m_1} \phi_1^*(\theta_1) = \dots = \mathbb{E}_{m_{D-1}} \phi_{D-1}^*(\theta_{D-1}) = 0. \quad (2.7)$$

Let  $\mathbf{E}(m)$  denote the space of  $\phi$  such that Eq. (2.7) holds. We find the optimal potentials from Eq. (2.6) by maximizing the Lagrangian dual problem,

$$\phi^* = \arg \max_{\phi \in \mathbf{E}(m)} \sum_{i=1}^D \mathbb{E}_{m_i} [\phi_i(\theta_i)] - \mathbb{E}_m \left[ \exp \left( \sum_{i=1}^D \phi_i(\theta_i) + \frac{1}{\lambda + 1} \ell(\mathbf{x}; \theta) \right) \right]. \quad (2.8)$$

See Section C for the derivation.

Finally, we solve Eq. (2.8) with a block coordinate ascent algorithm called the *Sinkhorn algorithm* (Cuturi, 2013). Given the marginals  $m^t$  at time  $t$ , the Sinkhorn algorithm iteratively updates each  $\phi_i$ ,

$$\phi_i^{t+1} = \arg \max_{\phi_i \in L_0^1(m_i^t)} \mathbb{E}_{m_i^t} \phi_i(\theta_i) + \sum_{j=1}^{i-1} \mathbb{E}_{m_j^t} \phi_j^{t+1}(\theta_j) + \sum_{j=i+1}^D \mathbb{E}_{m_j^t} \phi_j^t(\theta_j) - \mathbb{E}_{m^t} [\exp(\Lambda^{t+1}(\theta_i, \theta_{-i}))], \quad (2.9)$$

where

$$\Lambda^{t+1}(\theta_i, \theta_{-i}) := \sum_{j=1}^{i-1} \phi_j^{t+1}(\theta_j) + \sum_{j=i+1}^D \phi_j^t(\theta_j) + \frac{1}{\lambda + 1} \ell(\mathbf{x}; \theta_i, \theta_{-i}).$$

To solve for Eq. (2.9), the update has an explicit formula:

$$\phi_i^{t+1}(\theta_i) = -\log \mathbb{E}_{m_{-i}^t} \exp(\Lambda^{t+1}(\theta_i, \theta_{-i})) + \eta_i^t, \quad \forall \theta_i \in \Theta_i, \quad (2.10)$$

where

$$\eta_i^t = \begin{cases} \mathbb{E}_{m_i^t} \log \mathbb{E}_{m_{-i}^t} \exp(\Lambda^{t+1}(\theta_i, \theta_{-i})), & \text{for } i \leq D - 1, \\ 0 & \text{otherwise.} \end{cases}$$

The updated EOT potentials satisfy the identifiability constraints (2.7). The solution  $q_\lambda^*$  calculated with these EOT potentials is a valid probability distribution.

In practice, the expectations required of Eq. (2.10) might be difficult to compute. In our algorithm, we approximate them with Monte Carlo.

Specifically, given samples  $\theta_i^1, \dots, \theta_i^N \sim m_i^t$  for  $i \in [D]$ , we approximate  $\phi_i^{t+1}$  with the Monte Carlo estimate  $\hat{\phi}_i^{t+1}$ :

$$\hat{\phi}_i^{t+1}(\theta_i) = -\log \left( \frac{1}{N^{D-1}} \sum_{J \in [N]^{D-1}} \exp(\hat{\Lambda}^{t+1}(\theta_i, \theta_{-i}^J)) \right) + \hat{\eta}_i^T, \quad (2.11)$$

where

$$\hat{\Lambda}^{t+1}(\theta_i, \theta_{-i}) := \sum_{j=1}^{i-1} \hat{\phi}_j^{t+1}(\theta_j) + \sum_{j=i+1}^D \hat{\phi}_j^t(\theta_j) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta_i, \theta_{-i}).$$

and

$$\hat{\eta}_i^t = \begin{cases} \frac{1}{N} \sum_{k=1}^N \log \left( \frac{1}{N^{D-1}} \sum_{J \in [N]^{D-1}} \exp(\hat{\Lambda}^{t+1}(\theta_i, \theta_{-i}^J)) \right), & \text{for } i \leq D-1, \\ 0 & \text{otherwise.} \end{cases}$$

**One-step expressivity-corrected mean-field VI.** Algorithm 1 implements the entropic correction in a single round of updates. In the first stage, it computes a set of pseudomarginals  $\{\tilde{m}_i\}_{i \in [D]}$ , and draws samples from them. In the second stage, it uses those samples in a multi-marginal Sinkhorn algorithm to compute the optimal EOT coupling.

---

**Algorithm 1:** Expressivity-corrected mean-field variational inference

---

**Input:**

- Data  $\mathbf{x}$
- Likelihood function  $\ell(x; \theta)$
- Prior distribution  $\pi(\theta)$
- Expressivity regularization parameter  $\lambda > 0$

Compute marginals from a mean-field algorithm

$$\tilde{m}_1, \dots, \tilde{m}_D \leftarrow \text{mean field inference}(\mathbf{x}, \ell, \pi).$$

Iteratively compute EOT potentials

$$\tilde{\phi}_{\lambda,1}, \dots, \tilde{\phi}_{\lambda,D} \leftarrow \text{Sinkhorn}(\mathbf{x}, \ell, \tilde{m}_{1:D}, \lambda) \quad \text{Eqs. (2.9) and (2.10)}$$

---

**Output:**  $\tilde{q}_\lambda(\theta) = \exp \left( \sum_{i=1}^D \tilde{\phi}_{\lambda,i}(\theta_i) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta) \right) \prod_{i=1}^D \tilde{m}_i(\theta_i)$

---

Note that in the first stage, we can use any algorithm for approximating the posterior marginals, e.g., variational inference (Blei et al., 2017), expectation propagation (EP)

(Minka, 2013), or MCMC (Robert and Casella, 2004). Ideally, the first step of Algorithm 1 would produce accurate estimates of the marginals of the exact posterior.

In practice, we recommend using an approximate method that yields overdispersed marginals, such as EP, because the additional variability often improves downstream coupling approximations. Intuitively, it produces more variation in the initial samples of  $\{\theta_i^1, \dots, \theta_i^N\}_{i \in [D]}$  for the Monte Carlo step (2.11), which leads to better downstream approximations. We demonstrate this empirically in Section 3.

Algorithm 1 only outputs an approximate solution to the full  $\Xi$ -VI problem in Eq. (2.4). However, by coupling the marginals  $\tilde{m}_1, \dots, \tilde{m}_D$ , the final estimate  $\tilde{q}_\lambda(\theta)$  is guaranteed to be *at least as good as* the initial approximation  $\tilde{m}(\theta) := \prod_{j=1}^D m_j(\theta_j)$  in terms of KL divergence to the exact posterior. The reason is that  $\tilde{q}_\lambda$  maximizes the regularized ELBO in Eq. (1.4) over the coupling  $\mathcal{C}(\tilde{m})$ .

With a large number of variables, Algorithm 1 is computationally challenging because the cost of averaging over  $N^{D-1}$  term in step (2.11) scales exponentially in  $D$ . For this reason, a stochastic (minibatch) approximation to step (2.11) is necessary in practice when either  $N$  or  $D$  is large. In Section A, we outline conditions on the likelihood for the algorithm to be polynomial-time solvable. Specifically, we provide polynomial-time complexity guarantees in two settings: (i) graphical models with bounded treewidth, and (ii) models in which the likelihood evaluated at the sample points  $\{\theta_i^1, \dots, \theta_i^N\}_{i \in [D]}$  forms a low-rank and sparse tensor. In the first setting, we show that the algorithm converges in time polynomial in the dimension  $D$ , but exponential in the treewidth and inversely proportional to the regularization parameter  $\lambda$ .

### 2.3 $\Xi$ -VI Solution and Connection to Generalized Bayes

In this section, we discuss the structure of the  $\Xi$ -VI solution and its connection to existing theories of generalized Bayesian methods (Knoblauch et al., 2022).

As shown in Eq. (2.6), the  $\Xi$ -VI solution consists of three components: (i) a scaled log-likelihood term, (ii) a set of potential functions  $\{\phi_{\lambda,i}^*\}_{i=1}^D$ , and (iii) a product of marginals  $m_\lambda^*(\theta) = \prod_{i=1}^D m_{\lambda,i}^*(\theta_i)$ . The regularization parameter  $\lambda$  controls the temperature of the likelihood term  $\ell(\mathbf{x}; \theta)$ .

Intuitively,  $\lambda$  divides a sample size of  $n$  between the true posterior and the mean-field solution by a factor of  $1/(\lambda + 1)$  and  $\lambda/(\lambda + 1)$ , respectively. It thus quantifies the tradeoff between the likelihood and a product distribution. Higher  $\lambda$  allows the variational posterior to be close to the mean field, while lower  $\lambda$  allows the solution to better approximate the exact posterior (but at computational cost). When  $\lambda = 0$ , the likelihood term is untempered—the variational solution is the exact posterior. When  $\lambda = \infty$ , the solution matches the mean-field variational posterior. The curve of measures  $\{q_\lambda^*, \lambda \in \mathbb{R}_+\}$  smoothly interpolates between the mean-field variational posterior and the true posterior.

We can view  $q_\lambda^*$  as a nonlinear tilt of a  $1/(\lambda + 1)$ -tempered posterior (Miller and Dunson, 2018; Bhattacharya et al., 2019). Posteriors of this form have been studied extensively in statistical learning theory for addressing model misspecification (Grünwald, 2012; Grünwald and van Ommen, 2017), and in machine learning they have been studied for their predictive performance where they are called “cold posteriors” (Aitchison, 2021; Wenzel et al., 2020;

McLatchie et al., 2025). Write

$$f_{\lambda,i}^*(\theta_i) := \phi_{\lambda,i}^*(\theta_i) + \log m_{\lambda}^*(\theta_i) - \log \pi_i(\theta_i).$$

Then we can represent  $q_{\lambda}^*$  as a nonlinear tilt of the tempered posterior  $q_0^{*\lambda}$ ,

$$q_{\lambda}^*(\theta) \propto \exp \left( \sum_{i=1}^D f_{\lambda,i}^*(\theta_i) \right) q_0^{*\lambda}, \quad \text{where} \quad q_0^{*\lambda}(\theta) \propto \exp \left( \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta) \right) \pi(\theta). \quad (2.12)$$

Wainwright et al. (2008) shows that the mean-field variational posterior of the quadratic interaction model amounts to a linear tilting of the prior. Eq. (2.12) extends this result, where  $f_{\lambda,i}^*(\theta_i)$  is the tilting function.

### 3 Examples

We apply  $\Xi$ -VI to three statistical models: a multivariate Gaussian model, a high-dimensional Bayesian linear regression, and a hierarchical Bayesian model on the 8-schools data (Gelman et al. (1995), Section 5.5).

- In the multivariate Gaussian example,  $\Xi$ -VI is explicitly solvable. This example illustrates the limitations of mean-field VI (Blei et al., 2017), and demonstrates how  $\Xi$ -VI improves it.
- In high-dimensional Bayesian linear regression, mean-field VI produces valid inference under weak covariate interactions (Mukherjee and Sen, 2022; Mukherjee et al., 2023), but fail when the interaction among the covariates is strong (Qiu, 2024; Celentano et al., 2023a). Again,  $\Xi$ -VI improves on the classical approach.
- Our analysis of the Bayesian hierarchical model shows how  $\Xi$ -VI provides more accurate posterior inferences on a real-world dataset.

#### 3.1 Multivariate Gaussian distributions

We first apply  $\Xi$ -VI to approximating a multivariate Gaussian with the family of all Gaussian distributions. In this demonstration, no algorithm is needed because  $\Xi$ -VI admits a closed form solution. In general, it is well known that mean-field VI underestimates the marginal variance of its target posterior (Blei et al., 2017). Here we show how  $\Xi$ -VI interpolates between the mean-field and the target posterior, and strictly outperforms mean-field VI in covariance estimation.

Assume that the exact posterior is a multivariate normal,  $q_0^* := N(\mu_0, \Sigma_0)$  with  $D$ -dimensional mean vector  $\mu_0$  and a  $D \times D$  full-rank covariance matrix  $\Sigma_0$ . The  $\Xi$ -VI formulation is

$$q_{\lambda}^* = \arg \min_{q \in \mathcal{N}(\mu, \Sigma)} D_{\text{KL}}(q \parallel q_0^*) + \lambda \Xi(q). \quad (3.1)$$

The next result establishes the self-consistency equations for the  $\Xi$ -VI solution and establishes upper and lower bounds for the approximating covariance:

**Proposition 1** Suppose we solve the Gaussian  $\Xi$ -VI problem (3.1) with  $\mathcal{N}(\mu_0, \Sigma_0)$  the exact posterior and  $\lambda > 0$ . Then the minimizer  $q_\lambda^* = \mathcal{N}(\mu^*, \Sigma^*)$  where  $\mu^*, \Sigma^*$  satisfy the following fixed point equations:

$$\mu^* = \mu_0, \quad (\Sigma^*)^{-1} = \frac{1}{\lambda+1} \Lambda_0 + \frac{\lambda}{\lambda+1} (\Sigma_{diag}^*)^{-1}.$$

For any matrix norm  $\|\cdot\|$ , the following bounds hold:

$$\left\| \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{\lambda}{\lambda+1} \Sigma_{0,diag}^{-1} \right]^{-1} - \Sigma_0 \right\| \leq \|\Sigma^* - \Sigma_0\| \leq \left\| \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{\lambda}{\lambda+1} \Lambda_{0,diag} \right]^{-1} - \Sigma_0 \right\|.$$

The proof can be found in Section E. Our result shows that  $\Lambda^*$  is a convex combination of the true precision  $\Lambda_0$  and the inverse of the variational marginal variances. As the regularizer  $\lambda \rightarrow \infty$ , the off-diagonal elements of  $\Lambda^*$  converge to 0 while the diagonal elements approach those of  $\Lambda_0$ .

The weight  $\lambda$  controls the approximation error of a variational posterior covariance by combining the marginal precisions of the exact posterior and the mean-field precision with weights  $\frac{1}{\lambda+1}$  and  $\frac{\lambda}{\lambda+1}$ , respectively. For any  $\lambda < \infty$ , the  $\Xi$ -variational posterior offers a tighter approximation than the naive mean field. To see this, we note that  $D_{\text{KL}}(q_\lambda^* \| q_0^*) \leq D_{\text{KL}}(q_\lambda^* \| q_0^*) + \lambda \Xi(q_\lambda^*) \leq D_{\text{KL}}(q_\infty^* \| q_0^*)$ , where  $\Xi(q_\lambda^*) \geq 0$  implies the first inequality and the optimality of  $q_\lambda^*$  implies the second inequality.

As a concrete demonstration of these ideas, we study a bivariate Gaussian posterior. Here, the  $\Xi$ -variational posterior has an analytical solution that can be exactly computed (see Proposition 38 in the Section E).

Figure 1 illustrates the interpolation, where the regularization downweights the off-diagonal entries of the precision matrix by a factor of  $1/(\lambda+1)$ . It shows  $\hat{q}_\lambda$  fitted to a bivariate Gaussian, for different values of  $\lambda$ . The left panel shows  $\hat{q}_\lambda$  as a smooth interpolation between the true posterior and the mean-field variational posterior. Increasing  $\lambda$  smoothly reduces posterior dependence, with a sharp structural change only at  $\lambda = \infty$ . The right panel paints a quantitative picture of this interpolation: when  $\lambda \leq 10^{-1}$ , the  $\Xi$ -variational posterior closely approximates the covariance values of the exact bivariate Gaussian posterior. For  $\lambda \geq 10^1$ , the covariance is close to zero, which indicates proximity to the mean-field variational posterior. Both plots suggest that  $\hat{q}_\lambda$  undergoes a "phase transition" phenomenon at  $\lambda \in [10^{-1}, 10^1]$ .

### 3.2 Bayesian Linear Regression with Laplace Prior

$\Xi$ -VI involves a tradeoff between statistical accuracy and computational complexity: as the regularization increases away from the mean-field solution, the quality of VI approximation improves at the cost of increased computational complexity.

To study this, we consider a Bayesian linear model with Laplace prior,

$$\mathbf{y} = \mathbf{X}\theta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n), \quad \theta_i \sim \text{Laplace}(0, 1). \quad (3.2)$$

The Laplace prior has density  $\pi(\theta_i) = \frac{1}{2b} \exp\left(-\frac{|\theta_i|}{b}\right)$ .

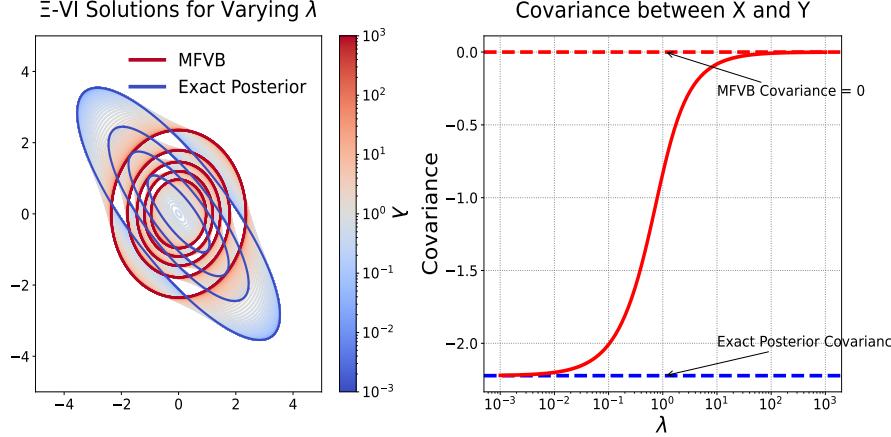


Figure 1:  $\Xi$ -VI solutions for a bivariate Gaussian posterior for varying  $\lambda$ . The left panel illustrates the transition of the variational posterior  $q_\lambda^*$  from closely approximating the exact posterior (at low  $\lambda$ ) to resembling the mean-field approximation (at high  $\lambda$ ). The right panel shows the covariance between the two normal coordinates versus  $\lambda$  on a log scale. Note that the  $\Xi$ -variational approximation to the covariance is very accurate up to a critical  $\lambda$  ( $\approx 10^{-1}$ ), after which it degrades rapidly to 0.

We simulate a dataset consisting of  $n = 100$  observations and  $d = 12$  features. The true regression coefficients are drawn randomly from a 12-dimensional standard Gaussian distribution, and  $\sigma^2 = 1$ . Columns (1, 2, 3, 8, 9) of  $\mathbf{X}$  are generated from a standard Gaussian distribution. Then we set each of features (4, 5, 6, 11, 12) equal to each of features (1, 2, 3, 8, 9) plus a standard Gaussian noise. This setup aims to simulate realistic multicollinearity. Finally, we generate the response  $\mathbf{y}$  using model (3.2). With this simulated data, we calculate an ‘‘exact’’ posterior with a long-run MCMC algorithm of 3,000 iterations. The MCMC draws produce an  $\hat{R}$  of below 1.01 across coefficients (Gelman et al., 1995), which is below the typical threshold of 1.1 for satisfactory mixing.

Since coupling all 12 coefficients is computationally expensive, we couple groups of coefficients in the EOT step. We adopt a naive grouping approach where features (1, 2, 3), (4, 5, 6), (7, 8, 9), (10, 11, 12) are grouped together. This effectively reduces the computational cost by reducing a twelve-dimensional coupling problem into a four-dimensional one. While it is beneficial to use an informed grouping, any choice of grouping will improve the approximation accuracy of MFVI.

For each dimension, we use  $N = 20$  points to represent each pseudomarginal. Recent work by Frazier et al. (2024) shows that generalized Bayes posteriors can be sensitive to the number of pseudo-samples used in stochastic likelihood estimation. In our setting, the posterior  $\hat{q}_\lambda$  depends on the EOT potentials computed via the Sinkhorn algorithm (Eq. (2.6)). To assess this sensitivity, Table 1 in Section F reports the approximation quality of  $\Xi$ -VI for various  $N$  under  $\lambda = 10$ . Larger  $N$  offers only modest accuracy gains while the memory cost scales as  $O(N^D)$  (see Eq. (2.11)) and runtime increases sharply. Empirically,  $N = 20$  is sufficient to accurately compute  $\Xi$ -VI under a reasonable computational budget.

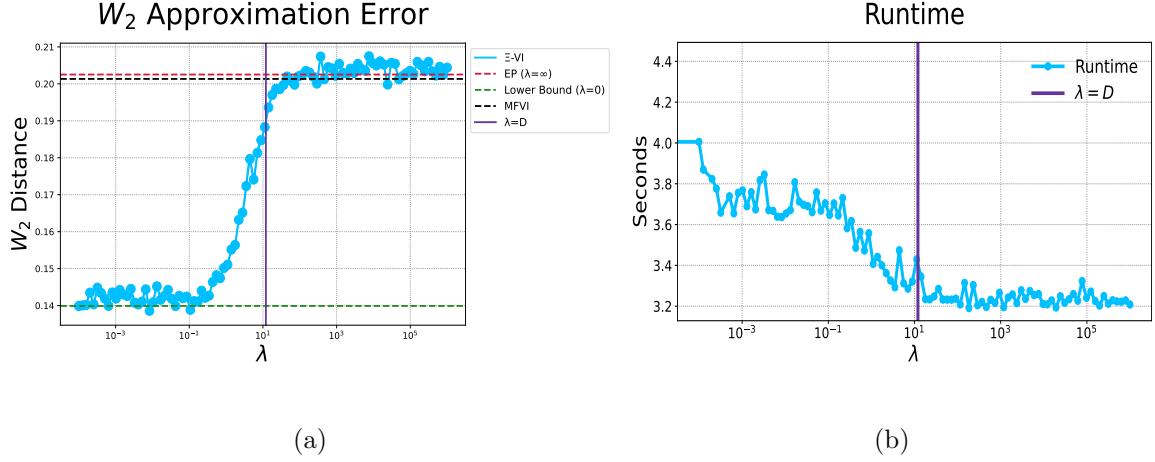


Figure 2: **Left.** accuracy of  $\Xi$ -VI for Laplace linear regression, measured in  $W_2$  across values of  $\lambda$ . **Right.** runtime of  $\Xi$ -VI for Laplace linear regression with a Sinkhorn error of  $10^{-5}$ , measured in seconds across values of  $\lambda$ .

With this simulated data, we use Algorithm 1 to compute the  $\Xi$ -VI approximation. In the first step, we use expectation propagation (EP) to compute the pseudomarginals. For the analysis, we chose 100  $\lambda$  values on a logarithmic scale from  $10^{-4}$  to  $10^6$ , and represented the variational posterior for each  $\lambda$  by 2,000 sample points. Figure 2(a) shows the approximation errors of  $\Xi$ -VI as a function of  $\lambda$ , measured using the Wasserstein distance ( $W_2$ ). These distances are computed between the posterior distributions sampled via MCMC and those obtained from  $\Xi$ -VI. The  $\Xi$ -VI approximation errors are benchmarked against the baseline errors of EP at  $\lambda = \infty$ , mean-field VI, and the theoretical lower bound at  $\lambda = 0$ . A vertical line at  $\lambda = D$ , the number of features, marks an inflection point where the posterior variational approximation error transitions from rapidly converging to the EP error ( $\lambda \leq D$ ) to relatively stable ( $\lambda > D$ ).

Figure 2(b) reports the runtime of the approximate coordinate-ascent algorithm for Laplace linear regression, measured in seconds until convergence. The  $\lambda$  values are displayed on a logarithmic scale to highlight the performance over several orders of magnitude. The runtime decreases sharply for  $\lambda \leq D$  and stabilizes once  $\lambda > D$ . The inflection in both the approximation error and runtime plots suggests that a regularization strength around  $\lambda = D$  offers a balanced tradeoff between approximation accuracy and computational complexity.

### 3.3 Hierarchical Model

The 8-schools model (Gelman et al. (1995), Section 5.5) is a classical example of a hierarchical Bayesian model. Each of the 8 schools run a randomized trial to assess the effect of tutoring on a standardized test. Each school provides separate estimates for the mean  $y_i$  and standard deviation  $\sigma_i$  of their respective treatment effects.

Let  $\theta_j$  be the treatment effect in school  $j$ . We treat the outcomes from each school as independent:

$$\begin{aligned} y_j | \theta_j &\sim N(\theta_j, \sigma_j^2), \quad \theta_j | \mu, \tau \sim \mathcal{N}(\mu, \tau^2), \quad 1 \leq j \leq 8, \\ \mu &\sim N(0, 5), \quad \tau \sim \text{halfCauchy}(0, 5). \end{aligned} \quad (3.3)$$

where  $\mu$  and  $\tau$  are the global parameters common to all schools,  $\theta_j$  is a local parameter specific to school  $j$ . The target of posterior inference are  $\{\theta_j\}_{j=1}^8$ ,  $\mu$  and  $\tau^2$ .

To match the  $\Xi$ -VI formulation in Section 2.2, we define  $z_j := (\theta_j - \mu)/\tau$  and rewrite the model as follows:

$$\begin{aligned} y_j | \mu, z_j, \tau &\sim N(\mu + \tau z_j, \sigma_j^2), \\ z_j &\sim N(0, 1), \quad \mu \sim N(0, 5), \quad \tau \sim \text{halfCauchy}(0, 5). \end{aligned} \quad (3.4)$$

This reparameterization transforms the joint prior of  $z_j$ 's,  $\mu$ , and  $\tau$  into a product distribution.

We apply Algorithm 1 to solve the  $\Xi$ -VI problem for this model, expressed as:

$$q_\lambda^* \in \arg \min_{q(z, \mu, \tau)} \mathbb{E}_q \left[ \sum_{j=1}^8 \frac{(y_j - \tau z_j - \mu)^2}{2\sigma_j^2} \right] + D_{\text{KL}}(q \parallel \pi) + \lambda \Xi(q). \quad (3.5)$$

In the first step, we use automatic differentiation variational inference (ADVI, (Kucukelbir et al., 2017; Carpenter et al., 2015)) to compute a set of pseudomarginals. In the second step, we use the Sinkhorn algorithm to solve the EOT problem:

$$q_\lambda^* = \arg \min_{q(z, \mu, \tau) \in \mathcal{C}(q_\infty^*)} \mathbb{E}_q \left[ \sum_{j=1}^8 \frac{\tau^2 z_j^2 + 2(\mu - y_j)\tau z_j}{2\sigma_j^2} \right] + (\lambda + 1) D_{\text{KL}}(q \parallel \hat{m}_\lambda). \quad (3.6)$$

The problem (3.6) is solvable in polynomial time using the Sinkhorn algorithm, as detailed in Proposition 8 of Section A. Assumption 1 of Proposition 8 is upheld due to efficient storage of the cost tensor as third-order tensors. Ultimately, we derive the joint distribution  $q_\lambda^*(\theta_1, \dots, \theta_8, \mu, \tau)$  by setting  $\theta_j = \mu + \tau z_j$  based on the optimal coupling in (3.6).

To benchmark the performance of our VI methods, we compute the true posterior using MCMC draws with 4 chains for 1000 tune and 5000 draw iterations. For each of the VI methods, we represent the approximate posterior with 10,000 sample points.

$\Xi$ -VI captures the dependency among the variables in the posterior. Figure 3 compares the strength of association between  $\theta_1$  and  $\theta_7$  under the true posterior, mean-field variational posterior and  $\Xi$ -variational posteriors when  $\lambda \in \{0, 1, 10, 1000\}$ . The true posterior shows a strong positive correlation between  $\theta_1$  and  $\theta_7$ , which is effectively captured by  $\Xi$ -VI at small  $\lambda$ . As  $\lambda$  increases, the correlation decreases to the MFVI level that attains a slope estimate of 0.19.

$\Xi$ -VI excels in inference that involves multiple variables in the posterior. Figure 4 illustrates credible intervals for maximum and minimum treatment effects across schools, comparing  $\Xi$ -VI with MFVI, normalizing flow variational inference (NFVI), Stein variational gradient descent (SVGD), and full-rank ADVI.  $\Xi$ -VI achieves more accurate interval width and coverage accuracy for both max and min treatment effects compared to other VI methods. Specifically, for the maximum treatment effect, while MFVI, NFVI, and full-rank ADVI produce overly large or small intervals, SVGD results in overly small intervals.

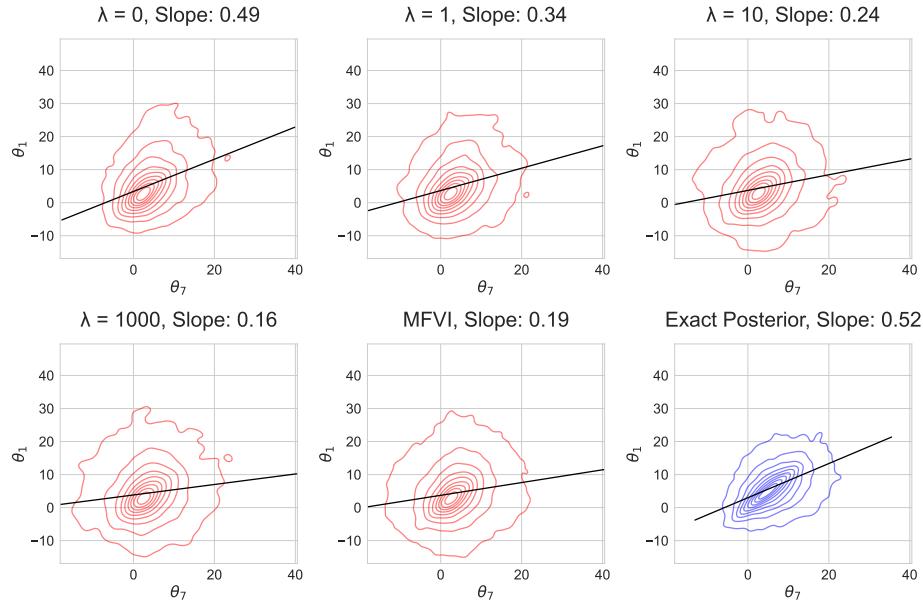


Figure 3: Contour plots for the joint distribution of  $\theta_1$  and  $\theta_7$  across various variational approximation of the Eight School model. The subplots compare the exact posterior distribution with  $\Xi$ -variational posteriors for varying  $\lambda$  values, and the MFVI approximation. A linear regression fitted slope of  $\theta_7$  over  $\theta_1$  is provided for each subplot. Each subplot includes a linear regression line showing the fitted slope of  $\theta_7$  over  $\theta_1$ .

In contrast,  $\Xi$ -VI closely approximates the true 95% posterior credible interval. For the minimum treatment effect, none of the VI methods precisely capture the true posterior interval. MFVI, NFVI, and full-rank ADVI produce intervals with a downward-shifted center, SVGD offers considerably undersized intervals, and  $\Xi$ -VI generates reasonably-sized intervals with less downward shift compared to MFVI.

Now we show the computation–statistical tradeoff of  $\Xi$ -VI in the 8-schools model. We evaluate our procedure on 100  $\lambda$  values logarithmically spaced from  $10^{-3}$  to  $10^5$ . Figure 5(a) illustrates the approximation errors of the  $\Xi$ -variational posterior relative to the exact posterior, measured using KL divergence and  $W_2$  distance. These errors are benchmarked against those of MFVI at  $\lambda = \infty$  and a theoretical lower bound at  $\lambda = 0$ . A vertical line at  $\lambda = D = 10$  marks a critical transition: errors remain relatively stable for  $\lambda < 1$  and approach MFVI for  $\lambda \geq 100$ . Notably, the normalizing-flow VI also performs reasonably well for this model and matches the performance of  $\Xi$ -VI at  $\lambda = 1$  in  $W_2$  distance and at  $\lambda = D$  in KL distance. Figure 5(b) shows the runtime of Algorithm 1 for the 8-schools model, measured in seconds to reduce the Sinkhorn error (Algorithm 2) below  $10^{-5}$ . The regularization strength  $\lambda$  is plotted on a logarithmic scale. The plot shows a sharp decline right before and right after  $\lambda = D$ . The phase transition in both plots confirms that a choice of  $\lambda = D$  offers a balance in the tradeoff between computational efficiency and approximation accuracy. However, a computational-statistical gap exists in this model: while  $\lambda < 1$  yields a closer approximation to the exact posterior, optimal runtime is only achieved for  $\lambda > 10$ .

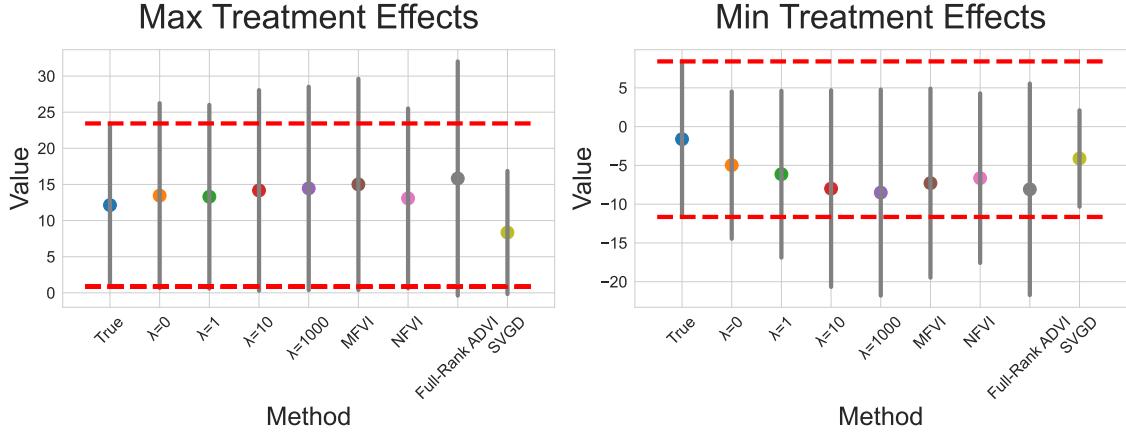


Figure 4: Comparison of the 95% posterior credible intervals for the maximum and minimum treatment effects across schools in the Eight School model. The sequence from left to right includes the exact posterior,  $\Xi$ -VI with  $\lambda \in \{0, 1, 10, 1000\}$ , MFVI, normalizing flow (NFVI), full-rank ADVI (Full-rank ADVI), and Stein variational gradient descent (SVGD).

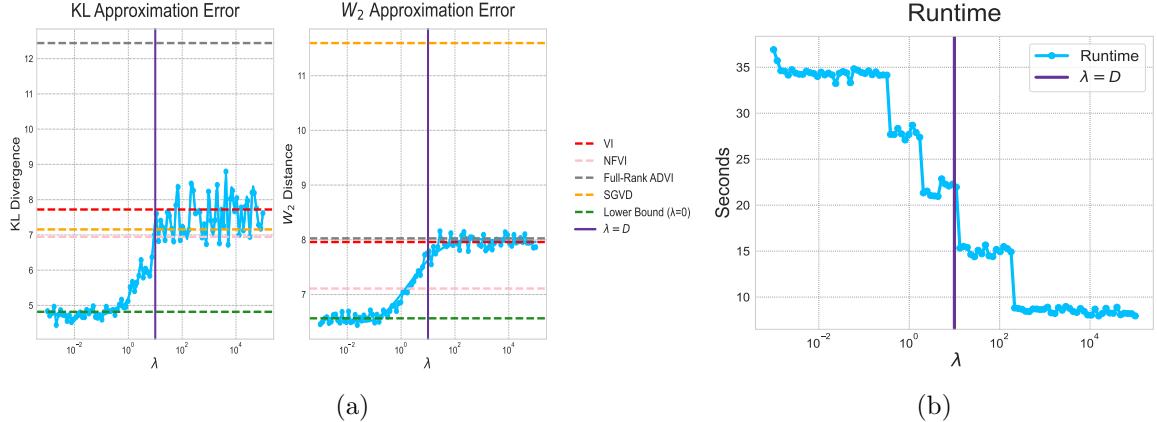


Figure 5: **Left.** approximation accuracy for the Eight School model of  $\Xi$ -VI across varying  $\lambda$  compared with other VI methods, measured in KL divergence and  $W_2$  distance. **Right.** runtime for the Eight School model as a function of varying  $\lambda$  with a Sinkhorn error of  $10^{-5}$ , measured in seconds.

Finally, note that in the 8-schools model, the MFVI produces overdispersed results after we apply the reparametrization. Generally, we recommend using overdispersed pseudomarginals in Algorithm 1. The advantage comes from an intuitive understanding of the one-step EOT correction: it seeks overlaps between the pseudomarginals and the exact posterior to effectively capture the dependency information present in the exact posterior. When the pseudomarginals are underdispersed, the one-step EOT correction still leads to underdispersed samples. With overdispersed pseudomarginals, the one-step EOT coupling compensates for

the overdispersion by subsampling points from the marginals that reflect the dependency structure of the exact posterior distributions, as seen in Figure 3.

## 4 Asymptotic Theory

In this section, we study the asymptotic theory of  $\Xi$ -variational posterior  $q_{\lambda_n}^*$  in two regimes.

- In the regime where the parameter dimension  $D$  grows with  $n$  (Section 4.1), we provide asymptotic approximation guarantees for  $q_{\lambda_n}^*$  and identify the range of  $\lambda_n$  for which  $\Xi$ -VI matches either exact posterior inference or a product distribution, under the assumption of a *compact parameter space*.
- In the finite-dimensional regime (Section 4.2), where  $D$  is held fixed, we establish the full set of frequentist guarantees for  $\Xi$ -VI, including posterior consistency and a Bernstein–von Mises (BvM) theorem. Under standard regularity conditions and a potentially unbounded parameter space, we show that  $q_{\lambda_n}^*$  converges to a Gaussian distribution as  $n \rightarrow \infty$ .

We set some notations for the theory. For two positive sequences  $a_n$  and  $b_n$ , we write  $a_n \lesssim b_n$  or  $a_n = O(b_n)$  or  $b_n \gtrsim a_n$  if there exists a constant  $C > 0$  such that  $a_n \leq Cb_n$  for all  $n$ . The constant  $C$  does not depend on  $n$ . The relation  $a_n \asymp b_n$  holds if both  $a_n \lesssim b_n$  and  $b_n \lesssim a_n$  are true. We write  $a_n \prec b_n$  or  $a_n = o(b_n)$  if  $a_n \leq c_n b_n$  for all  $n$ , for some sequence  $c_n$  that converges to zero,  $c_n \rightarrow 0$ . We write  $a_n \succ b_n$  if  $b_n = o(a_n)$ .

Let  $\mathbb{P}(\Theta)$  denote the set of probability measures on the Euclidean metric space  $(\Theta, \|\cdot\|)$  that admit Lebesgue densities and  $\mathbb{P}_p(\Theta) := \{q \in \mathbb{P}(\Theta) : \mathbb{E}_q[\|\theta\|^p] < \infty\}$ . For  $p \geq 1$ , the ( $p^{\text{th}}$ )-Wasserstein distance is defined as  $W_p(q_0, q_1) := (\inf_{q \in \mathcal{C}(q_0, q_1)} \mathbb{E}_q[\|X - Y\|^p])^{1/p}$ . The space  $(\mathbb{P}_2(\Theta), W_2)$  forms a metric space (Villani, 2009). We use  $\text{BW}(\mathbb{R}^D)$  to denote the subspace of  $\mathbb{P}_2(\mathbb{R}^D)$  consisting of Gaussian distributions, known as the Bures–Wasserstein space (Bhatia et al., 2019). We use  $\mathcal{C}^2(\Theta)$  to denote the space of twice continuously differentiable functions on  $\Theta$ .

**Assumption 1 (Standing Assumptions)** *Let  $(\Theta, \|\cdot\|)$  be a metric space where  $\Theta \subseteq \mathbb{R}^D$  is equipped with the Euclidean norm. The prior  $\pi$  admits a Lebesgue density of the form  $\pi(\theta) = \exp(\nu_0(\theta))$  with  $\nu_0 \in \mathcal{C}^2(\Theta)$ , and the exact posterior  $q_0^*$  lies in the Wasserstein space  $(\mathbb{P}_2(\Theta), W_2)$ .*

We make explicit the dependence on  $n$  of the regularizer  $\lambda_n$  and the data  $\mathbf{x}^{(n)}$ . Under this setup, the  $\Xi$ -variational posterior is given by

$$q_{\lambda_n}^* = \arg \min_{q \in \mathbb{P}_2(\Theta)} \left[ -\ell(\mathbf{x}^{(n)}; \theta) \right] + D_{\text{KL}}(q \parallel \pi) + \lambda_n \Xi(q). \quad (4.1)$$

### 4.1 Asymptotic Approximation Guarantees with Growing Dimension

To explain the changes in approximation accuracy observed in Figure 2(a) and Figure 5(a), we develop asymptotic approximation guarantees for the  $\Xi$ -variational posterior  $q_{\lambda_n}^*$  in the regime where  $D_n \rightarrow \infty$  as  $n \rightarrow \infty$ . Our primary goal is to characterize the scaling behavior of the regularization parameter  $\lambda_n$  under which  $q_{\lambda_n}^*$  either behaves like a product distribution or like the exact posterior. We begin with a theorem for general models with compact parameter space and then focus on the high-dimensional linear model.

**General Case.** For any function  $f : \Theta \rightarrow \mathbb{R}$ , let  $\omega_\Theta(f) := \sup_{\theta \in \Theta} f(\theta) - \inf_{\theta \in \Theta} f(\theta)$  denote its oscillation. We now state the main result.

**Theorem 2** Suppose that Assumption 1 holds, and let  $\Theta = [-1, 1]^D$ . Assume further that  $\ell(\mathbf{x}^{(n)}; \cdot) \in \mathcal{C}^2(\Theta)$ . Define

$$a := \omega_\Theta(\ell(\mathbf{x}^{(n)}; \theta)), \quad b_i := \omega_\Theta([\nabla \ell(\mathbf{x}^{(n)}; \theta)]_i), \quad c_{ij} := \begin{cases} \omega_\Theta([\nabla^2 \ell(\mathbf{x}^{(n)}; \theta)]_{ij}) & \text{for } i = j, \\ \sup_{\theta \in \Theta} |[\nabla^2 \ell(\mathbf{x}^{(n)}; \theta)]_{ij}| & \text{for } i \neq j. \end{cases} \quad (4.2)$$

When  $\lambda_n \succ D^{-1/2} \max \left( \sqrt{a \sum_{i=1}^D c_{ii}}, \sqrt{\sum_{i=1}^D b_i^2}, \sqrt{\sum_{i=1}^D \sum_{j=1}^D c_{ij}^2}, D^{1/2} \right)$ , there exists a sequence of product distributions  $m_{\lambda_n}^*$  such that, for any 1-Lipschitz function  $\psi : \mathbb{R} \mapsto \mathbb{R}$ , as  $n \rightarrow \infty$ , the  $\Xi$ -VI optimizer  $q_{\lambda_n}^*$  satisfies

$$\sup_{\mathbf{x}^{(n)} \in \mathbb{X}^n} \left| \frac{1}{D} \sum_{i=1}^D \left( \mathbb{E}_{q_{\lambda_n}^*} [\psi(\theta_i)] - \mathbb{E}_{m_{\lambda_n}^*} [\psi(\theta_i)] \right) \right| \xrightarrow{P_{\theta_0}} 0. \quad (4.3)$$

When  $\lambda_n \prec D\Xi^{-1}(q_0^*)$ , for any 1-Lipschitz function  $\psi : \mathbb{R} \mapsto \mathbb{R}$ , as  $n \rightarrow \infty$ ,

$$\sup_{\mathbf{x}^{(n)} \in \mathbb{X}^n} \mathbb{E}_{q_{\lambda_n}^*} \left[ \left( \frac{1}{D} \sum_{i=1}^D \psi(\theta_i) - \frac{1}{D} \sum_{i=1}^D \mathbb{E}_{q_0^*} [\psi(\theta_i)] \right)^2 \right] \xrightarrow{P_{\theta_0}} 0. \quad (4.4)$$

The compactness of the parameter space is essential for establishing approximation guarantees for variational inference in high dimensions. This assumption and the proofs in Section E build on the theory of nonlinear large deviations (Chatterjee and Dembo, 2016; Yan, 2020). The choice  $\Theta = [-1, 1]^D$  and corresponding tools have been adopted in the theoretical analysis of mean-field variational inference for Potts models (Basak and Mukherjee, 2017), linear models (Mukherjee and Sen, 2022; Mukherjee et al., 2023), and, in a related form, for latent variable models (Zhong et al., 2025). Our results also extend to any compact subset  $\Theta$  of  $\mathbb{R}^D$ .

As  $\theta \in [-1, 1]^D$ , all first- and second-order derivatives of  $\ell(\mathbf{x}^{(n)}; \cdot)$  are uniformly bounded, which controls the oscillation quantities  $a$ ,  $b_i$ , and  $c_{ij}$ . For unbounded parameter spaces, alternative results are available under stronger shape constraints—when both the likelihood and the prior are log-concave (Lacker et al., 2024). We do not pursue this direction here to keep the presentation focused.

When  $D$  is fixed, the convergence in Eq. (4.3) and Eq. (4.4) imply that the expectations of the averages  $\frac{1}{D} \sum_{i=1}^D \psi(\theta_i)$  under the pairs of posteriors  $(q_{\lambda_n}^*, m_{\lambda_n}^*)$  and  $(q_{\lambda_n}^*, q_0^*)$  are the same up to an asymptotically negligible error, respectively. When  $D$  is fixed and  $\Theta$  is compact, the results follow directly whenever the variational posteriors  $q_{\lambda_n}^*$ ,  $m_{\lambda_n}^*$ , and  $q_0^*$  are *uniformly (weakly) consistent* at the true parameter  $\theta_0$ , since weak consistency is equivalent to convergence in the bounded-Lipschitz metric (see Remark 6.3 of Ghosal and van der Vaart (2017)).

Eq. (4.3) defines a *mean-field regime*, where a product measure matches  $q_{\lambda_n}^*$  in any 1-Lipschitz statistic (first-order statistics). This regime characterizes when  $\Xi$ -VI can be replaced by MFVI. The critical scaling term in the threshold is  $\sqrt{\sum_{i,j} c_{ij}^2}$ , as the other

terms are typically well controlled. Roughly, the equivalence between  $\Xi$ -VI and MFVI is determined by comparing  $\lambda_n$  to the  $(D^{-1/2})$ -scaled Frobenius norm of the Fisher information.

Eq. (4.4) defines a *Bayes optimal regime*, where  $\Xi$ -VI asymptotically recovers 1-Lipschitz statistic of the exact posterior. If the dimension  $D = O(1)$  as  $n$  increases and the exact posterior achieves consistency, then  $\Xi(q_0^*)$  converges to zero, and Eq. (4.4) holds for any bounded sequence of  $\lambda_n$ . When  $D$  grows with  $n$  but at a slow rate (e.g.  $D \lesssim n^{-1/3}$ ), we may still expect a form of posterior consistency to hold the Bayesian optimal regime to contain non-trivial choices of  $\lambda_n$ .

To match the computational complexity in Section A, we provide sufficient conditions for  $\lambda_n \succ D$  to be in the mean-field regime.

**Corollary 3** *In the setting of Theorem 2, if  $a \lesssim D, b_i \lesssim D, c_{ii} \lesssim D$  for  $i \in [D]$  and  $c_{ij} \lesssim 1$  for  $i \neq j$ , then for  $\lambda_n \succ D$ , there exists a product distribution  $m_{\lambda_n}^*$  such that, for any 1-Lipschitz function  $\psi : \mathbb{R} \mapsto \mathbb{R}$ , as  $n \rightarrow \infty$ ,*

$$\sup_{\mathbf{x}^{(n)} \in \mathbb{X}^n} \left| \frac{1}{D} \sum_{i=1}^D \left( \mathbb{E}_{q_{\lambda_n}^*} [\psi(\theta_i)] - \mathbb{E}_{m_{\lambda_n}^*} [\psi(\theta_i)] \right) \right| \xrightarrow{P_{\theta_0}} 0. \quad (4.5)$$

The result establishes the asymptotic equivalence between  $q_{\lambda_n}^*$  and a product measure for  $\lambda \succ D$ , which provides the following computational insight: when  $\lambda_n$  is large, the  $\Xi$ -variational posterior can be replaced by the mean-field approximation to the posterior. To meet the assumptions of Corollary 3, it suffices that 1) the gradient and diagonal Hessian of the log-likelihood scale slower than  $D$  entry-wise and 2) the off-diagonal Hessian is uniformly bounded.

Now we consider the example of high-dimensional linear regression models.

**High-Dimensional Linear Model.** We observe  $\{(x_i, y_i) : 1 \leq i \leq n\}$ ,  $y_i \in \mathbb{R}$ ,  $x_i \in \mathbb{R}^D$ . Let  $\mathbf{y} = [y_1, \dots, y_n] \in \mathbb{R}^n$  and  $\mathbf{X} = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times D}$ . We consider a high-dimensional Bayesian linear regression model where both  $n, D$  are tending to infinity:

$$\mathbf{y} = \mathbf{X}\theta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n), \quad \theta \sim \pi. \quad (4.6)$$

The  $\Xi$ -VI problem for Bayesian linear model is given by

$$q_{\lambda_n}^* = \arg \min_{q \in \mathbb{P}_2(\Theta)} \mathbb{E}_q \left[ \frac{\|\mathbf{y} - \mathbf{X}\theta\|^2}{2\sigma^2} \right] + D_{\text{KL}}(q \parallel \pi) + \lambda_n \Xi(q). \quad (4.7)$$

For the matrix  $\mathbf{B} := \sigma^{-2} \mathbf{X}^\top \mathbf{X}$ , let  $\mathbf{B}_{\text{diag}}$  and  $\mathbf{B}_{\text{off}}$  denote its diagonal and off-diagonal components, respectively. Our main result characterizes the asymptotic behavior of the optimizer  $q_{\lambda_n}^*$  of Eq. (4.7).

**Theorem 4** *Suppose that Assumption 1 holds, and let  $\Theta = [-1, 1]^D$ . Assume further that there exist constants  $\kappa_1 \geq 0$  and  $\kappa_2 > 0$  such that  $\nabla^2 \nu_0 \preceq -\kappa_1 I_D$  and  $\mathbf{X}^\top \mathbf{X} \succeq \kappa_2 I_D$ . The following results hold:*

*When  $\lambda_n \succ \sqrt{\text{tr}(\mathbf{B}_{\text{off}}^2)}$ , there exists a sequence of product distributions  $m_{\lambda_n}^*$  such that, as  $n \rightarrow \infty$ ,*

$$\sup_{\mathbf{y}^{(n)} \in \mathbb{R}^n} W_2(q_{\lambda_n}^*, m_{\lambda_n}^*) \xrightarrow{P_{\theta_0}} 0. \quad (4.8)$$

When  $\lambda_n \succ \sqrt{D^{-1} \text{tr}(\mathbf{B}_{\text{off}}^2)}$ , there exists a sequence of product distributions  $m_{\lambda_n}^*$  such that, for any 1-Lipschitz function  $\psi : \mathbb{R} \mapsto \mathbb{R}$ , as  $n \rightarrow \infty$ ,

$$\sup_{\mathbf{y}^{(n)} \in \mathbb{R}^n} \mathbb{E}_{q_{\lambda_n}^*} \left[ \left( \frac{1}{D} \sum_{i=1}^D \psi(\theta_i) - \frac{1}{D} \sum_{i=1}^D \mathbb{E}_{m_{\lambda_n}^*} [\psi(\theta_i)] \right)^2 \right] \xrightarrow{P_{\theta_0}} 0. \quad (4.9)$$

When  $\lambda_n \prec (\kappa_1 + \kappa_2) [\text{tr}(\text{Cov}_{q_0^*}(\mathbf{B}_{\text{off}}\theta))]^{-1}$ , as  $n \rightarrow \infty$ ,

$$\sup_{\mathbf{y}^{(n)} \in \mathbb{R}^n} W_2(q_{\lambda_n}^*, q_0^*) \xrightarrow{P_{\theta_0}} 0. \quad (4.10)$$

When  $\lambda_n \prec D(\kappa_1 + \kappa_2) [\text{tr}(\text{Cov}_{q_0^*}(\mathbf{B}_{\text{off}}\theta))]^{-1}$ , for any 1-Lipschitz function  $\psi : \mathbb{R} \mapsto \mathbb{R}$ , as  $n \rightarrow \infty$ ,

$$\sup_{\mathbf{y}^{(n)} \in \mathbb{R}^n} \mathbb{E}_{q_{\lambda_n}^*} \left[ \left( \frac{1}{D} \sum_{i=1}^D \psi(\theta_i) - \frac{1}{D} \sum_{i=1}^D \mathbb{E}_{q_0^*} [\psi(\theta_i)] \right)^2 \right] \xrightarrow{P_{\theta_0}} 0. \quad (4.11)$$

Eq. (4.8) and Eq. (4.9) form the *mean-field regimes*. When  $\lambda_n$  scales faster than  $\sqrt{\text{tr}(\mathbf{B}_{\text{off}}^2)}$ ,  $q_{\lambda_n}^*$  converges in the Wasserstein metric to a product distribution. This means all moment statistics can be asymptotically transported between  $\Xi$ -VI and MFVI. When  $\lambda_n$  scales faster than  $\sqrt{\text{tr}(\mathbf{B}_{\text{off}}^2)/D}$ , we can transport any 1-Lipschitz statistic between  $\Xi$ -VI and MFVI. As  $\lambda_n$  increases,  $q_{\lambda_n}^*$  shares more distributional information with  $m_{\lambda_n}^*$ .

Eq. (4.10) and Eq. (4.11) define the *Bayes optimal regimes*. When  $\lambda_n$  increases more slowly than  $(\kappa_1 + \kappa_2) [\text{tr}(\text{Cov}_{q_0^*}(\mathbf{B}_{\text{off}}\theta))]^{-1}$ ,  $q_{\lambda_n}^*$  converges to the exact posterior in the Wasserstein metric. By relaxing a factor of  $D$ ,  $q_{\lambda_n}^*$  achieves asymptotic Bayes optimality for all 1-Lipschitz statistics. Since  $\text{tr}(\text{Cov}_{q_0^*}(\mathbf{B}_{\text{off}}\theta)) \leq \|\mathbf{B}_{\text{off}}\|_2 \text{tr}(\text{Cov}_{q_0^*}(\theta))$ , the Bayes optimal regime is large when  $\|\mathbf{B}_{\text{off}}\|_2$  is small. When  $\mathbf{B}_{\text{off}} = 0$ , any choice of  $\lambda_n$  falls within the Bayes optimal regime.

The log-concavity assumptions in Theorem 4 require the log-prior  $\nu_0 \in \mathcal{C}^2(\Theta)$  to be  $\kappa_1$ -strongly concave, and the log-likelihood  $\ell(\mathbf{x}^{(n)}; \cdot)$  to be in  $\mathcal{C}^2(\Theta)$  and  $\kappa_2/(2\sigma^2)$ -strongly concave. The log-concavity parameters  $\kappa_1$  and  $\kappa_2$  are not required for the mean-field regimes but appear in the Bayes-optimal regimes. Larger values of  $\kappa_1$  and  $\kappa_2$  enlarge the range of  $\lambda_n$  for which Bayes-optimal behavior is satisfied.

The mean-field regime (Eq. (4.8)) and the Bayes optimal regime (Eq. (4.10)) can both hold for sufficiently large  $\lambda_n$ . In that case, the  $\Xi$ -VI solution  $q_{\lambda_n}^*$  can be computed efficiently via MFVI while still closely approximating the exact posterior. For instance, in a linear regression setup where  $\mathbf{X}^\top \mathbf{X}$  is diagonal, the exact posterior  $q_0^*$  is itself a product measure. Then the upper bound in (4.10) is infinite, whereas the lower bound in (4.8) is zero. In this case, any choice of  $\lambda_n$  satisfies both (4.8) and (4.10) simultaneously.

For 1-Lipschitz statistics, Eq. (4.9) and Eq. (4.11) both hold if

$$\sqrt{\text{tr}(\mathbf{B}_{\text{off}}^2)} \|\mathbf{B}_{\text{off}}\|_2^2 \text{tr}(\text{Cov}_{q_0^*}(\theta)) \prec D^{3/2}(\kappa_1 + \kappa_2). \quad (4.12)$$

This criterion is satisfied, for example, when  $\text{tr}(\mathbf{B}_{\text{off}}^2) \prec D$ ,  $\|\mathbf{B}_{\text{diag}}\|_2 \lesssim 1$ , and  $\text{tr}(\text{Cov}_{q_0^*}(\theta)) \lesssim D$ , which recovers the Bayes optimal condition for MFVI (Mukherjee and Sen, 2022). But

our criterion is more flexible: for example, it is also satisfied when  $\text{tr}(\mathbf{B}_{\text{off}}^2) \lesssim D$ ,  $\|\mathbf{B} - \mathbf{B}'_{\text{diag}}\|_2 \lesssim 1$ , and  $\text{tr}(\text{Cov}_{q_0^*}(\theta)) \prec D$ .

When no choice of  $\lambda_n$  satisfies the overlap criterion, there is a gap between the mean-field and Bayes optimal regimes. Achieving accurate posterior inference thus requires paying an additional computational cost that scales inversely with  $\lambda_n$ , as discussed in Section A.

Let the eigenvalues of  $\mathbf{B}_{\text{off}}$  be  $\eta_D \geq \dots \geq \eta_1$ . Then  $\text{tr}(\mathbf{B}_{\text{off}}^2) = \sum_{i=1}^D \eta_i^2$ , and the mean-field regime Eq. (4.8) corresponds to  $\lambda_n \succ \sqrt{\sum_{i=1}^D \eta_i^2}$ . To match the complexity bound in Section A, we provide sufficient conditions for  $\lambda_n \succ D$  to be in the mean-field regime Eq. (4.8).

**Corollary 5** *Let  $\eta_D \geq \dots \geq \eta_1$  denote the eigenvalues of  $\mathbf{B}_{\text{off}}$ . In the setting of Theorem 4, if  $\sum_{i=1}^D \eta_i^2 \lesssim D^2$  as  $n \rightarrow \infty$ , then for any  $\lambda_n \succ D$  there exists  $m_{\lambda_n}^* \in \mathbb{M}(\Theta)$  such that  $\sup_{\mathbf{y}^{(n)} \in \mathbb{R}^n} W_2(q_{\lambda_n}^*, m_{\lambda_n}^*) \xrightarrow{P_{\theta_0}} 0$ .*

Data preprocessing often involves normalizing features to have unit variances. Thus, the requirement that  $\sum_{i=1}^D \eta_i^2 \lesssim D^2$  is often met in practice.

## 4.2 Asymptotic Normality in Finite Dimension

The second part of the theory deals with posterior consistency and asymptotic normality of  $\Xi$ -variational posteriors for finite-dimensional models. The asymptotic normality results state that depending on the limit of  $\lambda_n$ ,  $\Xi$ -variational posterior converges in the limit to one of three quantities: the mean-field minimizer of a normal distribution, the normal distribution itself, or a  $\Xi$ -variational normal approximation.

We consider the setting where the observations  $X_1, \dots, X_n \stackrel{iid}{\sim} P_{\theta_0}$ , where  $\theta_0$  lies in the interior of a Borel set  $\Theta \subseteq \mathbb{R}^D$ . Unlike Section 4.1, here  $\Theta$  is allowed to be unbounded. We posit the following assumptions.

**Assumption 2 (Prior Mass)** *The function  $\nu_0$  is bounded in a neighborhood of  $\theta_0$ . For some  $C > 0$ , we have*

$$\sup_{\|\theta - \theta_0\|_2 \leq Cn^{-1/2}} \|\nabla \nu_0(\theta)\|_2 \lesssim \sqrt{n}, \quad \text{and} \quad \sup_{\|\theta - \theta_0\|_2 \leq Cn^{-1/2}} \|\nabla^2 \nu_0(\theta)\|_2 \lesssim n.$$

**Assumption 3 (Consistent Testability Assumptions)** *For every  $\epsilon > 0$ , there exists a sequence of tests  $\phi_n$  such that*

$$\int \phi_n(x) p_{\theta_0}(x) dx \rightarrow 0, \quad \sup_{\theta: \|\theta - \theta_0\|_2 \geq \epsilon} \int (1 - \phi_n(x)) p_{\theta}(x) dx \rightarrow 0.$$

**Assumption 4 (Local Asymptotic Normality (LAN) Assumptions)** *For every compact set  $K \subset \mathbb{R}^D$ , there exist random vectors  $\Delta_{n,\theta_0}$  bounded in probability and a nonsingular matrix  $V_{\theta_0}$  such that*

$$\sup_{h \in K} \left| \ell(\mathbf{x}^{(n)}; \theta_0 + \delta_n h) - \ell(\mathbf{x}^{(n)}; \theta_0) - h^T V_{\theta_0} \Delta_{n,\theta_0} + \frac{1}{2} h^T V_{\theta_0} h \right| \xrightarrow{P_{\theta_0}} 0,$$

where  $\delta_n$  is a  $D \times D$  diagonal matrix whose entries converge to 0 as  $n \rightarrow \infty$ . For  $D = 1$ , we commonly take  $\delta_n = n^{-1/2}$ .

The first assumption ensures the prior is light-tailed. It is satisfied by, for example, the flat prior or a sub-Gaussian prior.

The second assumption guarantees the existence of a sequence of uniformly consistent tests for  $H_0 : \theta = \theta_0$  versus  $H_1 : \|\theta - \theta_0\|_2 \geq \epsilon$  based on the data. This condition is satisfied if there exists a consistent sequence of estimators  $T_n$  for  $\theta$  and we set  $\phi_n(x) := I\{T_n - \theta \geq \epsilon/2\}$ , or when the Hellinger distance between  $\{p_\theta : \|\theta - \theta_0\|_2 \geq \epsilon\}$  and  $p_{\theta_0}$  is lower bounded by some positive constant  $\delta$  (Ghosal and van der Vaart, 2017).

The third assumption states that the log-likelihood is locally well-approximated (up to a vanishing error) by that of a normal location model centered at  $\theta_0$  under an appropriate rescaling. The rescaling sequence  $\delta_n$  is exactly the posterior contraction rate. In standard finite-dimensional, correctly specified models, we typically have  $\delta_n = n^{-1/2}$  (Ghosal and van der Vaart, 2017).

In line with Assumption 4, we consider a change of variable:

$$h := \delta_n^{-1}(\theta - \theta_0 - \delta_n \Delta_{n,\theta_0}), \quad \text{for } \theta \sim q_{\lambda_n}^*. \quad (4.13)$$

Our first result states that under this change of variable,  $\Xi$ -variational posterior satisfies a Bernstein-von-Mises phenomenon with a phase transition.

**Theorem 6 (Bernstein von-Mises Theorem)** *Suppose that Assumptions 1 to 4 hold, and let  $\theta_0$  be an interior point of a Borel set  $\Theta$ . Let  $\tilde{q}_{\lambda_n}$  be the distribution of the rate-adjusted parameter  $h$  defined in Eq. (4.13). The distribution  $\tilde{q}_{\lambda_n}$  converges in the Wasserstein metric to a normal distribution under the following three regimes:*

1. If  $\lambda_n \rightarrow \infty$ , then  $W_2(\tilde{q}_{\lambda_n}, N(0, ((V_{\theta_0})_{\text{diag}})^{-1})) \xrightarrow{P_{\theta_0}} 0$ , where  $(V_{\theta_0})_{\text{diag}}$  is the diagonal submatrix of  $V_{\theta_0}$ .
2. If  $\lambda_n \rightarrow 0$ , then  $W_2(\tilde{q}_{\lambda_n}, N(0, V_{\theta_0}^{-1})) \xrightarrow{P_{\theta_0}} 0$ .
3. If  $\lim_{n \rightarrow \infty} \lambda_n = \lambda_\infty$  for some  $\lambda_\infty \in (0, \infty)$ , then

$$W_2\left(\tilde{q}_{\lambda_n}, \arg \min_{q \in \mathbb{P}_2(\mathbb{R}^D)} D_{KL}(q \parallel N(0, V_{\theta_0}^{-1})) + \lambda_\infty \Xi(q)\right) \xrightarrow{P_{\theta_0}} 0.$$

The result aligns well with intuition. When  $\lambda_n$  diverges,  $q_{\lambda_n}^*$  converges to the mean-field approximation. When  $\lambda_n$  approaches zero, the constraint set in the Lagrangian dual problem increases to include the true limiting posterior  $N(0, V_{\theta_0}^{-1})$ . When  $\lambda_n$  converges to some finite value  $\lambda_\infty$ , the  $\Xi$ -variational posterior converges to the Gaussian limit of the exact posterior. In the regime where  $\lim_{n \rightarrow \infty} \lambda_n$  does not exist but  $\lambda_n = O(1)$ , the  $\Xi$ -variational posterior converges to a “biased” estimate of the true Gaussian posterior  $N(0, V_{\theta_0}^{-1})$  along a subsequence of  $\lambda_n$  that converges as  $n \rightarrow \infty$ .

The Bernstein von-Mises Theorem implies the (weak) posterior consistency for  $q_{\lambda_n}^*$ .

**Corollary 7** *Under the assumptions of Theorem 6,  $\Xi$ -variational posterior is consistent in  $[P_{\theta_0}]$ -probability in the sense that  $W_2(q_{\lambda_n}^*, \delta_{\theta_0}) \xrightarrow{P_{\theta_0}} 0$ .*

The convergence in Corollary 7 is stated in the Wasserstein metric, which is slightly stronger than the typical metric used in posterior consistency results. The convergence in the Wasserstein metric is equivalent to weak convergence plus the convergence of the second moments (Theorem 5.11, Santambrogio (2015)). Thus, posterior consistency and the Bernstein–von Mises theorem (Theorem 6) can be framed in terms of the weak convergence and  $L_2$  convergence for the corresponding measures.

## 5 Discussion

We introduced  $\Xi$ -VI, a new way of performing variational inference that extends MFVI through entropic regularization. We characterized the asymptotic normality of  $\Xi$ -variational posteriors in low-dimensional scenarios and analyzed the tradeoff between computational complexity and statistical fidelity in high-dimensional settings. On both simulated and empirical datasets, we demonstrated its advantages over traditional MFVI. Further, our method explicitly connects VI to entropic optimal transport, using the Sinkhorn algorithm to improve the fidelity of variational approximations.

One question prompted by our work is to understand the fundamental limits of high-dimensional Bayesian models. It is known that many high-dimensional problems exhibit a gap between what is statistically achievable (in a minimax sense) and what is achievable by a polynomial-time algorithm, such as sparse PCA (Wang et al., 2016) and matrix denoising (Chandrasekaran and Jordan, 2013). However, characterizing a *statistical-computational gap* is a newer topic in probabilistic machine learning.

The theoretical results in Section 4 identify distinct asymptotic regimes that correspond to the exact posterior and the mean-field approximation. The transition between these regimes echoes classical phase transitions in spin glass models (Montanari and Sen, 2024). While our analysis focuses on a regression setting, similar techniques could be extended to models such as the Ising model or the quadratic interaction model. It would also be interesting to explore the connection between  $\Xi$ -VI and the rich literature on PAC-Bayes and generalized Bayes learning (Alquier and Guedj, 2018; Alquier, 2021; Husain and Knoblauch, 2022; Wild et al., 2022, 2023), and to investigate how recent results in PAC-Bayes theory (Alquier, 2024) interact with the observed phase transitions—especially given that the target distribution is a general Gibbs measure. Another promising direction is to extend the analysis in Section 4.1 beyond compact  $\Theta$  to unbounded parameter spaces, for example using the tools of Lacker et al. (2024), which require strong log-concavity of the exact posterior.

Another challenge is to solve  $\Xi$ -VI efficiently for models with high-dimensional parameters. The difficulty lies in implementing the multimarginal Sinkhorn algorithm efficiently for large  $D$  in polynomial time. A stochastic approximation of the Sinkhorn step in Algorithm 1 could help mitigate these computational costs. Future work may focus on developing user-friendly algorithmic tools to enable scalable applications of  $\Xi$ -VI.

While this paper implemented the examples using EP and BBVI for approximating posterior marginals, advanced mean-field methods such as the TAP approach may be preferable in certain contexts, such as spiked covariance models (Fan et al., 2021) and high-dimensional Bayesian linear regression (Celentano et al., 2023a). Exploring  $\Xi$ -VI combined

with the TAP method is a promising avenue for future research, potentially yielding more accurate approximations of the true posterior.

### Acknowledgments

David M. Blei was supported by NSF IIS-2127869, NSF DMS-2311108, ONR N00014-24-1-2243, and the Simons Foundation. We thank Sumit Mukherjee, Marcel Nutz, Yixin Wang, Eli Weinstein, Kaizheng Wang, and Chenyang Zhong for helpful discussions. We also thank the three anonymous reviewers for their many insightful comments and pointers that helped to improve the paper.

## References

Laurence Aitchison. A statistical theory of cold posteriors in deep neural networks. In *9th International Conference on Learning Representations*, 2021.

Pierre Alquier. Non-exponentially weighted aggregation: regret bounds for unbounded loss functions. In *International Conference on Machine Learning*, pages 207–218, 2021.

Pierre Alquier. User-friendly Introduction to PAC-Bayes Bounds. *Foundations and Trends in Machine Learning*, 17, 2024.

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107, 2018. ISSN 15730565. .

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *Annals of Statistics*, 48(3):1475–1497, 2020. ISSN 0090-5364,2168-8966. .

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17:Paper No. 239, 41, 2016.

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. *Advances in Neural Information Processing Systems*, 30, 2017.

Jason M Altschuler and Enric Boix-Adsera. Hardness results for multimarginal optimal transport problems. *Discrete Optimization*, 42:100669, 2021.

Jason M Altschuler and Enric Boix-Adsera. Polynomial-time algorithms for multimarginal optimal transport problems with structure. *Mathematical Programming*, 199:1107–1178, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and In the Space of Probability Measures*. Springer Science & Business Media, 2005.

Anirban Basak and Sumit Mukherjee. Universality of the mean-field for the Potts model. *Probability Theory and Related Fields*, 168:557–600, 2017.

James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media, 2013. ISBN 147574286X.

Kush Bhatia, Nikki Lijing Kuang, Yi-An Ma, and Yixin Wang. Statistical and computational trade-offs in variational inference: A case study in inferential model selection. *arXiv preprint arXiv:2207.11208*, 2022.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37:165–191, 2019.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. Bayesian fractional posteriors. *Annals of Statistics*, 47:39–66, 2019.

Anirban Bhattacharya, Debdeep Pati, and Yun Yang. On the convergence of coordinate ascent variational inference. *Annals of Statistics*, 53(3):929–962, 2025.

Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Annals of Statistics*, 41(4):1922–1943, 2013.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78:1103–1130, 2016.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017.

Andrea Braides. *Local Minimization, Variational Evolution and  $\Gamma$ -Convergence*, volume 2094. Springer, 2014.

Guillaume Carlier. On the linear convergence of the multimarginal Sinkhorn algorithm. *SIAM Journal on Optimization*, 32:786–794, 2022.

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 2015.

Michael Celentano, Zhou Fan, Licong Lin, and Song Mei. Mean-field variational inference with the TAP free energy: Geometric and statistical properties in linear models. *arXiv preprint arXiv:2311.08442*, 2023a.

Michael Celentano, Zhou Fan, and Song Mei. Local convexity of the TAP free energy and AMP convergence for  $Z_2$ -synchronization. *Annals of Statistics*, 51:519–546, 2023b.

Venkat Chandrasekaran and Michael I Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110:E1181–E1190, 2013.

S. Chatterjee and A. Dembo. Nonlinear large deviations. *Advances in Mathematics*, 299:396–450, 2016.

Giovanni Conforti and Luca Tamanini. A formula for the time derivative of the entropic cost and applications. *Journal of Functional Analysis*, 280:108964, 2021.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, pages 146–158, 1975.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein space. In *International Conference on Machine Learning*, pages 7960–7991, 2023.

Stephan Eckstein and Marcel Nutz. Quantitative stability of regularized optimal transport and convergence of Sinkhorn’s algorithm. *SIAM Journal on Mathematical Analysis*, 54: 5922–5948, 2022.

Jiaojiao Fan, Isabel Haasler, Johan Karlsson, and Yongxin Chen. On the complexity of the optimal transport problem with graph-structured cost. *International Conference on Artificial Intelligence and Statistics*, pages 9147–9165, 2022.

Zhou Fan, Song Mei, and Andrea Montanari. TAP free energy, spin glasses and variational inference. *Annals of Probability*, 49, 2021.

David T Frazier, Jeremias Knoblauch, Jack Jewson, and Christopher Drovandi. Exact sampling of Gibbs measures with estimated losses. *arXiv preprint arXiv:2404.15649*, 2024.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 1995.

Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An instability in variational inference for topic models. *International Conference on Machine Learning*, pages 2221–2231, 2019.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press, 2017.

Ryan Giordano, Tamara Broderick, and Michael I Jordan. Covariances, robustness and variational Bayes. *Journal of Machine Learning Research*, 19, 2018.

Peter Grünwald. The safe Bayesian: learning the learning rate via the mixability gap. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.

Peter Grünwald and Thijs van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12, 2017.

Peter Hall, John T Ormerod, and M P Wand. Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, pages 369–389, 2011a.

Peter Hall, Tung Pham, Matt P Wand, and Shen SJ Wang. Asymptotic normality and valid inference for Gaussian variational approximation. *Annals of Statistics*, pages 2502–2532, 2011b.

Hisham Husain and Jeremias Knoblauch. Adversarial interpretation of Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 553–572, 2022.

Yiheng Jiang, Sinho Chewi, and Aram-Alexandre Pooladian. Algorithms for mean-field variational inference via polyhedral optimization in the Wasserstein space. *Foundations of Computational Mathematics*, pages 1–52, 2025.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

Kyurae Kim, Jisu Oh, Kaiwen Wu, Yian Ma, and Jacob R Gardner. On the convergence of black-box variational inference. *Advances in Neural Information Processing Systems*, 36:1–2, 2023.

Jeremias Knoblauch. Frequentist consistency of generalized variational inference. *arXiv:1912.04946*, 2019.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on Bayes rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23:1–109, 2022.

Alexey Kroshnin, Nazarii Tupitsa, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Cesar Uribe. On the complexity of approximating Wasserstein barycenters. *International Conference on Machine Learning*, pages 3530–3540, 2019.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:430–474, 2017.

Daniel Lacker, Sumit Mukherjee, and Lane Chun Yeung. Mean field approximations via log-concavity. *International Mathematics Research Notices*, 2024:6008–6042, 2024. ISSN 1073-7928.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35, 5 2022.

Tianyi Lin, Nhat Ho, Marco Cuturi, and Michael I Jordan. On the complexity of approximating multimarginal optimal transport. *Journal of Machine Learning Research*, 23:1–43, 2022.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, pages 2378–2386, 2016.

F Locatello, R Khanna, and J Ghosh. Boosting variational inference: an optimization perspective. *International Conference on Artificial Intelligence and Statistics*, 2018.

Yann McLatchie, Edwin Fong, David T Frazier, and Jeremias Knoblauch. Predictive performance of power posteriors. *Biometrika*, page asaf034, 2025.

Andrew C Miller, Nicholas J Foti, and Ryan P Adams. Variational boosting: Iteratively refining posterior approximations. *International Conference on Machine Learning*, pages 2420–2429, 2017.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22:161–168, 2021.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

Thomas P Minka. Expectation propagation for approximate Bayesian inference. *arXiv preprint arXiv:1301.2294*, 2013.

Andrea Montanari and Subhabrata Sen. A Friendly Tutorial on Mean-Field Spin Glass Techniques for Non-Physicists. *Foundations and Trends in Machine Learning*, 17, 2024. ISSN 19358245. .

Soumendu Sundar Mukherjee and Purnamrita Sarkar. Mean field for the stochastic blockmodel: Optimization landscape and convergence issues. *Advances in Neural Information Processing Systems*, 2018.

Sumit Mukherjee and Subhabrata Sen. Variational inference in high-dimensional linear regression. *Journal of Machine Learning Research*, 23(304):1–56, 2022.

Sumit Mukherjee, Bodhisattva Sen, and Subhabrata Sen. A mean field approach to empirical Bayes estimation in high-dimensional linear regression. *arXiv preprint arXiv:2309.16843*, 2023.

Marcel Nutz. *Introduction to Entropic Optimal Transport*. Lecture notes, Columbia University, 2021.

Manfred Opper and David Saad. *Advanced Mean Field Methods: Theory and Practice*. MIT press, 2001.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173:361–400, 2000.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Sean Plummer, Debdeep Pati, and Anirban Bhattacharya. Dynamics of coordinate ascent variational inference: A case study in 2D Ising models. *Entropy*, 22:1263, 2020.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge university press, 2025.

Jiaze Qiu. Sub-optimality of the naive mean field approximation for proportional high-dimensional linear regression. *Advances in Neural Information Processing Systems*, 36, 2024.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. *International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. *International Conference on Machine Learning*, pages 324–333, 2016.

Kolyan Ray and Botond Szabó. Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281, 2022. ISSN 0162-1459,1537-274X. .

Kolyan Ray, Botond Szabó, and Gabriel Clara. Spike and slab variational Bayes for high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 33: 14423–14434, 2020.

Jack Raymond and Federico Ricci-Tersenghi. Improving variational methods via pairwise linear response identities. *Journal of Machine Learning Research*, 18, 2017. ISSN 15337928.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *32nd International Conference on Machine Learning*, 2, 2015.

C Robert and G Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 2004.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 55. Springer, 2015.

Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. *Advances in Neural Information Processing Systems*, 28, 2015.

Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. *Advances in Neural Information Processing Systems*, 30, 2017.

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2014.

C Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Martin J Wainwright, Michael I Jordan, et al. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Bo Wang and D M Titterington. Lack of consistency of mean field and variational Bayes approximations for state space models. *Neural Processing Letters*, 20:151–170, 2004.

Tengyao Wang, Quentin Berthet, Richard J Samworth, et al. Statistical and computational trade-offs in estimation of sparse principal components. *Annals of Statistics*, 44:1896–1930, 2016.

Yixin Wang and David M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

Florian Wenzel, Kevin Roth, Bastiaan S. Veeling, Jakub Świk atkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the Bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, 2020.

Veit D. Wild, Robert Hu, and Dino Sejdinovic. Generalized variational inference in function spaces: Gaussian measures meet Bayesian deep learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Veit David Wild, Sahra Ghalebikesabi, Dino Sejdinovic, and Jeremias Knoblauch. A rigorous link between deep ensembles and (variational) Bayesian methods. *Advances in Neural Information Processing Systems*, 36:39782–39811, 2023.

Eric P Xing, Michael I Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. *arXiv preprint arXiv:1212.2512*, 2012.

Zuheng Xu and Trevor Campbell. The computational asymptotics of Gaussian variational inference and the Laplace approximation. *Statistics and Computing*, 32, 2022.

J. Yan. Nonlinear large deviations: Beyond the hypercube. *Annals of Applied Probability*, 30:812–846, 2020.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya.  $\alpha$ -variational inference with statistical guarantees. *Annals of Statistics*, 48(2):886–905, 2020.

Rentian Yao and Yun Yang. Mean field variational inference via Wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*, 2022.

Arnold Zellner. Optimal information processing and Bayes’s theorem. *The American Statistician*, 42(4):278–280, 1988.

Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *Annals of Statistics*, 48(5):2575–2598, 2020.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *Annals of Statistics*, 48(4):2180–2207, 2020.

Chenyang Zhong, Sumit Mukherjee, and Bodhisattva Sen. Variational inference for latent variable models in high dimensions. *arXiv preprint arXiv:2506.01893*, 2025.

## Appendix A. Implementation and Computational Complexity

Algorithm 1 is a two-stage algorithm to approximate the  $\Xi$ -variational posterior. A natural question to ask is how tractable is it in high-dimensional models? In the first stage, existing methods for the pseudomarginal computation (such as mean-field variational inference) are known run in polynomially with respect to the parameter dimension (Jiang et al., 2025). But the second stage is a Sinkhorn algorithm, which is not necessarily scalable (Altschuler and Boix-Adsera, 2023). In this section, we discuss sufficient conditions for our algorithm to run in polynomial time.

### A.1 The Sinkhorn algorithm

We use discrete points to represent distributions. The set of marginals  $\{m_i\}_{i \in [D]}$  is a nonnegative matrix  $\mathbf{M} = (\mathbf{M}_1, \dots, \mathbf{M}_D) \in \mathbb{R}^{N \times D}$ , where each  $\mathbf{M}_i$  contains  $N$  design points. The EOT potentials  $\{\phi_i\}_{i \in [D]}$  are represented by a matrix  $\mathbf{F} = (\mathbf{F}_1, \dots, \mathbf{F}_D) \in \mathbb{R}^{N \times D}$ . The negative loss  $-\ell(\mathbf{x}^n; \theta)$  is represented as a cost tensor  $\mathbf{C} = (C_{i_1, \dots, i_D}) \in (\mathbb{R}^N)^{\otimes D}$ , and the variational posterior  $\mathbf{q}$  is a nonnegative tensor  $\mathbf{Q} = (Q_{i_1, \dots, i_D}) \in (\mathbb{R}^N)^{\otimes D}$ .

Numerically, the EOT problem (2.5) can be posed as a linear program with  $DN$  marginal constraints and  $N^D$  decision variables:

$$\min_{\mathbf{Q} > 0, r_i(\mathbf{Q}) = r_i(\mathbf{M})} \langle \mathbf{C}, \mathbf{Q} \rangle + (\lambda + 1) \langle \log \mathbf{Q} - \log \mathbf{M}, \mathbf{Q} \rangle.$$

The Sinkhorn algorithm (2) returns the potentials  $\mathbf{F}^*$  by iteratively performing log-sum-exp updates between  $\mathbf{F}_1, \dots, \mathbf{F}_D$ .

### A.2 The complexity of the Sinkhorn algorithm

Assuming the cost tensor is uniformly bounded, the best-known complexity bound for the multimarginal Sinkhorn algorithm is  $O(D^3 N^D (\lambda + 1)^{-2})$  (Lin et al., 2022). Unfortunately, the exponential dependence on  $D$  cannot be improved in general (Kroshnin et al., 2019). Polynomial-time solvability requires structural assumptions on  $\mathbf{C}$ . Altschuler and Boix-Adsera (2021) show that if the cost tensor has bounded treewidth or is the sum of a low-rank tensor and a sparse tensor, then multimarginal EOT is solvable in  $\text{poly}(N, D)$  time.

**Proposition 8 (Altschuler and Boix-Adsera (2023))** *Let  $\mathbf{C} \in (\mathbb{R}^N)^{\otimes D}$  satisfy one of the following:*

1.  $\mathbf{C}$  has graphical structure with constant junction tree width  $\omega$ ; or
2.  $\mathbf{C} = \mathbf{R} + \mathbf{S}$  where  $\mathbf{R}$  has constant multilinear rank and  $\mathbf{S}$  has  $\text{poly}(N, D)$  sparsity.

*Then for any  $\lambda \geq 0$ , Algorithm 2 terminates in  $\text{poly}(N, D, \mathbf{C}_{\max}/\epsilon, (\lambda + 1)^{-1})$  time.*

**Remark 9** *The bounded-treewidth assumption ensures polynomial-time solvability of the junction-tree algorithm (Wainwright et al., 2008). Examples include state-space models, topic models, and sparse linear regression.*

**Remark 10** *The “low-rank plus sparse” structural assumption is less commonly used in Bayesian inference. Roughly, it requires the posterior to be well approximated by a mixture of product distributions up to a  $\text{poly}(N, D)$ -sparse error term.*

---

**Algorithm 2:** (Multimarginal) Sinkhorn Algorithm

---

**Input:** Cost tensor  $\mathbf{C}$ , marginals  $\mathbf{M}$ , tolerance  $\epsilon$ , regularization parameter  $\lambda$ .**Initialize:**  $\mathbf{F}_i = -\frac{1}{D(\lambda+1)}\mathbf{1} - \log \mathbf{M}_i$  for all  $i \in [D]$ .**while**  $\mathbf{E} > \epsilon$  **do**    Select the greedy coordinate  $j = \arg \min_{i \in [D]} \|r_i(\mathbf{Q}) - \mathbf{M}_i\|_1$ .**for**  $i \in [D]$  **do**

Update

$$\mathbf{F}_i = \begin{cases} \mathbf{F}_i - \log(r_i(\mathbf{Q})) + \log(\mathbf{M}_i), & i = j, \\ \mathbf{F}_i, & \text{otherwise.} \end{cases}$$

Compute

$$r_i(\mathbf{Q}) = \frac{\sum_{1 \leq k_\ell \leq N, \ell \neq i} \exp \left[ \sum_{\ell=1}^D \mathbf{F}_{\ell k_\ell} - \frac{1}{\lambda+1} \mathbf{C}_{k_1 \dots k_D} \right] \mathbf{M}_{k_1 \dots k_D}}{\sum_{1 \leq k_\ell \leq N, \forall \ell} \exp \left[ \sum_{\ell=1}^D \mathbf{F}_{\ell k_\ell} - \frac{1}{\lambda+1} \mathbf{C}_{k_1 \dots k_D} \right] \mathbf{M}_{k_1 \dots k_D}}.$$

**end**    Set  $\mathbf{E} = \sum_{i=1}^D \|r_i(\mathbf{Q}) - \mathbf{M}_i\|_1$ .**end****Output:** An  $N \times D$  matrix  $\mathbf{F}$ .

For general graphs  $G$  with bounded treewidth, Fan et al. (2022) propose implementing the Sinkhorn algorithm using junction-tree message passing, yielding the following complexity bound:

**Corollary 11** *Let  $\mathbf{C} \in (\mathbb{R}^N)^{\otimes D}$  have constant treewidth  $\omega$ . Then Algorithm 2 implemented with the junction tree method converges in*

$$O(D^3 N^{\omega+1} (\lambda+1)^{-1} \epsilon^{-1})$$

*iterations.*

This result is adapted from Theorem 4 of Fan et al. (2022), which shows that computational cost decreases as  $\lambda$  increases. Polynomial dependence on  $D$  requires only that the treewidth  $\omega(G)$  grows slower than  $\log D$ , so multimarginal EOT may still be feasible for “locally tree-like” graphs.

Counterintuitively, when  $\lambda$  grows faster than  $D$ , the computational complexity decreases as  $D$  increases. However, as Corollaries 3 and 5 show in Section 4, in this regime the variational posterior collapses to the naive mean-field approximation.

**Remark 12** *There are well-known models that violate the assumptions in Proposition 8. For example, an Ising model on a complete  $D \times D$  graph has treewidth  $D$ , and its cost tensor is neither low rank nor sparse. Multimarginal Sinkhorn becomes NP-hard in this setting (Altschuler and Boix-Adsera, 2021). A practical workaround is to group variables and couple only the groupwise marginals. For instance, in an Ising model with 100 variables, one may group the first 50 and last 50 variables, yielding two 50-dimensional marginals. This still*

yields a strict improvement over MFVI. Moreover, when the number of groups is fixed, the Sinkhorn complexity becomes polynomial because it is insensitive to the dimensionality within each group (Altschuler et al., 2017).

### A.3 Finite-Sample Convergence

Here we analyze convergence properties of the  $\Xi$ -variational posterior  $q_\lambda^*$  when  $\lambda$  tends to 0 or  $\infty$ , while keeping  $n$  fixed. Understanding this setting justifies the stability of the algorithmic output after we replace the marginals of  $q_\lambda^*$  with a set of pseudomarginals in Algorithm 1. Moreover, the convergence results of the  $\Xi$ -variational posterior for both large and small  $\lambda$  values are useful from a classical Bayesian perspective that treats the observed data as known (Berger, 2013).

We show that  $\Xi$ -variational posterior converges to the mean-field variational posterior as  $\lambda$  tends to infinity, and converges to the exact posterior as  $\lambda$  tends to zero. Then we establish a stability property for  $q_\lambda^*$  when we replace its marginals with another set of marginals, which helps justify Algorithm 1.

Let us define a cost function  $C_\lambda$  over  $\lambda \in \bar{\mathbb{R}}_+$  as  $C_\lambda := \max_{q \in \mathbb{P}_2(\Theta)} \text{ELBO}(q) - \lambda \Xi(q)$ .

**Limits as  $\lambda \rightarrow \infty$  or  $\lambda \rightarrow 0$ .** We start with the convergence of  $q_\lambda^*$  and  $C_\lambda$  as  $\lambda$  tends to infinity.

**Theorem 13** *Assume that  $D_{KL}(q \parallel q_0^*) < \infty$  for some  $q \in \mathbb{M}(\Theta)$ . For each  $\lambda \in \bar{\mathbb{R}}_+$ , define the set  $\mathbb{Q}_\lambda$  as the set of minimizers for the functional  $q \mapsto D_{KL}(q \parallel q_0^*) + \lambda \Xi(q)$  in  $\mathbb{P}_2(\Theta)$ . If  $q_\infty^* \in \mathbb{Q}_\infty$ , there exists a sequence  $q_\lambda^* \in \mathbb{Q}_\lambda$  such that  $\lim_{\lambda \rightarrow \infty} W_2(q_\infty^*, q_\lambda^*) = 0$ . Furthermore, the  $\Xi$ -VI cost converges to the mean-field ELBO, i.e.  $\lim_{\lambda \rightarrow \infty} |C_\lambda - C_0| = 0$ .*

The result shows that every mean-field variational posterior is an accumulation point of some sequence of  $\Xi$ -variational posteriors. This type of result is called "large-time limits" in the optimal transport literature. When the likelihood is quadratic, it is possible to prove an exponential rate of convergence for  $C_\lambda$  under more restrictive conditions (Conforti and Tamanini, 2021). However, this setting is uninteresting for Bayesian inference and we do not pursue it in this paper.

As  $\lambda$  tends to zero, we provide analog results to show that  $\Xi$ -variational posterior converges to the exact posterior in the Wasserstein metric.

**Theorem 14** *Assume that  $\Xi(q_0^*) < \infty$   $[P_{\theta_0}]$ -almost surely. For each  $\lambda \in \bar{\mathbb{R}}_+$ , define the set  $\mathbb{Q}_\lambda$  as the set of minimizers for the functional  $q \mapsto D_{KL}(q \parallel q_0^*) + \lambda \Xi(q)$  in  $\mathbb{P}_2(\Theta)$ . If  $q_\lambda^* \in \mathbb{Q}_\lambda$  converges as  $\lambda \rightarrow 0$  in the Wasserstein metric, then  $\lim_{\lambda \rightarrow 0} W_2(q_0^*, q_\lambda^*) = 0$ . Furthermore, the  $\Xi$ -VI cost converges to the true posterior ELBO, i.e.  $\lim_{\lambda \rightarrow 0} |C_\lambda - C_0| = 0$ .*

**Algorithmic Stability.** Let  $m_\lambda^*$  be the product of marginals of  $q_\lambda^*$ . In Section 2.2, we produce to replace idealized Algorithm 3 with a simple, efficient approximate Algorithm 1. A natural question to ask is whether the solution is stable after we replace  $m_\lambda^*$  with pseudomarginals  $\tilde{m}$ . To answer this question, we leverage the tools of quantitative stability from OT theory (Eckstein and Nutz, 2022). We make two assumptions: a Lipschitz cost assumption, and transportation cost inequality.

**Assumption 5 (Lipschitz Cost Assumption)** *We assume that there exists a constant  $L \geq 0$  and  $\phi_i : \Theta_i \rightarrow \mathbb{R}$  such that for all  $q \in \mathcal{C}(m_\lambda^*)$  and  $\tilde{q} \in \mathcal{C}(\tilde{m})$ ,*

$$\left| \int_{\Theta} \left( \ell(\mathbf{x}^{(n)}; \theta) - \sum_{i=1}^D \phi_i(\theta_i) \right) (q(\theta) - \tilde{q}(\theta)) d\theta \right| \leq L W_2(q, \tilde{q}). \quad (\text{A.1})$$

This assumption is slightly more general than the Lipschitzness of  $\ell(\mathbf{x}^{(n)}; \cdot)$  minus additive correction factors. As an example, the Gaussian likelihood satisfies Assumption 5 (Lemma 3.5, Eckstein and Nutz (2022)).

**Assumption 6 (Transportation Cost Inequality)** *A product distribution  $m$  over  $\Theta$  satisfies the transportation cost inequality if there exists a constant  $C$  such that*

$$W_2(q_1, q_2) \leq \sqrt{D_{KL}(q_1 \parallel q_2)}, \quad \text{for all } q_1, q_2 \in \mathcal{C}(m).$$

Assumption 6 is standard in high-dimensional statistics (Wainwright, 2019). When  $\Theta$  is compact, the assumption follows from Pinsker's inequality. Otherwise, this assumption holds when each marginal has a finite exponential moment.

We now state the main stability result which upper bounds the approximation error of Algorithm 1 using the approximation error of the pseudomarginals.

**Theorem 15 (Stability of Algorithm 1)** *Let Assumption 5 hold with a Lipschitz constant  $L$ . Let  $m_\lambda^*$  be the marginals of  $\Xi$ -variational posterior  $q_\lambda^*$  and  $\tilde{m} \in \mathbb{M}(\Theta)$  be another product distribution. Suppose  $m_\lambda^*$  satisfies Assumption 6 with a fixed constant  $C$ . Then for the one-step approximation  $\tilde{q}_\lambda$  defined in Algorithm 1 with pseudomarginals  $\tilde{m}$ , the following upper bound holds:*

$$W_2(q_\lambda^*, \tilde{q}_\lambda) \leq W_2(m_\lambda^*, \tilde{m}) + 2CL^{\frac{1}{4}} W_2^{\frac{1}{4}}(m_\lambda^*, \tilde{m}). \quad (\text{A.2})$$

The proof uses an OT technique called shadowing. See Section E for details.

The result highlights the stability of Algorithm 1, as the approximation error of  $q_\lambda^*$  is only Lipschitz in the approximation error of the pseudomarginals. If  $\tilde{m}$  is close enough to  $m_\lambda^*$  in terms of the  $W_2$  metric, the output of Algorithm 1 is guaranteed to well approximate true variational posterior  $q_\lambda^*$ .

**Corollary 16** *Assume Assumption 5 with Lipschitz constant  $L$ , and Assumption 6 for the pseudomarginals. Then the following limits hold:*

1. *Let  $q_\lambda^{*(\infty)}$  be the optimizer of Eq. (2.5) with marginals  $\{q_{\infty,i}^*\}_{i \in [D]}$ . Then  $\lim_{\lambda \rightarrow \infty} W_2(q_\lambda^{*(\infty)}, q_\lambda^*) = 0$ .*
2. *Let  $q_\lambda^{*(0)}$  be the optimizer of Eq. (2.5) with marginals  $\{q_{0,i}^*\}_{i \in [D]}$ . Then  $\lim_{\lambda \rightarrow 0} W_2(q_\lambda^{*(0)}, q_\lambda^*) = 0$ .*

The Corollary is a consequence of Theorem 13, Theorem 14, and Theorem 15. As  $\lambda$  tends to 0 or  $\infty$ , the error of replacing the idealized Algorithm 3 with Algorithm 1 vanishes when we use exact posterior marginals or mean-field variational posteriors, respectively. If we plug in a consistent estimate of the exact posterior marginals (e.g. TAP approximation of a linear model with i.i.d. Gaussian design (Celentano et al., 2023a)), then Algorithm 1 asymptotically recovers the exact posterior as  $\lambda$  tends to zero.

#### A.4 Full Coordinate Ascent Algorithm

In this section, we present a full coordinate descent algorithm to exactly optimize the  $\Xi$ -VI objective.

##### A.4.1 OUTER VARIATIONAL PROBLEM

We now derive steps to solve the outer variational problem of Eq. (2.4). Treating  $\phi_1, \dots, \phi_D$  as fixed, we optimize over the marginals  $m_i$ 's:

$$\min_{m \in \mathbb{M}(\Theta)} \mathbb{E}_m \left[ \underbrace{(\lambda + 1) \left( \sum_{i=1}^D \phi_i(\theta_i) \right) \exp \left( \sum_{i=1}^D \phi_i(\theta_i) + \frac{1}{\lambda + 1} \ell(\mathbf{x}; \theta) \right)}_{\text{surrogate loss}} \right] + D_{\text{KL}}(m \parallel \pi). \quad (\text{A.3})$$

Eq. (A.3) is equivalent to a mean-field VI problem with a surrogate log-likelihood. To solve Eq. (A.3), we follow a method based on coordinate ascent variational inference (CAVI) (Blei et al., 2017).

Denote  $\Theta_{-i} := \prod_{j \neq i} \Theta_j$ ,  $\theta_{-i} := (\theta_{[D] \setminus \{i\}})$ , and

$$m_{-i}^t := \prod_{j < i} m_j^{t+1}(\theta_j) \prod_{j > i} m_j^t(\theta_j).$$

Define  $\nu_i^{t+1}(\theta_i)$  as

$$\nu_i^{t+1}(\theta_i) := \mathbb{E}_{m_{-i}^t} \left[ (\lambda + 1) \left( \sum_{i=1}^D \phi_i^{t+1}(\theta_i) \right) \exp \left( \sum_{i=1}^D \phi_i^{t+1}(\theta_i) + \frac{1}{\lambda + 1} \ell(\mathbf{x}; \theta_i, \theta_{-i}) \right) \right]. \quad (\text{A.4})$$

Given marginals  $m^t = (m_1^t, \dots, m_D^t)$  and potentials  $\phi^{t+1} = (\phi_1^{t+1}, \dots, \phi_D^{t+1}) \in E(m^t)$ , CAVI iteratively updates each marginal  $i \in [D]$  by solving:

$$m_i^{t+1} = \arg \min_{m_i \in \mathbb{M}(\Theta_i)} \mathbb{E}_{m_i} [\nu_i^{t+1}(\theta_i)] + D_{\text{KL}}(m_i \parallel \pi_i).$$

This yields the explicit update

$$m_i^{t+1}(\theta_i) \propto \pi_i(\theta_i) \exp(-\nu_i^{t+1}(\theta_i)), \quad \forall \theta_i \in \Theta_i. \quad (\text{A.5})$$

##### A.4.2 FULL COORDINATE ASCENT ALGORITHM

Algorithm 3 presents the full coordinate ascent algorithm. It monitors the change in the ELBO as the criterion for convergence, which is equivalent (up to a constant) to the KL divergence between the variational posterior and the exact posterior.

Unfortunately, Algorithm 3 is difficult to implement because we cannot compute the expectations required in Eq. (2.10) or Eq. (A.4). When the  $\phi_i$ 's are represented implicitly, there is no practically stable MFVI procedure for such implicit log-likelihoods, especially in high-dimensional models.

**Algorithm 3:** Coordinate Ascent Algorithm

---

**Input:** Log-likelihood  $\ell(\mathbf{x}; \theta)$ , prior  $\pi$ , tolerance  $\epsilon$ , regularization parameter  $\lambda$ .

**Initialize:** marginals  $m_1^0, \dots, m_D^0$ ; EOT potentials  $\phi_1^0, \dots, \phi_D^0$ ;  $t = 0$ .

**while** *ELBO* has not converged **do**

**for**  $i \in [D]$  **do**  
| Update  $\phi_i^{t+1}(\theta_i)$  using Eq. (2.10).

**end**

**for**  $i \in [D]$  **do**  
| Update  $m_i^{t+1}(\theta_i)$  using Eq. (A.5). // Challenging step

**end**

Compute

$$q^{t+1}(\theta) = \exp\left(\sum_{i=1}^D \phi_i^{t+1}(\theta_i) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta)\right) \prod_{i=1}^D m_i^{t+1}(\theta_i).$$

Compute

$$\text{ELBO}(q^{t+1}) = \mathbb{E}_{q^{t+1}}[\ell(\mathbf{x}; \theta) + \log \pi(\theta)] - \mathbb{E}_{q^{t+1}}[\log q^{t+1}(\theta)].$$

Increment  $t = t + 1$ .

**end**

**Output:**  $q(\theta)$ .

---

## Appendix B. Support Results

**Lemma 17 (Gibbs variational principle)** *For probability measures  $\mu$  on  $\Theta$ ,*

$$\log \mathbb{E}_\mu[\exp(f(\theta))] = \sup_{\nu \in \mathbb{P}(\Theta)} \{\mathbb{E}_\nu[f(\theta)] - D_{KL}(\nu \parallel \mu)\}. \quad (\text{B.1})$$

**Lemma 18** *Let  $A$  be a symmetric, positive definite matrix. For all  $j \in [D]$ , it holds that  $(A^{-1})_{jj} \geq \frac{1}{A_{jj}}$ .*

**Proof** Given that  $A$  is symmetric and positive definite, there exists another symmetric positive definite matrix  $B$  such that  $B^2 = A$ . We note that  $A_{jj} = e_j^T A e_j = e_j^T B^T B e_j = \|B e_j\|_2^2$  and similarly,  $(A^{-1})_{jj} = \|B^{-1} e_j\|_2^2$ .

By the Cauchy–Schwarz inequality, we have

$$\langle B e_j, B^{-1} e_j \rangle^2 \leq \|B e_j\|_2^2 \|B^{-1} e_j\|_2^2 = A_{jj} (A^{-1})_{jj}.$$

However,

$$\langle B e_j, B^{-1} e_j \rangle = e_j^T (B^{-1})^T B e_j = e_j^T B^{-1} B e_j = e_j^T e_j = 1.$$

Therefore, we have  $A_{jj} (A^{-1})_{jj} \geq 1$ , which simplifies to  $(A^{-1})_{jj} \geq \frac{1}{A_{jj}}$ . This completes the proof. ■

**Lemma 19** For  $q \in \mathbb{P}(\Theta)$ , the following variational characterization of its expressivity holds:

$$\Xi(q) = \sup_f \mathbb{E}_q[f] - \log \mathbb{E}_{q_i} \mathbb{E}_{q_{-i}} [\exp(f(\theta_i, \theta_{-i})) \mid \theta_i].$$

**Proof** Apply Donsker–Varadhan lemma. We obtain

$$\Xi(q) = \sup_f \mathbb{E}_q[f] - \log \mathbb{E}_{q_i} \mathbb{E}_{q_{-i}} [\exp(f(\theta_i, \theta_{-i}))].$$

Apply Donsker–Varadhan lemma again to  $D_{KL}(q(\theta_i, \theta_{-i})$  ■

**Theorem 20 (Theorem 5.11, Santambrogio (2015))** In the space  $\mathbb{P}_p(\mathbb{R}^d)$ , we have  $W_p(\mu_n, \mu) \rightarrow 0$  if and only if  $\mu_n \rightarrow \mu$  weakly and

$$\int |x|^p d\mu_n \rightarrow \int |x|^p d\mu,$$

where  $p > 0$  is a given exponent.

**Lemma 21** Let  $q_n$  be a sequence of measures in  $\mathbb{P}_2(\Theta)$ . If  $W_2(q_n, q) \rightarrow 0$  for some  $q \in \mathbb{P}_2(\Theta)$ , then

$$\liminf_{n \rightarrow \infty} \Xi(q_n) \geq \Xi(q).$$

Let  $q_0$  be another measure in  $\mathbb{P}_2(\Theta)$ . We have

$$\liminf_{n \rightarrow \infty} D_{KL}(q_n \parallel q_0) \geq D_{KL}(q \parallel q_0).$$

**Proof** The second property follows from the fact that functional  $D_{KL}(\cdot \parallel q_0^*)$  is continuous in the Wasserstein metric (Proposition 7.1, Santambrogio (2015)). For any  $q_n \xrightarrow{W_2} q$ , Theorem 20 implies that  $q_n$  weakly converge to  $q_0$ . The convergence  $W_2(q_n, q_0) \rightarrow 0$  implies the convergence  $W_2(q_{n,i}, q_{0,i}) \rightarrow 0$  for each  $i \in [D]$ . Since  $D_{KL}$  is lower semicontinuous in both arguments (Theorem 4.8, Polyanskiy and Wu (2025)), we get

$$\liminf_{n \rightarrow \infty} \Xi(q_n) = \liminf_{n \rightarrow \infty} D_{KL} \left( q_n \parallel \prod_{i=1}^D q_{n,i} \right) \geq D_{KL} \left( q_0 \parallel \prod_{i=1}^D q_{0,i} \right) = \Xi(q), \quad (\text{B.2})$$

where  $D$  is fixed with respect to  $n$ . ■

**Definition 22 (Shadow)** Let  $p \in [1, \infty]$  and  $m, \tilde{m}$  be product measures within  $\mathbb{P}_p(\Theta)$ . Assume  $\kappa_i \in \mathcal{C}(m_i, \tilde{m}_i)$  is a coupling that achieves  $W_p(m_i, \tilde{m}_i)$  and let  $\kappa_i = m_i \otimes K_i$  represent a disintegration. For a given  $q \in \mathcal{C}(m)$ , its shadow  $q^s \in \mathcal{C}(\tilde{m})$  is defined as the second marginal of  $q \otimes K \in \mathbb{P}(\Theta \times \Theta)$ , where the kernel  $K : \Theta \rightarrow \mathbb{P}(\Theta)$  is constructed as a direct sum  $K(x) = K_1(x_1) \otimes \dots \otimes K_D(x_D)$ .

Given a coupling  $q \in \mathcal{C}(m)$ , its shadow  $q^s$  satisfies the following properties.

**Lemma 23 (Lemma 3.2, Eckstein and Nutz (2022))** *For product distributions  $m, \tilde{m} \in \mathbb{P}_2(\Theta)$  and coupling  $q \in \mathcal{C}(m)$ , its shadow  $q^s \in \mathcal{C}(\tilde{m})$  satisfies*

$$W_2(q, q^s) = W_2(m, \tilde{m}), \quad D_f(q^s \parallel \tilde{m}) \leq D_f(q \parallel m),$$

where  $D_f(\cdot)$  is any  $f$ -divergence.

**Theorem 24 (Theorem 12, Lin et al. (2022))** *Let  $\{\phi^t\}_{t \geq 0}$  be the iterates generated by Algorithm 2. The number of iterations  $t$  required to reach the stopping criterion  $E \leq \epsilon'$  is upper bounded by:*

$$t \leq 2 + \frac{2D^2 [\|\mathbf{C}\|_\infty / (\lambda + 1) - \log(\min_{1 \leq i \leq N, 1 \leq j \leq D} m_{ij})]}{\epsilon'}.$$

## Appendix C. Proofs of Section 2

**Proof** [Derivation of Eq. (2.5)] Let  $m$  be given. Then to optimize  $q^*$ , we have

$$\begin{aligned} q^*(\theta) &= \arg \min_{q \in \mathcal{C}(m)} \mathbb{E}_q[-\ell(\mathbf{x}; \theta)] + \lambda D_{\text{KL}}(q \parallel m) + D_{\text{KL}}(q \parallel \pi) \\ &= \arg \min_{q \in \mathcal{C}(m)} \mathbb{E}_q[-\ell(\mathbf{x}; \theta)] + \lambda D_{\text{KL}}(q \parallel m) + D_{\text{KL}}(q \parallel m) + D_{\text{KL}}(m \parallel \pi) \\ &= \arg \min_{q \in \mathcal{C}(m)} \mathbb{E}_q[-\ell(\mathbf{x}; \theta)] + (\lambda + 1) D_{\text{KL}}(q \parallel m). \end{aligned}$$

The first line uses the fact that  $m$  is a product distribution. The third line drops  $D_{\text{KL}}(m \parallel \pi)$  as it does not depend on  $q$ .  $\blacksquare$

The next result states that the solution to the EOT problem (2.5) has a unique representation.

**Theorem 25 (Structure Theorem for Multi-Marginal EOT)** *Assume that*

$$\inf_{q \in \mathcal{C}(m)} \{-\mathbb{E}_q[\ell(\mathbf{x}; \theta)] + (\lambda + 1)\mathbb{E}(q)\} < \infty \quad \text{and} \quad \sup_{\theta \in \Theta} \ell(\mathbf{x}; \theta) < \infty.$$

*Then there exists a unique minimizer  $q^*$  to the inner variational problem (2.5) that is absolutely continuous with respect to  $m$  (denoted  $q^* \ll m$ ), and the following hold:*

(1) *There exist measurable functions  $\phi_i^* : \Theta_i \rightarrow \mathbb{R}$  for  $i \in [D]$  such that*

$$q^*(\theta) = \exp \left( \sum_{i=1}^D \phi_i^*(\theta_i) + \frac{1}{\lambda + 1} \ell(\mathbf{x}; \theta) \right) m(\theta), \quad (\text{C.1})$$

*$m$ -almost surely. The collection  $\phi^* := (\phi_1^*, \dots, \phi_D^*)$  is referred to as the EOT potentials. Each  $\phi_i^*$  is  $m_i$ -almost surely measurable and unique up to an additive constant. Moreover, if  $\mathbb{E}_{m_i}[\phi_i^*] \geq 0$ , then  $\phi_i^* \in L^\infty(m_i)$  for all  $i \in [D]$ .*

(2) *Conversely, suppose  $q \in \mathcal{C}(m)$  admits a density of the form in Eq. (C.1),  $m$ -almost surely, for some functions  $\phi_i : \Theta_i \rightarrow \mathbb{R}$ . Then  $q$  minimizes the inner variational problem in Eq. (2.5), and the functions  $\phi_i$  are the EOT potentials.*

This result first appeared heuristically in Carlier (2022). For  $D = 2$ , the uniform boundedness assumption can be relaxed to  $\ell(\mathbf{x}; \cdot)$  being integrable (Theorem 4.2, Nutz (2021)).

**Proof** [Proof of Theorem 25] Define the auxiliary distribution  $q_{\text{aux}} \in \mathbb{P}(\Theta)$  as

$$q_{\text{aux}}(\theta) = \mathcal{Z}_n(\lambda)^{-1} \exp\left(\frac{1}{\lambda+1}\ell(\mathbf{x}; \theta)\right) m(\theta), \quad (\text{C.2})$$

where  $\mathcal{Z}_n(\lambda)$  is the normalizing constant. Since  $\sup_{\theta \in \Theta} \ell(\mathbf{x}; \theta) < \infty$ ,  $\mathcal{Z}_n(\lambda) < \infty$ , and hence  $q_{\text{aux}}$  is well-defined and absolutely continuous with respect to  $m$ .

Minimizing the objective function in Eq. (2.5) is equivalent to minimizing the KL loss to  $q_{\text{aux}}$ .

$$\begin{aligned} & \min_{q \in \mathcal{C}(m)} \mathbb{E}_q [-\ell(\mathbf{x}; \theta)] + (\lambda + 1) D_{\text{KL}}(q \parallel m) \\ &= \min_{q \in \mathcal{C}(m)} \mathbb{E}_q \left[ -\frac{1}{\lambda+1} \ell(\mathbf{x}; \theta) \right] + D_{\text{KL}}(q \parallel m) \\ &= \min_{q \in \mathcal{C}(m)} \int_{\Theta} q(\theta) \log \frac{q(\theta)}{\exp\left(\frac{1}{\lambda+1}\ell(\mathbf{x}; \theta)\right) m(\theta)} d\theta \\ &= \min_{q \in \mathcal{C}(m)} D_{\text{KL}}(q \parallel q_{\text{aux}}) - \log \mathbb{E}_m \exp\left(\frac{1}{\lambda+1}\ell(\mathbf{x}; \theta)\right) \\ &\stackrel{C}{=} \min_{q \in \mathcal{C}(m)} D_{\text{KL}}(q \parallel q_{\text{aux}}). \end{aligned} \quad (\text{C.3})$$

Since  $\mathcal{C}(m)$  is displacement convex and the KL functional is displacement convex (Villani, 2009), the solution is unique.

Let  $q^*$  be the optimizer. Then, by the method of Lagrange multipliers, there exist dual variables  $\phi_i^* \in L^\infty(m_i)$  such that

$$q^* = \arg \min_{q \in \mathbb{P}(\Theta)} D_{\text{KL}}(q \parallel q_{\text{aux}}) + \sum_{i=1}^D (\mathbb{E}_{m_i}[\phi_i^*] - \mathbb{E}_{q_i}[\phi_i^*]).$$

This is equivalent to

$$q^*(\theta) \propto \exp\left(\sum_{i=1}^D \phi_i^*(\theta_i) + \frac{1}{\lambda+1}\ell(\mathbf{x}; \theta)\right) m(\theta), \quad (\text{C.4})$$

after normalizing. Since the potentials  $\phi_i^*$  are uniformly bounded, the resulting normalization constant is finite. By adding the normalizing constant to  $\phi_D$ , we obtain

$$q^*(\theta) = \exp\left(\sum_{i=1}^D \phi_i^*(\theta_i) + \frac{1}{\lambda+1}\ell(\mathbf{x}; \theta)\right) m(\theta).$$

”only if” direction: Assume that the optimal coupling  $q^*$  is given by

$$q^* = \exp\left(\sum_{i=1}^D \phi_i(\theta_i) + \frac{1}{\lambda+1}\ell(\mathbf{x}; \theta)\right) m(\theta).$$

where  $\phi = (\phi_1, \dots, \phi_D) \in \prod_{i=1}^D L^1(m_i)$  are some potential functions.

Plugging the solution in the EOT primal problem, for each  $i$  and  $[m_i]$ -a.s.  $\theta_i$ , the potentials satisfy a set of fixed point equations called the *Schrödinger system*:

$$\exp(\phi_i(\theta_i)) \int_{\Theta_{-i}} \exp \left( \sum_{j \neq i} \phi_j(\theta_j) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta) \right) m_{-i}(\theta_{-i}) d\theta_{-i} = 1. \quad (\text{C.5})$$

The Schrödinger system (Eq. (C.5)) satisfies the Euler-Lagrange optimality condition for the primal EOT problem (Carlier, 2022). Precisely, the EOT potentials solve

$$\max_{\phi \in \prod_{i=1}^D L^1(m_i)} \sum_{i=1}^D \mathbb{E}_{m_i} \phi_i - \mathbb{E}_m \left[ \exp \left( \sum_{i=1}^D \phi_i(\theta_i) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta) \right) \right],$$

which is the dual problem to the multimarginal EOT problem (Eq. (2.5)). Since the EOT problem is convex (Nutz, 2021), the primal-dual gap closes, which means the probability measure  $q$  defined under  $\phi$  solves Eq. (2.5).

To see that  $\phi_i \in L^\infty(\Theta_i)$  for  $i \in [D]$ . Assume that  $\mathbb{E}_{m_i}[\phi_i] \geq 0$ , which is possible under the Euler-Lagrange condition:

$$\sum_{i=1}^D \mathbb{E}_{m_i} \phi_i = \min_{q \in \mathcal{C}(m)} D_{\text{KL}}(q \| q_{\text{aux}}) \geq 0$$

By Eq. (C.5), we apply Jensen's inequality to obtain that

$$\begin{aligned} \phi_i(\theta_i) &= -\log \int_{\Theta_{-i}} \exp \left( \sum_{j \neq i} \phi_j(\theta_j) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta) \right) dm_{-i}(\theta_{-i}) \\ &\leq -\mathbb{E}_{m_{-i}} \left[ \sum_{j \neq i} \phi_j(\theta_j) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta) \right] \leq -\frac{1}{\lambda+1} \mathbb{E}_{m_{-i}} [\ell(\mathbf{x}; \theta)], \end{aligned}$$

thus  $\sup_{\theta_i \in \Theta_i} \phi_i(\theta_i) \leq -\sup_{\theta \in \Theta} |\ell(\mathbf{x}; \theta)| / (\lambda+1)$  for all  $i \in [D]$ .

For the other direction, since  $\sup_{\theta \in \Theta} \ell(\mathbf{x}; \theta) < \infty$ , we have

$$\begin{aligned} \phi_i(\theta_i) &= -\log \int_{\Theta_{-i}} \exp \left( \sum_{j \neq i} \phi_j(\theta_j) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta) \right) dm_{-i}(\theta_{-i}) \\ &\geq -\frac{\sup_{\theta \in \Theta} \ell(\mathbf{x}; \theta)}{\lambda+1} - \log \int_{\Theta_{-i}} \exp \left( \sum_{j \neq i} \phi_j(\theta_j) \right) dm_{-i}(\theta_{-i}). \end{aligned}$$

Since the right-hand side of the inequality does not depend on  $\theta_i$ ,  $\inf_{\theta_i \in \Theta_i} \phi_i(\theta_i) > -\infty$  as long as  $\sum_{j \neq i} \phi_j(\theta_j) < \infty$  holds  $[m_{-i}]$ -almost surely. Since  $\sup_{\theta_i \in \Theta_i} \phi_i(\theta_i) \leq -\sup_{\theta \in \Theta} \ell(\mathbf{x}; \theta) / (\lambda+1)$ , we have that  $\inf_{\theta_i \in \Theta_i} \phi_i(\theta_i) > -\infty$  for all  $i \in [D]$ .  $\blacksquare$

**Proof** [Proof of Eq. (A.3)] We make the following derivation,

$$\begin{aligned}
& \min_{m \in \mathbb{M}(\Theta)} \min_{q \in \mathcal{C}(m)} \mathbb{E}_q[-\ell(\mathbf{x}; \theta)] + (\lambda + 1) D_{\text{KL}}(q \parallel m) + D_{\text{KL}}(m \parallel \pi) \\
&= \min_{m \in \mathbb{M}(\Theta)} (\lambda + 1) \int_{\Theta} q^*(\theta) \log \frac{q^*(\theta)}{\exp(\frac{1}{\lambda+1} \ell(\mathbf{x}; \theta)) m(\theta)} d\theta + D_{\text{KL}}(m \parallel \pi) \\
&= \min_{m \in \mathbb{M}(\Theta)} (\lambda + 1) \int_{\Theta} q^*(\theta) \log \frac{\exp\left(\sum_{i=1}^D \phi_i(\theta_i) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta)\right)}{\exp(\frac{1}{\lambda+1} \ell(\mathbf{x}; \theta))} d\theta + D_{\text{KL}}(m \parallel \pi) \\
&= \min_{m \in \mathbb{M}(\Theta)} \mathbb{E}_m \left[ \underbrace{(\lambda + 1) \left( \sum_{i=1}^D \phi_i(\theta_i) \right) \exp\left(\sum_{i=1}^D \phi_i(\theta_i) + \frac{1}{\lambda+1} \ell(\mathbf{x}; \theta)\right)}_{\text{surrogate loss}} \right] + D_{\text{KL}}(m \parallel \pi).
\end{aligned}$$

■

## Appendix D. Proofs of Section A

**Proof** [Proof of Proposition 8] By Theorem 14 of Lin et al. (2022), Algorithm 2 reaches the stopping criterion  $\mathbf{E} \leq \epsilon$  in  $t$  iterations, where  $t$  satisfies

$$t \leq 2 + \frac{2D^2 \|\mathbf{C}\|_{\infty} - \log(\max_{ij} \mathbf{M}_{ij})}{\epsilon(\lambda + 1)}.$$

This implies

$$t \asymp \text{poly}(D, \mathbf{C}_{\max}/\epsilon, (\lambda + 1)^{-1}). \quad (\text{D.1})$$

Algorithm 2 calls the following oracle  $D$  times:

$$\text{Compute } r_i(\mathbf{Q}) = \frac{\sum_{1 \leq k_j \leq N, j \neq i} \exp\left[\sum_{j=1}^D \mathbf{F}_{jk_j} - \frac{1}{\lambda+1} \mathbf{C}_{k_1 \dots k_D}\right] \mathbf{M}_{k_1 \dots k_D}}{\sum_{1 \leq k_j \leq N, j \in [D]} \exp\left[\sum_{j=1}^D \mathbf{F}_{jk_j} - \frac{1}{\lambda+1} \mathbf{C}_{k_1 \dots k_D}\right] \mathbf{M}_{k_1 \dots k_D}}. \quad (\text{D.2})$$

All other steps are computed in linear time.

By Theorems 5.5 and 7.4 of Altschuler and Boix-Adsera (2023), this oracle can be computed in  $\text{poly}(N, D)$  time. Repeating the oracle evaluation  $Dt$  times and using Eq. (D.1), the algorithm terminates in  $\text{poly}(N, D, \mathbf{C}_{\max}/\epsilon, (\lambda + 1)^{-1})$  time. ■

**Proof** [Proof of Corollary 11] We consider Algorithm 1 of Fan et al. (2022). The algorithm implements the marginalization step of Algorithm 2 using the sum–product method. Consider a graph  $G = ([D], E, K)$ , where  $[D]$ ,  $E$ , and  $K$  denote the nodes, edges, and maximal cliques, respectively. If the log-likelihood  $\ell(\mathbf{x}^{(n)}; \theta)$  factorizes according to  $G$ , by the Hammersley–Clifford theorem,  $\ell(\mathbf{x}; \theta) = \sum_{\alpha \in K} \ell_{\alpha}(\theta_{\alpha})$  where each  $\ell_{\alpha}$  is defined on  $\prod_{j \in \alpha} \Theta_j$ .

Define  $\mathbf{C}_{k_\alpha}$  as the tensor of  $\ell_\alpha(\theta_\alpha)$  evaluated at support points  $(\theta_i^{(s)}, i \in \alpha)_{s \in [N]}$ . The full cost tensor decomposes as

$$\mathbf{C}_{k_1, \dots, k_D} = \sum_{\alpha \in K} \mathbf{C}_{k_\alpha}.$$

Let  $t$  be the number of iterations for Algorithm 1 of Fan et al. (2022) to reach tolerance  $\epsilon$ . By Theorem 1 of Fan et al. (2022),

$$\mathbb{E}[t] = O\left(D^2 \max_{\alpha \in K} \|\mathbf{C}_{k_\alpha}\|_\infty (\lambda + 1)^{-1} \epsilon^{-1}\right).$$

Let  $\mathcal{T}$  be a minimal junction tree for  $G$ . Marginalizing over each factor in  $\mathcal{T}$  costs  $O(N^{\omega(G)})$ , and message passing costs  $O(d(\mathcal{T}) N^{\omega(G)})$ , where  $d(\mathcal{T})$  is the average leaf distance in  $\mathcal{T}$ .

Since  $\max_{\alpha \in K} \|\mathbf{C}_{k_\alpha}\|_\infty$  is uniformly bounded, the sum-product implementation of the Sinkhorn algorithm requires  $O(d(\mathcal{T}) N^{\omega(G)} D^2 (\lambda + 1)^{-1} \epsilon^{-1})$  iterations. Because  $d(\mathcal{T}) \leq D$ , the complexity is also  $O(N^{\omega(G)} D^3 (\lambda + 1)^{-1} \epsilon^{-1})$ .  $\blacksquare$

## Appendix E. Proofs of Section 4

### Proofs of Section 4.2

We define the set  $\tilde{\Theta}_n$  as the set of all  $h$  defined in Eq. (4.13), and  $\mathbb{H}(q) := \int_{\Theta} q(\theta) \log q(\theta) d\theta$  as the Boltzmann's  $H$ -functional (Villani, 2009).

**Lemma 26 (Transformation Identities)** *For  $h := \delta_n^{-1}(\theta - \theta_0 - \delta_n \Delta_{n,\theta_0})$  where  $\theta \sim q$ , we have*

$$q(\theta) = |\det(\delta_n)|^{-1} \tilde{q}(h), \quad \text{and} \quad q_i(\theta) = \delta_{n,ii}^{-1} \tilde{q}_i(h) \text{ for } i \in [D],$$

Moreover, we have

$$\mathbb{H}(q) = \mathbb{H}(\tilde{q}) - \log |\det(\delta_n)|, \quad \text{and} \quad \mathbb{H}(q_i) = \mathbb{H}(\tilde{q}_i) - \log \delta_{n,ii},$$

and

$$\Xi(q) = \Xi(\tilde{q}) + \log |\det(\delta_n)| - \sum_{i=1}^D \log \delta_{n,ii},$$

and for any distribution  $q_1, q_2$  over  $\Theta$ , we have

$$D_{KL}(q_1 \parallel q_2) = D_{KL}(\tilde{q}_1 \parallel \tilde{q}_2),$$

where  $\tilde{q}_1, \tilde{q}_2$  are densities defined via Eq. (4.13).

**Proof** [Proof of Lemma 26] We obtain the first equality by applying the change-of-variables formula to Eq. (4.13).

$$q(\theta) = |\det(\delta_n)|^{-1} \tilde{q}(h), \quad \text{and} \quad q_i(\theta) = \delta_{n,ii}^{-1} \tilde{q}_i(h), i \in [D],$$

For the second equality, we have

$$\mathbb{H}(\tilde{q}) = \int |\det(\delta_n)| q(\theta) \log(|\det(\delta_n)| q(\theta)) dh = \int q(\theta) \log q(\theta) d\theta + \log |\det(\delta_n)|^{-1}.$$

The univariate case follows from this.

For the third equality, we can write

$$\begin{aligned} \Xi(\tilde{q}) &= \mathbb{H}(\tilde{q}) - \sum_{i=1}^D \mathbb{H}(\tilde{q}_i) \\ &= \mathbb{H}(q) - \sum_{i=1}^D \mathbb{H}(\tilde{q}_i) - \log |\det(\delta_n)| + \sum_{i=1}^D \log \delta_{n,ii} = \Xi(q) - \log |\det(\delta_n)| + \sum_{i=1}^D \log \delta_{n,ii}. \end{aligned}$$

For the fourth equality, we have

$$\begin{aligned} D_{\text{KL}}(\tilde{q}_1 \parallel \tilde{q}_2) &= \mathbb{H}(\tilde{q}_1) - \int \log \tilde{q}_2(h) \tilde{q}_1(h) dh \\ &= \mathbb{H}(q_1) + \log |\det(\delta_n)| - \int q_2(\theta) q_1(\theta) d\theta - \log |\det(\delta_n)| \\ &= D_{\text{KL}}(q_1 \parallel q_2). \end{aligned}$$

This concludes the proof. ■

To establish the Bernstein von-Mises theorem, we introduce the tool of  $\Gamma$ -convergence (Braides, 2014).

**Definition 27 ( $\Gamma$ -Convergence)** *Let  $X$  be a metric space and consider a set of functionals  $F_\varepsilon : X \rightarrow \mathbb{R}$  indexed by  $\varepsilon > 0$ . A limiting functional  $F_0$  exists and is called the  $\Gamma$ -limit of  $F_\varepsilon$  as  $\varepsilon \rightarrow 0$ , if the following conditions are met:*

1. **Liminf Inequality:** For all  $x \in X$  and for every sequence  $x_\varepsilon \rightarrow x$ ,

$$F_0(x) \leq \liminf_{\varepsilon \rightarrow 0} F_\varepsilon(x_\varepsilon).$$

2. **Limsup Inequality / Existence of a Recovery Sequence:** For each  $x \in X$ , there exists a sequence  $\bar{x}_\varepsilon \rightarrow x$  such that

$$F_0(x) \geq \limsup_{\varepsilon \rightarrow 0} F_\varepsilon(\bar{x}_\varepsilon).$$

The first condition requires  $F_0$  to be asymptotically upper bounded by  $F_\varepsilon$ . When paired with the second condition, it ensures that  $F_0(x) = \lim_{\varepsilon \rightarrow 0} F_\varepsilon(\bar{x}_\varepsilon)$ , thereby confirming that the lower bound is tight.

**Definition 28 (Equi-Coerciveness of Functionals)** *A sequence of functionals  $F_\varepsilon : X \rightarrow \mathbb{R}$  is said to be equi-coercive if for every bounded sequence  $x_\varepsilon$  with  $F_\varepsilon(x_\varepsilon) \leq t$ , there exists a subsequence  $x_j$  of  $x_\varepsilon$  and a converging sequence  $x'_j$  satisfies  $F_{\varepsilon_j}(x'_j) \leq F_{\varepsilon_j}(x_j) + o(1)$ .*

Equi-coerciveness ensures the existence of a precompact minimizing sequence for  $F_\varepsilon$ , which helps establish the convergence  $x_\varepsilon \rightarrow x$ .

**Theorem 29 (Fundamental Theorem of  $\Gamma$ -Convergence)** *Let  $X$  be a metric space and  $F_\varepsilon$  an equi-coercive sequence of functionals. If  $F = \Gamma\text{-}\lim_{\varepsilon \rightarrow 0} F_\varepsilon$ , then*

$$\arg \min_{x \in X} F = \lim_{\varepsilon \rightarrow 0} \arg \min_{x \in X} F_\varepsilon.$$

This theorem implies that if minimizers  $x_\varepsilon$  for all  $F_\varepsilon$  exist, the sequence converges, potentially along a subsequence, to a minimizer of  $F$ . We note that the converse is not necessarily true; there may exist minimizers for  $F$  which are not limits of minimizers for  $F_\varepsilon$ .

Note that when  $\delta_n = \lambda_n^{1/2} \delta_n$ , we have  $|\det(\delta_n)| = \lambda_n^{D/2} |\det(\delta_n)|$  and  $\delta_{n,ii} = \lambda_n^{1/2} \delta_{n,ii}$ .

We can explicitly characterize the transformed variational posterior:

$$\tilde{q}_{\lambda_n}(h) := |\det(\delta_n)| q_{\lambda_n}^*(\theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0}), \quad (\text{E.1})$$

where  $q_\lambda^*$  is the original  $\Xi$ -variational posterior.

**Lemma 30** *Under Definition Eq. (E.1), the distribution  $\tilde{q}_\lambda$  solves the following variational problem*

$$\tilde{q}_\lambda = \arg \min_{q \in \mathbb{P}_2(\Theta)} D_{KL}(q \parallel \tilde{q}_0) + \lambda_n \Xi(q).$$

This Lemma is a direct consequence of the transformation identities (Lemma 26) and Eq. (4.1), thus the proof is omitted.

**Proof** [Proof of Theorem 6] WLOG, we assume that  $\Theta = \mathbb{R}^D$ . Otherwise, we use the same proof by adding an indicator of the minimizing set to the sequence of functionals.

**Regime 1:**  $\lambda_n \rightarrow \infty$ . It suffices to show

$$F_n(q) := D_{KL}(q \parallel \tilde{q}_0) + \lambda_n \Xi(q),$$

$\Gamma$ -converge to

$$F_0(q) := D_{KL}(q \parallel N(0, V_{\theta_0}^{-1})) + \infty \Xi(q),$$

in  $[P_{\theta_0}]$ -probability as  $n \rightarrow \infty$ .

By Theorem 29,  $\Gamma$  convergence implies  $W_2(\tilde{q}_{\lambda_n}, \arg \min_{q \in \mathbb{P}_2(\Theta)} F_0(q)) \xrightarrow{P_{\theta_0}} 0$ , where  $q_0$  is the minimizer of  $F_0$ .

To prove the  $\Gamma$ -convergence, we rewrite  $F_n$ .

$$\begin{aligned} F_n(q) &:= D_{KL}(q \parallel \tilde{q}_0) + \lambda_n \Xi(q) \\ &= \mathbb{E}_q \left[ -\ell \left( \mathbf{x}^{(n)}; \theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0} \right) \right] + D_{KL}(q \parallel \tilde{\pi}) + \log |\det(\delta_n)| + \lambda_n \Xi(q) \\ &\quad + \int \pi(\theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0}) \ell(\mathbf{x}^{(n)}; \theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0}) dh. \\ &= -\ell(\mathbf{x}^{(n)}; \theta_0) + \mathbb{E}_q \left[ \frac{1}{2} h^T V_{\theta_0} h \right] + \mathbb{H}(q) - \mathbb{E}_q [\log \pi(\theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0})] + \lambda_n \Xi(q) \\ &\quad + \log \int \pi(\theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0}) \ell(\mathbf{x}^{(n)}; \theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0}) dh + o_P(1). \end{aligned}$$

Applying LAN expansion and Laplace approximation to the log-normalizer, we have

$$\begin{aligned} & \log \int \pi(\theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0}) \ell(\mathbf{x}^{(n)}; \theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0}) dh \\ &= \frac{D}{2} \log 2\pi - \frac{1}{2} \log \det(V_{\theta_0}) + \log \pi(\theta_0) + \ell(\mathbf{x}^{(n)}; \theta_0) + o_P(1). \end{aligned}$$

After cancellation, we have

$$\begin{aligned} F_n(\mathbf{q}) &= \mathbb{E}_{\mathbf{q}} \left[ \frac{1}{2} h^T V_{\theta_0} h \right] + \mathbb{H}(\mathbf{q}) + \lambda_n \Xi(\mathbf{q}) + \left( \frac{D}{2} \log 2\pi - \frac{1}{2} \log \det(V_{\theta_0}) \right) \\ &\quad - \{ \mathbb{E}_{\mathbf{q}} [\log \pi(\theta_0 + \delta_n h + \delta_n \Delta_{n,\theta_0})] - \log \pi(\theta_0) \} + o_P(1). \end{aligned}$$

Using Assumption 2 to bound the prior tail via Taylor expansion, we have an expression for  $F_n$

$$\begin{aligned} F_n(\mathbf{q}) &= \mathbb{E}_{\mathbf{q}} \left[ \frac{1}{2} h^T V_{\theta_0} h \right] + \mathbb{H}(\mathbf{q}) + \lambda_n \Xi(\mathbf{q}) + \frac{D}{2} \log 2\pi - \frac{1}{2} \log \det(V_{\theta_0}) + o_P(1). \quad (\text{E.2}) \\ &= D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) + \lambda_n \Xi(\mathbf{q}) + o_P(1). \end{aligned}$$

Now we rewrite  $F_0(\mathbf{q})$ .

$$\begin{aligned} F_0(\mathbf{q}) &:= D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) + \infty \Xi(\mathbf{q}) \\ &= \mathbb{E}_{\mathbf{q}} \left[ \frac{1}{2} h^T V_{\theta_0} h \right] + \mathbb{H}(\mathbf{q}) + \infty \Xi(\mathbf{q}) + \frac{D}{2} \log 2\pi - \frac{1}{2} \log \det(V_{\theta_0}). \\ &= F_n(\mathbf{q}) + \infty \Xi(\mathbf{q}) + o_P(1). \end{aligned}$$

Now we prove the  $\Gamma$  convergence.

First, we verify the liminf inequality. Let  $\mathbf{q}_n \xrightarrow{W_2} \mathbf{q}$ . When  $\mathbf{q}$  is not mean-field, we have:

$$\begin{aligned} \liminf_{n \rightarrow \infty} F_n(\mathbf{q}_n) &\geq \liminf_{n \rightarrow \infty} \left\{ D_{\text{KL}}(\mathbf{q}_n \parallel N(0, V_{\theta_0}^{-1})) + \lambda_n \Xi(\mathbf{q}_n) \right\} - \epsilon \\ &\geq \liminf_{n \rightarrow \infty} D_{\text{KL}}(\mathbf{q}_n \parallel N(0, V_{\theta_0}^{-1})) + \liminf_{n \rightarrow \infty} \lambda_n \liminf_{n \rightarrow \infty} \Xi(\mathbf{q}_n) - \epsilon \\ &\geq D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) + \infty \Xi(\mathbf{q}) - \epsilon = \infty \geq F_0(\mathbf{q}). \end{aligned}$$

The second inequality follows from the definition of liminf. The third line is due to Lemma 21, which states that the KL functional and  $\Xi$  functional are lower semicontinuous.

$$\begin{aligned} \liminf_{n \rightarrow \infty} F_n(\mathbf{q}_n) &\geq \liminf_{n \rightarrow \infty} \left\{ D_{\text{KL}}(\mathbf{q}_n \parallel N(0, V_{\theta_0}^{-1})) + \lambda_n \Xi(\mathbf{q}_n) \right\} - \epsilon \\ &\geq \liminf_{n \rightarrow \infty} D_{\text{KL}}(\mathbf{q}_n \parallel N(0, V_{\theta_0}^{-1})) - \epsilon \\ &\geq D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) - \epsilon. \end{aligned}$$

Since this holds for all  $\epsilon$ , we verified that  $\liminf_{n \rightarrow \infty} F_n(\mathbf{q}_n) \geq F_0(\mathbf{q})$ .

Next, we show the existence of a recovery sequence. When  $q$  is not mean-field,  $F_0(q) = +\infty$ , and the limsup inequality is automatically satisfied. When  $q$  is mean-field, choose  $q_n := q$ , then:

$$\limsup_{n \rightarrow \infty} F_n(q_n) = \limsup_{n \rightarrow \infty} D_{\text{KL}}(q \parallel N(0, V_{\theta_0}^{-1})) + o_P(1) \leq F_0(q).$$

Thus,  $F_0$  is the  $\Gamma$ -limit of the sequence  $F_n$ .

Next we prove that the sequence  $F_n$  is eqi-coercive. Take  $n_j \rightarrow \infty$  and  $q_{n_j}$  such that  $F_{n_j}(q_{n_j}) \leq t$  for all  $j$ . Then  $\lambda_{n_j} \Xi(q_{n_j})$  is bounded as  $\lambda_{n_j} \rightarrow \infty$ , thus  $\Xi(q_{n_j}) = o(1)$ . Using this and Eq. (E.2), we have

$$D_{\text{KL}}(q_{n_j} \parallel N(0, V_{\theta_0}^{-1})) \leq t + 1, \quad \text{for sufficiently large } j.$$

Since  $D_{\text{KL}}(\cdot \parallel N(0, V_{\theta_0}^{-1}))$  is a Wasserstein (geodesically) convex functional, it is coercive by Lemma 2.4.8 of Ambrosio et al. (2005). This implies that the set  $\{q \in \mathbb{P}_2(\Theta) \mid D_{\text{KL}}(q \parallel \tilde{q}_0) \leq t + 1\}$  is compact under the Wasserstein metric, thus  $q_{n_j}$  has a subsequence  $q'_{n_j}$  that converges to  $q^*$  in the Wasserstein metric and  $D_{\text{KL}}(q^* \parallel \tilde{q}_0) \leq t + 1$ . Thus we have  $F_{n_j}(q'_{n_j}) \leq F_{n_j}(q'_{n_j}) + o(1)$  by Eq. (E.2) where  $q'_{n_j}$  is a converging subsequence of  $q_{n_j}$ . This verifies the equi-coercivity of  $F_n$ .

Lastly, we note that  $F_0$  attains its minimum at  $N(\Delta_{n, \theta_0}, V_{\theta_0}^{'-1})$  where  $V_{\theta_0}^{'-1}$  is the MFVI covariance. As a result of Theorem 29, we conclude that the desired convergence takes place:

$$D_{\text{KL}}(q_{n_j} \parallel N(0, V_{\theta_0}^{-1})) \leq t + 1, \quad \text{for sufficiently large } j.$$

**Regime 2:**  $\lambda_n \rightarrow 0$ .

In this regime, we show that the functionals

$$F_n(q) := D_{\text{KL}}(q \parallel \tilde{q}_0) + \lambda_n \Xi(q),$$

$\Gamma$ -converge to

$$F_0(q) := D_{\text{KL}}(q \parallel N(0, V_{\theta_0}^{-1})),$$

in  $[P_{\theta_0}]$ -probability as  $n \rightarrow \infty$ .

Given that  $F_n$  is defined analogous to Regime 1, we will skip the derivation:

$$\begin{aligned} F_n(q) &= \mathbb{E}_q \left[ \frac{1}{2} h^T V_{\theta_0} h \right] + \mathbb{H}(q) + \lambda_n \Xi(q) + \frac{D}{2} \log 2\pi - \frac{1}{2} \log \det(V_{\theta_0}) + o_P(1). \\ &= D_{\text{KL}}(q \parallel N(0, V_{\theta_0}^{-1})) + \lambda_n \Xi(q) + o_P(1). \end{aligned}$$

Now we prove the  $\Gamma$  convergence. First, we verify the liminf inequality. Let  $q_n \xrightarrow{W_2} q$ . We have:

$$\begin{aligned} \liminf_{n \rightarrow \infty} F_n(q_n) &\geq \liminf_{n \rightarrow \infty} \left\{ D_{\text{KL}}(q_n \parallel N(0, V_{\theta_0}^{-1})) + \lambda_n \Xi(q_n) \right\} - \epsilon \\ &\geq \liminf_{n \rightarrow \infty} D_{\text{KL}}(q_n \parallel N(0, V_{\theta_0}^{-1})) + \liminf_{n \rightarrow \infty} \lambda_n \liminf_{n \rightarrow \infty} \Xi(q_n) - \epsilon \\ &\geq D_{\text{KL}}(q \parallel N(0, V_{\theta_0}^{-1})) - \epsilon. \end{aligned}$$

Since this holds for all  $\epsilon$ , we verified that  $\liminf_{n \rightarrow \infty} F_n(\mathbf{q}_n) \geq F_0(\mathbf{q})$ .

For the recovery sequence, we take  $\mathbf{q}_n := \mathbf{q}$ . Since  $\mathbf{q}$  is absolutely continuous with respect to the product of its marginals,  $\Xi(\mathbf{q})$  is finite. Then we have:

$$\limsup_{n \rightarrow \infty} F_n(\mathbf{q}_n) = \limsup_{n \rightarrow \infty} D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) + o_P(1) \leq F_0(\mathbf{q}).$$

The equicoercivity of  $F_n$  follows from the argument in regime 1. By Theorem 29, we conclude with the desired convergence:

$$W_2(\tilde{\mathbf{q}}_{\lambda_n}, N(0, V_{\theta_0}^{'-1})) \rightarrow 0.$$

**Regime 3:**  $\lambda_n \rightarrow \lambda_\infty \in (0, \infty)$ .

In this regime, we show that the functionals

$$F_n(\mathbf{q}) := D_{\text{KL}}(\mathbf{q} \parallel \tilde{\mathbf{q}}_0) + \lambda_n \Xi(\mathbf{q}),$$

$\Gamma$ -converge to

$$F_0(\mathbf{q}) := D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) + \lambda_\infty \Xi(\mathbf{q}),$$

in  $[P_{\theta_0}]$ -probability as  $n \rightarrow \infty$ .

Recall that

$$\begin{aligned} F_n(\mathbf{q}) &= \mathbb{E}_{\mathbf{q}} \left[ \frac{1}{2} \mathbf{h}^T V_{\theta_0} \mathbf{h} \right] + \mathbb{H}(\mathbf{q}) + \lambda_n \Xi(\mathbf{q}) + \frac{D}{2} \log 2\pi - \frac{1}{2} \log \det(V_{\theta_0}) + o_P(1). \\ &= D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) + \lambda_n \Xi(\mathbf{q}) + o_P(1). \end{aligned}$$

Now we prove the  $\Gamma$  convergence. First, we verify the liminf inequality. Let  $\mathbf{q}_n \xrightarrow{W_2} \mathbf{q}$ . We have:

$$\begin{aligned} \liminf_{n \rightarrow \infty} F_n(\mathbf{q}_n) &\geq \liminf_{n \rightarrow \infty} \left\{ D_{\text{KL}}(\mathbf{q}_n \parallel N(0, V_{\theta_0}^{-1})) + \lambda_n \Xi(\mathbf{q}_n) \right\} - \epsilon \\ &\geq \liminf_{n \rightarrow \infty} D_{\text{KL}}(\mathbf{q}_n \parallel N(0, V_{\theta_0}^{-1})) + \liminf_{n \rightarrow \infty} \lambda_n \liminf_{n \rightarrow \infty} \Xi(\mathbf{q}_n) - \epsilon \\ &\geq D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) + \lambda_\infty \Xi(\mathbf{q}) - \epsilon. \end{aligned}$$

The second inequality follows from the definition of liminf, and the last inequality is due to Lemma 21, which states that the KL functional and  $\Xi$  functional are lower semicontinuous.

For the recovery sequence, we take  $\mathbf{q}_n := \mathbf{q}$ . As long as  $\Xi(\mathbf{q})$  is finite, we have:

$$\limsup_{n \rightarrow \infty} F_n(\mathbf{q}_n) = \limsup_{n \rightarrow \infty} D_{\text{KL}}(\mathbf{q} \parallel N(0, V_{\theta_0}^{-1})) + \lambda_n \Xi(\mathbf{q}) + o_P(1) = F_0(\mathbf{q}).$$

The equicoercivity of  $F_n$  follows from the argument in regime 1. By Theorem 29, we have the convergence:

$$W_2(\tilde{\mathbf{q}}_{\lambda_n}, \arg \min_{\mathbf{q} \in \mathbb{P}_2(\Theta)} F_0(\mathbf{q})) \rightarrow 0.$$

■

**Proof** [Proof of Corollary 7] Recall the definition of Wasserstein distance:

$$W_2(p, q) = \left( \inf_{\pi \in \mathcal{C}(p, q)} \mathbb{E}_{\pi} [\|X - Y\|^2] \right)^{1/2}.$$

Given the change-of-variables definition (Eq. (4.13)), we have

$$\begin{aligned} W_2(\tilde{q}_{\lambda_n}, N(\mu, \Sigma)) &= \left( \inf_{\pi \in \mathcal{C}(\tilde{q}_{\lambda_n}, N(\mu, \Sigma))} \mathbb{E}_{\pi} [\|h - h'\|^2] \right)^{1/2} \\ &= |\det(\delta_n)|^{-1} \left( \inf_{\pi \in \mathcal{C}(q_{\lambda_n}^*, N(\delta_n \mu + \theta_0 + \delta_n \Delta_{n, \theta_0}, \delta_n^T \Sigma \delta_n))} \mathbb{E}_{\pi} [\|\theta - \theta'\|^2] \right)^{1/2} \\ &= |\det(\delta_n)|^{-1} W_2(q_{\lambda_n}^*, N(\delta_n \mu + \theta_0 + \delta_n \Delta_{n, \theta_0}, \delta_n^T \Sigma \delta_n)). \end{aligned}$$

If  $W_2(\tilde{q}_{\lambda_n}, N(\mu, \Sigma))$  tends to 0, then  $W_2(q_{\lambda_n}^*, N(\delta_n \mu + \theta_0 + \delta_n \Delta_{n, \theta_0}, \delta_n^T \Sigma \delta_n))$  tends to 0. Since  $N(\delta_n \mu + \theta_0 + \delta_n \Delta_{n, \theta_0}, \delta_n^T \Sigma \delta_n)$  weakly converge to  $\delta_{\theta_0}$ , it converges to  $\delta_{\theta_0}$  in Wasserstein metric. By Theorem 6, we have  $q_{\lambda_n}^*$  converges in Wasserstein metric to  $\delta_{\theta_0}$ , as desired.  $\blacksquare$

## Proofs of Section 4.1

We first prove a useful proposition.

**Proposition 31 (Optimality to fixed point)** *Let  $m_{\lambda}^*(\theta) = \prod_{i=1}^D m_{\lambda, i}^*(\theta_i)$  be the product of optimal marginals, and  $\phi_{\lambda}^*$  be the optimal EOT potentials. Then  $m_{\lambda}^*$  and  $\phi_{\lambda}^*$  satisfy the fixed point equations:*

$$\begin{aligned} m_{\lambda, i}^*(\theta_i) &= Z_i^{-1} \exp(-(\lambda + 1)\phi_{\lambda, i}^*(\theta_i))\pi_i(\theta_i), \quad \text{and} \\ \phi_{\lambda, i}^*(\theta_i) &= -\log \int_{\Theta_{-i}} \exp \left( \frac{1}{\lambda + 1} \ell(\mathbf{x}^{(n)}; \theta) - \lambda \sum_{j \neq i} \phi_{\lambda, j}^*(\theta_j) \right) \prod_{j \neq i} \pi_j(\theta_j) d\theta_{-i} \\ &\quad + \sum_{j \neq i} \log \int_{\Theta_j} \exp(-(\lambda + 1)\phi_{\lambda, j}^*(\theta_j)) \pi_j(\theta_j) d\theta_j, \end{aligned} \tag{E.3}$$

where  $Z_i$ 's are the normalizing constants.

**Proof** [Proof of Proposition 31] Define  $f_{\lambda}(\theta)$  as follows,

$$f_{\lambda}(\theta) := -(\lambda + 1) \left( \sum_{i=1}^D \phi_{\lambda, i}^*(\theta_i) \right) \exp \left( \sum_{i=1}^D \phi_{\lambda, i}^*(\theta_i) + \frac{1}{\lambda + 1} \ell(\mathbf{x}^{(n)}; \theta) \right) + \sum_{i=1}^D \log \pi_i(\theta_i). \tag{E.4}$$

By Theorem 25 and the uniform boundedness of  $\ell(\mathbf{x}^{(n)}; \cdot)$ , the function  $f_{\lambda}$  is integrable with respect to  $m_{\lambda}^*$ . From the derivation in Section 2.2, the distribution  $m_{\lambda}^*$  attain the minimum,

$$\min_{m \in \mathbb{M}(\Theta)} -\mathbb{E}_m [f_{\lambda}(\theta)] + \mathbb{H}(m). \tag{E.5}$$

Define  $\hat{f}_{\lambda, i}(\theta_i) = \mathbb{E}_{m_{\lambda}^*}[f_{\lambda}(\theta) \mid \theta_i]$  for  $[m_{\lambda, i}^*]$ -a.s.  $\theta_i \in \Theta_i$ . Since  $m \in \mathbb{M}(\Theta)$ ,  $\mathbb{H}(m) = \sum_{i=1}^D \mathbb{H}(m_i)$  and by the tower property, we have

$$m_{\lambda, i}^*(\theta_i) := \arg \min_{m_i \in \mathbb{M}(\Theta_i)} \left( \mathbb{E}_{m_i} [-\hat{f}_{\lambda, i}(\theta_i)] + \mathbb{H}(m_i) \right). \tag{E.6}$$

By the Gibbs variational principle (Lemma 17), the minimum is uniquely attained by

$$m_{\lambda,i}^*(\theta_i) \propto \exp(\hat{f}_{\lambda,i}(\theta_i)). \quad (\text{E.7})$$

Recall that the optimal EOT potentials satisfy the Schrödinger system:

$$\phi_{\lambda,i}^*(\theta_i) = -\log \int_{\Theta_{-i}} \exp \left( \sum_{j \neq i} \phi_{\lambda,j}^*(\theta_j) + \frac{1}{\lambda+1} \ell(\mathbf{x}^{(n)}; \theta) \right) \prod_{j \neq i} m_{\lambda,j}^*(\theta_j) d\theta_{-i}. \quad (\text{E.8})$$

This allows us to simplify  $\hat{f}_{\lambda,i}(\theta_i)$ :

$$\begin{aligned} \hat{f}_{\lambda,i}(\theta_i) &= \mathbb{E}_{m_{\lambda}^*}[f_{\lambda}(\theta) \mid \theta_i] \\ &= \mathbb{E}_{m_{\lambda}^*} \left[ -(\lambda+1) \left( \sum_{i=1}^D \phi_{\lambda,i}^*(\theta_i) \right) \exp \left( \sum_{i=1}^D \phi_{\lambda,i}^*(\theta_i) + \frac{1}{\lambda+1} \ell(\mathbf{x}^{(n)}; \theta) \right) + \sum_{i=1}^D \log \pi_i(\theta_i) \mid \theta_i \right] \\ &= -(\lambda+1) \frac{\mathbb{E}_{m_{\lambda,-i}^*} \left[ \left( \sum_{i=1}^D \phi_{\lambda,i}^*(\theta_i) \right) \exp \left( \sum_{j \neq i} \phi_{\lambda,j}^*(\theta_j) + \frac{1}{\lambda+1} \ell(\mathbf{x}^{(n)}; \theta) \right) \right]}{\exp \left( -\phi_{\lambda,i}^*(\theta_i) \right)} \\ &\quad + \mathbb{E}_{m_{\lambda,-i}^*} \left[ \sum_{j \neq i} \log \pi_j(\theta_j) \right] + \log \pi_i(\theta_i) \\ &= -(\lambda+1)\phi_{\lambda,i}^*(\theta_i) + \log \pi(\theta_i) - (\lambda+1)\mathbb{E}_{\hat{h}_{\lambda}} \left[ \sum_{j \neq i} \phi_{\lambda,j}^*(\theta_j) \right] + C. \end{aligned}$$

where  $\hat{h}_{\lambda}(\theta_{-i}) \propto \exp \left( \sum_{j \neq i} \phi_{\lambda,j}^*(\theta_j) + \frac{1}{\lambda+1} \ell(\mathbf{x}^{(n)}; \theta) \right) \prod_{j \neq i} m_{\lambda,j}^*(\theta_j)$ . Since  $\hat{h}_{\lambda}(\theta_{-i}) \propto q_{\lambda}^*(\theta_{-i}, \theta_i)$ , we have for all  $\theta_i$ ,

$$\mathbb{E}_{\hat{h}_{\lambda}(\theta_{-i})} \left[ \sum_{j \neq i} \phi_{\lambda,j}^*(\theta_j) \right] = \mathbb{E}_{q_{\lambda}^*(\theta_{-i}, \theta_i)} \left[ \sum_{j \neq i} \phi_{\lambda,j}^*(\theta_j) \right] = \sum_{j \neq i} \mathbb{E}_{m_{\lambda,j}^*} [\phi_{\lambda,j}^*(\theta_j)]. \quad (\text{E.9})$$

The last equality uses the fact that  $m_{\lambda,i}^*$  is the  $i^{\text{th}}$  marginal of  $q_{\lambda}^*$ .

Since  $\sum_{j \neq i} \mathbb{E}_{m_{\lambda,j}^*} [-(\lambda+1)\phi_{\lambda,j}^*(\theta_j) + \log \pi_j(\theta_j)]$  does not depend on  $\theta_i$ , we obtain

$$m_{\lambda,i}^*(\theta_i) \propto \exp(-(\lambda+1)\phi_{\lambda,i}^*(\theta_i))\pi_i(\theta_i). \quad (\text{E.10})$$

Using Eq. (E.10), we conclude

$$\begin{aligned} \phi_{\lambda,i}^*(\theta_i) &= -\log \int_{\Theta_{-i}} \exp \left( \frac{1}{\lambda+1} \ell(\mathbf{x}^{(n)}; \theta) - \lambda \sum_{j \neq i} \phi_{\lambda,j}^*(\theta_j) \right) \prod_{j \neq i} \pi_j(\theta_j) d\theta_{-i} \\ &\quad + \sum_{j \neq i} \log \int_{\Theta_j} \exp(-(\lambda+1)\phi_{\lambda,j}^*(\theta_j)) \pi_j(\theta_j) d\theta_j. \end{aligned}$$

■

**Proof** [Proof of Theorem 2] Since the parameter space  $\Theta$  is compact and  $\ell(\mathbf{x}^{(n)}; \cdot) \in \mathbb{C}^2(\Theta)$ , both the gradient  $\nabla \ell(\mathbf{x}^{(n)}; \cdot)$  and the Hessian  $\nabla^2 \ell(\mathbf{x}^{(n)}; \cdot)$  are uniformly bounded over  $\Theta$ .

Let  $u_i := \frac{1}{2} \inf_{\theta \in \Theta} [\nabla^2 \ell(\mathbf{x}^{(n)}; \theta)]_{ii}$ ,  $v_i$  is chosen such that  $\sup_{\theta \in \Theta} |[\nabla \ell(\mathbf{x}^{(n)}; \theta)]_i - v_i - 2u_i \theta_i| = b_i$ , and  $w$  is chosen such that  $\sup_{\theta \in \Theta} |\ell(\mathbf{x}^{(n)}; \theta) - w - \sum_i v_i \theta_i - \sum_i u_i \theta_i^2| = a$ . We define a new log-likelihood  $\tilde{\ell}(\mathbf{x}^{(n)}; \theta)$  that shift  $\ell(\mathbf{x}^{(n)}; \theta)$  by a quadratic function:

$$\tilde{\ell}(\mathbf{x}^{(n)}; \theta) := \ell(\mathbf{x}^{(n)}; \theta) - w - \sum_{i=1}^D v_i \theta_i - \sum_{i=1}^D u_i \theta_i^2.$$

Some calculation yields that

$$\sup_{\theta \in \Theta} |\tilde{\ell}(\mathbf{x}^{(n)}; \theta)| = a, \quad \sup_{\theta \in \Theta} |[\nabla \tilde{\ell}(\mathbf{x}^{(n)}; \theta)]_i| = b_i, \quad \sup_{\theta \in \Theta} |[\nabla^2 \tilde{\ell}(\mathbf{x}^{(n)}; \theta)]_{ii}| = c_i.$$

Given the optimal  $m_{\lambda_n}^*$ , the inner variational problem is

$$q_{\lambda_n}^* = \arg \min_{q \in \mathcal{C}(m_{\lambda_n}^*)} \left\{ -\mathbb{E}_q[\tilde{\ell}(\mathbf{x}^{(n)}; \theta)] + (\lambda_n + 1) D_{\text{KL}}(q \parallel m_{\lambda_n}^*) \right\},$$

where we replace  $\ell(\mathbf{x}^{(n)}; \theta)$  with  $\tilde{\ell}(\mathbf{x}^{(n)}; \theta)$  since adding a tensorized function  $-(w + \sum_{i=1}^D v_i \theta_i + \sum_{i=1}^D u_i \theta_i^2)$  to the cost does not change the optimal EOT coupling.

By Theorem 25, we can write  $q_{\lambda_n}^*$  as follows,

$$q_{\lambda_n}^*(\theta) = \exp \left( \frac{1}{\lambda_n + 1} \tilde{\ell}(\mathbf{x}^{(n)}; \theta) + \sum_{i=1}^D \phi_{\lambda_n, i}^*(\theta_i) \right) \prod_{i=1}^D m_{\lambda_n, i}^*(\theta)$$

where  $m_{\lambda_n, i}^*$ 's are the marginals of  $q_{\lambda_n}^*$  and  $\phi_{\lambda_n, i}^*$ 's are the EOT potentials.

Define another product distribution  $\tilde{m}_{\lambda_n}(\theta) \propto \exp \left( \sum_{i=1}^D \phi_{\lambda_n, i}^*(\theta_i) \right) m_{\lambda_n, i}^*(\theta)$ . We can rewrite  $q_{\lambda_n}^*$  as the product of a tempered likelihood and a  $\tilde{m}_{\lambda_n}$ .

$$q_{\lambda_n}^*(\theta) = \frac{1}{\mathcal{Z}_D(\lambda_n)} \exp \left( \frac{1}{\lambda_n + 1} \tilde{\ell}(\mathbf{x}^{(n)}; \theta) \right) \tilde{m}_{\lambda_n}(\theta), \quad (\text{E.11})$$

where the normalizing constant is given by

$$\mathcal{Z}_D(\lambda_n) := \int_{\Theta} \exp \left( \frac{1}{\lambda_n + 1} \tilde{\ell}(\mathbf{x}^{(n)}; \theta) \right) \tilde{m}_{\lambda_n}(\theta) d\theta. \quad (\text{E.12})$$

First, we want to show that

$$\lim_{n \rightarrow \infty} \frac{1}{D} \left[ \log \mathcal{Z}_D(\lambda_n) - \sup_{m \in \mathbb{M}(\Theta), m \ll \tilde{m}_{\lambda_n}} \left\{ \frac{1}{\lambda_n + 1} \mathbb{E}_m[\tilde{\ell}(\mathbf{x}^{(n)}; \theta)] - \sum_{i=1}^D D_{\text{KL}}(m_i \parallel \tilde{m}_{\lambda_n, i}) \right\} \right] = 0. \quad (\text{E.13})$$

Let  $\|f\|_\infty$  denote the supremum norm of a function  $f$ . Fix some  $\epsilon > 0$ . Let  $\mathcal{S}_{\lambda_n}(\epsilon) \subset \Theta$  be a finite set such that for any  $\theta \in \Theta$ , there exists  $s \in \mathcal{S}_{\lambda_n}(\epsilon)$  satisfying

$$\sum_{i=1}^D \left\| \frac{1}{\lambda_n + 1} \partial_i \tilde{\ell}(\mathbf{x}^{(n)}; \theta) - s_i \right\|_\infty^2 \leq \epsilon^2 D. \quad (\text{E.14})$$

Denote by  $|\mathcal{S}_{\lambda_n}(\epsilon)|$  the cardinality of  $\mathcal{S}_{\lambda_n}(\epsilon)$ . Theorem 1 of Yan (2020) implies that

$$\begin{aligned} \log \mathcal{Z}_D(\lambda_n) - \sup_{m \in \mathbb{M}(\Theta), m \ll \tilde{m}_{\lambda_n}} \left[ \frac{1}{\lambda_n + 1} \mathbb{E}_m[\tilde{\ell}(\mathbf{x}^{(n)}; \theta)] - \sum_{i=1}^D D_{\text{KL}}(m_i \parallel \tilde{m}_{\lambda_n, i}) \right] \leq \\ 4 \left( \frac{4}{(\lambda_n + 1)^2} \left( a \sum_{i=1}^D c_{ii} + \sum_{i=1}^D b_i^2 \right) + \frac{8}{(\lambda_n + 1)^2} \sum_{i=1}^D \sum_{j=1}^D b_i c_{ij} + \frac{16}{(\lambda_n + 1)^{3/2}} \left( a \sum_{i=1}^D \sum_{j=1}^D c_{ij}^2 + \sum_{i=1}^D \sum_{j=1}^D b_i b_j c_{ij} \right) \right)^{1/2} \\ + 4 \left( \frac{1}{(\lambda_n + 1)^2} \sum_{i=1}^D b_i^2 + \epsilon^2 D \right)^{1/2} \left( \frac{8}{\lambda_n + 1} \left( \sum_{i=1}^D c_{ii}^2 \right)^{1/2} + 4D^{1/2}\epsilon \right) + \frac{4}{\lambda_n + 1} \sum_{i=1}^D c_{ii} + 2D\epsilon \\ + \log 2 + \log |\mathcal{S}_{\lambda_n}(\epsilon)|. \end{aligned} \quad (\text{E.15})$$

Consider  $\lambda_n \succ D^{-1/2} \max \left( \sqrt{a \sum_{i=1}^D c_{ii}}, \sqrt{\sum_{i=1}^D b_i^2}, \sqrt{\sum_{i=1}^D \sum_{j=1}^D c_{ij}^2}, D^{1/2} \right)$ . Then,

$$\begin{aligned} \frac{\sum_{i=1}^D \sum_{j=1}^D b_i c_{ij}}{(\lambda_n + 1)^2} &\leq \frac{\sqrt{\sum_{i=1}^D b_i^2} \sqrt{\sum_{i=1}^D \sum_{j=1}^D c_{ij}^2}}{(\lambda_n + 1)^2} = o(D), \\ \frac{\sum_{i=1}^D \sum_{j=1}^D b_i b_j c_{ij}}{(\lambda_n + 1)^3} &\leq \frac{\sqrt{\sum_{i=1}^D \sum_{j=1}^D c_{ij}^2} \sum_{i=1}^D b_i^2}{(\lambda_n + 1)^3} = o(D^{3/2}). \end{aligned}$$

by the Cauchy–Schwarz inequality. With the other terms being  $o(D)$ , we have

$$\begin{aligned} \log \mathcal{Z}_D(\lambda_n) - \sup_{m \in \mathbb{M}(\Theta), m \ll \tilde{m}_{\lambda_n}} \left[ \frac{1}{\lambda_n + 1} \mathbb{E}_m[\tilde{\ell}(\mathbf{x}^{(n)}; \theta)] - \sum_{i=1}^D D_{\text{KL}}(m_i \parallel \tilde{m}_{\lambda_n, i}) \right] \\ \leq o(D) + 2D\epsilon + \log 2 + \log |\mathcal{S}_{\lambda_n}(\epsilon)|. \end{aligned} \quad (\text{E.16})$$

To upper bound  $|\mathcal{S}_{\lambda_n}(\epsilon)|$ , we can construct an  $\epsilon$ -covering by covering  $[-\frac{b_i}{\lambda_n + 1}, \frac{b_i}{\lambda_n + 1}]$  with balls of size  $2\epsilon$ . We consider a candidate set  $\tilde{\mathcal{S}}_{\lambda_n}(\epsilon)$  as the product of these coverings. Since  $|\tilde{\mathcal{S}}_{\lambda_n}(\epsilon)| = \frac{\prod_{i=1}^D b_i}{(\lambda_n + 1)^D \epsilon^D}$ , we have

$$\log |\mathcal{S}_{\lambda_n}(\epsilon)| \leq \sum_{i=1}^D \log b_i - D \log(\lambda_n + 1) - D \log \epsilon.$$

Define  $\bar{b} := \sum_{i=1}^D b_i$ . Since  $\lambda_n \succ D^{-1/2} \sqrt{\sum_{i=1}^D b_i^2}$ , we have  $D\lambda_n \succ D^{1/2} \sqrt{\sum_{i=1}^D b_i^2} \geq D\bar{b}$ . By Jensen's inequality,

$$D \log(\lambda_n + 1) \succ D \log(\bar{b} + 1) \geq \sum_{i=1}^D \log(b_i).$$

To complete the bound of  $\log |\mathcal{S}_{\lambda_n}(\epsilon)|$ , we choose a specific sequence  $\epsilon_n := \sqrt{\frac{\bar{b}+1}{\lambda_n+1}}$ . The inequality above shows that  $\epsilon = o(1)$ .

Thus,

$$\log |\mathcal{S}_{\lambda_n}(\epsilon_n)| \leq D \log(\bar{b} + 1) - D \log(\lambda_n + 1) - D \log \epsilon_n = \frac{1}{2} D \log \left( \frac{\bar{b} + 1}{\lambda_n + 1} \right) \rightarrow -\infty. \quad (\text{E.17})$$

Plugging the definition of  $\epsilon_n$  into Eq. (E.16), we get

$$\log \mathcal{Z}_D(\lambda_n) - \sup_{m \in \mathbb{M}(\Theta), m \ll \tilde{m}_{\lambda_n}} \left[ \frac{1}{\lambda_n + 1} \mathbb{E}_m[\tilde{\ell}(\mathbf{x}^{(n)}; \theta)] - \sum_{i=1}^D D_{\text{KL}}(m_i \parallel \tilde{m}_{\lambda_n, i}) \right] = o(D). \quad (\text{E.18})$$

For any  $m \in \mathbb{M}(\Theta)$ , we have

$$\begin{aligned} D_{\text{KL}}(m \parallel q_{\lambda_n}^*) &= \int_{\Theta} m(\theta) \left[ \log \mathcal{Z}_D(\lambda_n) - \frac{1}{\lambda_n + 1} \tilde{\ell}(\mathbf{x}^{(n)}; \theta) + \log \frac{m(\theta)}{\prod_{i=1}^D \tilde{m}_{\lambda_n, i}(\theta_i)} \right] d\theta. \\ &= \log \mathcal{Z}_D(\lambda_n) - \frac{1}{\lambda_n + 1} \mathbb{E}_m[\tilde{\ell}(\mathbf{x}^{(n)}; \theta)] + \sum_{i=1}^D D_{\text{KL}}(m_i \parallel \tilde{m}_{\lambda_n, i}(\theta_i)). \end{aligned}$$

Eq. (E.18) implies that for  $m_{\lambda_n}^* \in \arg \min_{m \in \mathbb{M}(\Theta)} D_{\text{KL}}(m \parallel q_{\lambda_n}^*)$ , we have

$$D_{\text{KL}}(m_{\lambda_n}^* \parallel q_{\lambda_n}^*) = o(D). \quad (\text{E.19})$$

For any 1-Lipschitz function  $f$  under the  $L_1$  norm, consider the random variable  $f(\theta)$ , where  $\theta \sim q_{\lambda_n}^*$ . This variable satisfies the inequality  $\log \mathbb{E}_{q_{\lambda_n}^*} [\exp(\langle t, f(\theta) - \mathbb{E}_{q_{\lambda_n}^*} [f(\theta)] \rangle)] \leq 2D\|t\|_2^2$ , which is derived from the assumption that  $\Theta = [-1, 1]^D$ . Thus,  $q_{\lambda_n}^*$  is  $(4D)$ -subGaussian. By the  $T_1$ -transportation inequality (Theorem 4.8, Van Handel (2014)), for any  $m \in \mathbb{M}(\Theta)$ , the following upper bound holds:

$$W_1(m, q_{\lambda_n}^*) \leq \sqrt{8D \cdot D_{\text{KL}}(m \parallel q_{\lambda_n}^*)} = o(D), \quad (\text{E.20})$$

where  $W_1$  is the 1-Wasserstein distance.

Let  $m_{\lambda_n}^*$  denote the minimizer of the left hand side Eq. (E.19). Consider a function  $\psi$  that is 1-Lipschitz on  $\mathbb{R}$ . The function  $\theta \mapsto \sum_{i=1}^D \psi(\theta_i)$  is also 1-Lipschitz with respect to the  $L_1$  norm. This follows from the inequality:

$$\left| \sum_{i=1}^D \psi(\theta_i) - \sum_{i=1}^D \psi(\theta'_i) \right| \leq \sum_{i=1}^D |\psi(\theta_i) - \psi(\theta'_i)| \leq \sum_{i=1}^D |\theta_i - \theta'_i| \leq \|\theta - \theta'\|_1. \quad (\text{E.21})$$

Applying Kantorovich duality, we obtain the bound:

$$\sup_{\psi \in \text{Lip}(\mathbb{R})} \left| \sum_{i=1}^D \left( \mathbb{E}_{q_{\lambda_n}^*} [\psi(\theta_i)] - \mathbb{E}_{m_{\lambda_n}^*} [\psi(\theta_i)] \right) \right| \leq W_1(q_{\lambda_n}^*, m_{\lambda_n}^*) = o(D). \quad (\text{E.22})$$

Since the bound in Eq. (E.15) does not depend on the value of  $\mathbf{x}^{(n)}$ , we have

$$\sup_{\mathbf{x}^{(n)} \in \mathbb{X}^n} \left| \frac{1}{D} \sum_{i=1}^D \left( \mathbb{E}_{q_{\lambda_n}^*} [\psi(\theta_i)] - \mathbb{E}_{m_{\lambda_n}^*} [\psi(\theta_i)] \right) \right| \leq \frac{1}{D} W_1(q_{\lambda_n}^*, m_{\lambda_n}^*) \xrightarrow{P_{\theta_0}} 0. \quad (\text{E.23})$$

Consider the regime  $\lambda_n \prec D\Xi^{-1}(q_0^*)$ . Recall that the the  $\Xi$ -VI has the Lagrangian formulation as  $\min_{\Xi(q) \leq r(\lambda_n)} D_{\text{KL}}(q \parallel q_0^*)$  for some constant  $r(\lambda_n)$  depending on  $\lambda_n$ . If  $\Xi(q_0^*) \leq r(\lambda_n)$ , then  $q_{\lambda_n}^* = q_0^*$ , which implies  $\Xi(q_0^*) \geq \Xi(q_{\lambda_n}^*)$  for all  $\lambda_n$ . For fixed  $n$ , we have

$$D_{\text{KL}}(q_{\lambda_n}^* \parallel q_0^*) - D_{\text{KL}}(q_0^* \parallel q_0^*) \leq \lambda_n (\Xi(q_0^*) - \Xi(q_{\lambda_n}^*)) \leq \lambda_n \Xi(q_0^*) = o(D).$$

By the  $T_1$ -transportation inequality and Kantorovich duality, we have

$$\sup_{\mathbf{x}^{(n)} \in \mathbb{X}^n} \left| \frac{1}{D} \sum_{i=1}^D \left( \mathbb{E}_{q_{\lambda_n}^*} [\psi(\theta_i)] - \mathbb{E}_{q_0^*} [\psi(\theta_i)] \right) \right| \leq \frac{1}{D} W_1(q_{\lambda_n}^*, q_0^*) \lesssim \sqrt{D \cdot D_{\text{KL}}(q_{\lambda_n}^* \parallel q_0^*)} \xrightarrow{P_{\theta_0}} 0. \quad (\text{E.24})$$

■

**Proof** [Proof of Corollary 3] Under the assumptions,

$$\sqrt{\sum_{i=1}^D ac_{ii}/D} \lesssim D, \quad \sqrt{\sum_{i=1}^D b_i^2/D} \lesssim D, \quad \sqrt{\sum_{i,j} c_{ij}^2} \lesssim D.$$

When we plug these terms in the upper bounds (2), Eq. (4.3) follows as the desired result.

■

For the linear model, denote  $\mathbf{w} := \sigma^{-2} \mathbf{X}^T \mathbf{y}$  and  $d_i = [\mathbf{B}_{\text{diag}}]_{ii}$ , where  $[\mathbf{B}_{\text{diag}}]_{ii}$  is the  $i^{\text{th}}$  diagonal entry of matrix  $\mathbf{B}_{\text{diag}}$ . The next result shows  $\Xi$ -VI respects log-concavity of the exact posterior.

**Lemma 32** *Let the assumptions of Theorem 4 hold. For  $\lambda_n \in \bar{\mathbb{R}}_+$ , the solution  $q_{\lambda_n}^*$  to Eq. (4.1) is  $(\kappa_1 + \kappa_2)$ -log-concave. Moreover, for each  $i$ , the optimal EOT potential  $\phi_{\lambda_n, i}^*$  is  $\kappa_2/(\lambda_n + 1)$ -convex and marginal  $m_{\lambda_n, i}^*$  is  $(\kappa_1 + \kappa_2)$ -log-concave.*

**Proof** [Proof of Lemma 32] We first prove existence. By Lagrangian duality,  $\Xi$ -VI (Eq. (4.1)) is equivalent to  $\min_{\Xi(q) \leq r(\lambda_n)} D_{\text{KL}}(q \parallel q_0^*)$ . An optimizer of the latter problem exists because  $\Xi(\cdot)$  has weakly closed sublevel set in  $\mathbb{P}_2(\Theta)$  and because  $D_{\text{KL}}(\cdot \parallel q_0^*)$  has weakly compact sub-level sets.

Recall the  $\Xi$ -variational posterior be represented in term of optimal marginals  $m_{\lambda_n}^*$  and optimal EOT potentials  $\phi_{\lambda_n}^*$ :

$$q_{\lambda_n}^*(\theta) = \exp \left( \sum_{i=1}^D \phi_{\lambda_n, i}^*(\theta_i) + \frac{1}{\lambda_n + 1} \ell(\mathbf{x}^{(n)}; \theta) \right) m_{\lambda_n}^*(\theta). \quad (\text{E.25})$$

By Proposition 31,  $m_{\lambda_n}^*$  and  $\phi_{\lambda_n}^*$  satisfy the following fixed point equations:

$$\begin{aligned} m_{\lambda_n,i}^*(\theta_i) &= Z_i^{-1} \exp(-(\lambda_n + 1)\phi_{\lambda_n,i}^*(\theta_i))\pi_i(\theta_i), \quad \text{and} \\ \hat{\phi}_{\lambda_n,i}(\theta_i) &= -\log \int_{\Theta_{-i}} \exp \left( \frac{1}{\lambda_n + 1} \ell(\mathbf{x}^{(n)}; \theta) - \lambda_n \sum_{j \neq i} \phi_{\lambda_n,j}^*(\theta_j) \right) \prod_{j \neq i} \pi_j(\theta_j) d\theta_{-i} \quad (\text{E.26}) \\ &\quad + \sum_{j \neq i} \log \int_{\Theta_j} \exp(-(\lambda_n + 1)\phi_{\lambda_n,j}^*(\theta_j)) \pi_j(\theta_j) d\theta_j. \end{aligned}$$

Using equations Eq. (E.26) to replace  $m_{\lambda_n}^*$  in Eq. (E.25), the variational posterior  $q_{\lambda_n}^*$  satisfies

$$q_{\lambda_n}^*(\theta) \propto \exp \left( \frac{1}{\lambda_n + 1} \ell(\mathbf{x}^{(n)}; \theta) - \lambda_n \sum_{i=1}^D \phi_{\lambda_n,i}^*(\theta_i) \right) \pi(\theta). \quad (\text{E.27})$$

We now establish the log-concavity of  $q_{\lambda_n}^*$ . Applying Eq. (E.26) to Eq. (E.27), we get

$$q_{\lambda_n}^*(\theta) \propto \exp \left( \frac{1}{\lambda_n + 1} \ell(\mathbf{x}^{(n)}; \theta) + \lambda_n \sum_{i=1}^D \log \mathbb{E}_{\pi_{-i}} \exp \left( \frac{1}{\lambda_n + 1} \ell(\mathbf{x}^{(n)}; \theta) - \lambda_n \sum_{j \neq i} \phi_{\lambda_n,j}^*(\theta_j) \right) \right) \pi(\theta). \quad (\text{E.28})$$

For  $\alpha \in [0, 1]$  and  $\theta_i^0, \theta_i^1 \in \Theta_i$ , we have

$$-\hat{\phi}_{\lambda_n,i}(\alpha\theta_i^0 + (1 - \alpha)\theta_i^1) = \log \mathbb{E}_{\pi_{-i}} \left[ \exp \left( \frac{1}{\lambda_n + 1} \ell(\mathbf{x}^{(n)}; \alpha\theta_i^0 + (1 - \alpha)\theta_i^1, \theta_{-i}) - \lambda_n \sum_{j \neq i} \phi_{\lambda_n,j}^*(\theta_j) \right) \right] + C.$$

The log-likelihood is  $\kappa_2$ -concave, thus

$$\ell(\mathbf{x}^{(n)}; \alpha\theta_i^0 + (1 - \alpha)\theta_i^1, \theta_{-i}) \geq \alpha \ell(\mathbf{x}^{(n)}; \theta_i^0, \theta_{-i}) + (1 - \alpha) \ell(\mathbf{x}^{(n)}; \theta_i^1, \theta_{-i}) + \frac{\kappa_2 \alpha (1 - \alpha)}{2} (\theta_i^0 - \theta_i^1)^2.$$

By the Prékopa–Leindler inequality (Theorem 19.16, Villani (2009)), we have

$$\begin{aligned} &\mathbb{E}_{\pi_{-i}} \left[ \exp \left( \frac{\alpha \ell(\mathbf{x}^{(n)}; \theta_i^0, \theta_{-i}) + (1 - \alpha) \ell(\mathbf{x}^{(n)}; \theta_i^1, \theta_{-i})}{\lambda_n + 1} - \lambda_n \sum_{j \neq i} \phi_{\lambda_n,j}^*(\theta_j) \right) \right] \\ &\geq \mathbb{E}_{\pi_{-i}} \left[ \exp \left( \frac{\ell(\mathbf{x}^{(n)}; \theta_i^0, \theta_{-i})}{\lambda_n + 1} - \lambda_n \sum_{j \neq i} \phi_{\lambda_n,j}^*(\theta_j) \right) \right]^\alpha \mathbb{E}_{\pi_{-i}} \left[ \exp \left( \frac{\ell(\mathbf{x}^{(n)}; \theta_i^1, \theta_{-i})}{\lambda_n + 1} - \lambda_n \sum_{j \neq i} \phi_{\lambda_n,j}^*(\theta_j) \right) \right]^{1-\alpha}. \end{aligned}$$

Since the logarithmic function is concave, we conclude

$$\begin{aligned} &-\hat{\phi}_{\lambda_n,i}(\alpha\theta_i^0 + (1 - \alpha)\theta_i^1) \\ &\stackrel{\text{const}}{\geq} \log \mathbb{E}_{\pi_{-i}} \left[ \exp \left( \frac{\alpha \ell(\mathbf{x}^{(n)}; \theta_i^0, \theta_{-i}) + (1 - \alpha) \ell(\mathbf{x}^{(n)}; \theta_i^1, \theta_{-i})}{\lambda_n + 1} - \lambda_n \sum_{j \neq i} \phi_{\lambda_n,j}^*(\theta_j) \right) \right] + \frac{\kappa_2 \alpha (1 - \alpha)}{2(\lambda_n + 1)} (\theta_i^0 - \theta_i^1)^2 \\ &\geq -\alpha \phi_{\lambda_n,i}(\theta_i^0) - (1 - \alpha) \phi_{\lambda_n,i}(\theta_i^1) + \frac{\kappa_2 \alpha (1 - \alpha)}{2(\lambda_n + 1)} (\theta_i^0 - \theta_i^1)^2. \end{aligned}$$

Thus, the function  $-\hat{\phi}_{\lambda_n, i}(\cdot)$  is  $\kappa_2/(\lambda_n+1)$ -concave. By the fixed point representation (E.26),  $m_{\lambda_n}^*$  is  $(\kappa_2 + \kappa_1)$ -log-concave. Using the representation (E.27), we conclude that the distribution  $q_{\lambda_n}^*$  is  $(\kappa_2 + \kappa_1)$ -log-concave.  $\blacksquare$

We introduce some notations to streamline the two subsequent proofs.

**Definition 33 (Nonlinear quadratic tilt)** *Let  $\mu$  be a probability measure on  $\mathbb{R}$ . For  $(\phi, \gamma) \in L_1(\mathbb{R}) \in (0, \infty)$ , set*

$$c_\mu(\phi, \gamma) := \log \left[ \int \exp \left( -\phi(\theta) - \frac{\gamma}{2}\theta^2 \right) d\mu(\theta) \right], \quad (\text{E.29})$$

and define the probability distribution  $\mu_{\phi, \gamma}$  on  $\mathbb{R}$  by setting

$$\mu_{\phi, \gamma}(\theta) := \exp \left( -\phi(\theta) - \frac{\gamma}{2}\theta^2 - c_\mu(\phi, \gamma) \right) \mu(\theta), \quad \forall \theta \in \mathbb{R}. \quad (\text{E.30})$$

For any probability measure  $\mu$ , we have  $c_\mu(\phi, \gamma) < \infty$  for any  $(\phi, \gamma) \in L_1(\mathbb{R}) \in (0, \infty)$ . Given the base measure  $\mu$ , the tilted measure  $\mu_{\phi, \gamma}(\theta)$  has an exponential family density that has  $(\phi(\theta), \theta^2)$  as the sufficient statistics. We call  $\mu_{\phi, \gamma}$  a *nonlinear quadratic tilt* of  $\mu$ .

Using Theorem 25 and Proposition 31, we have

$$q_{\lambda_n}^*(\theta) = \frac{1}{\mathcal{Z}_D(\lambda_n)} \exp \left( -\frac{\theta^T \mathbf{B}_{\text{off}} \theta}{2(\lambda_n + 1)} + \mathbf{w}^T \theta + \sum_{i=1}^D c_{\pi_i} \left( \lambda_n \phi_{\lambda_n, i}^*, \frac{d_i}{\lambda_n + 1} \right) \right) \prod_{i=1}^D \pi_{i, \lambda_n \phi_{\lambda_n, i}^*, \frac{d_i}{\lambda_n + 1}}(\theta_i), \quad (\text{E.31})$$

Here  $c_{\pi_i} \left( \lambda_n \phi_{\lambda_n, i}^*, \frac{d_i}{\lambda_n + 1} \right)$  is defined in Eq. (E.29), and  $\pi_{i, \lambda_n \phi_{\lambda_n, i}^*, \frac{d_i}{\lambda_n + 1}}$  is the nonlinear quadratic tilt of  $\pi_i$  with parameters  $(\lambda_n \phi_{\lambda_n, i}^*, \frac{d_i}{\lambda_n + 1})$ . The constant  $\mathcal{Z}_D(\lambda_n)$  is defined as:

$$\mathcal{Z}_D(\lambda_n) := \int_{\Theta} \exp \left( -\frac{\theta^T \mathbf{B}_{\text{off}} \theta}{2(\lambda_n + 1)} + \mathbf{w}^T \theta \right) \prod_{i=1}^D \pi_{i, \lambda_n \phi_{\lambda_n, i}^*, \frac{d_i}{\lambda_n + 1}}(\theta_i) d\theta.$$

When  $\lambda_n = 0$ ,  $\mathcal{Z}_D(0)$  is the normalizing constant of the exact posterior. When  $\lambda_n > 0$ , we can view  $\mathcal{Z}_D(\lambda_n)$  as an approximation to  $\mathcal{Z}_D(0)$ .

The log-concavity of  $q_{\lambda_n}^*$  implies an upper bound of  $\Xi(q_{\lambda_n}^*)$  using the covariance matrix, the design matrix, and the regularization parameter.

**Lemma 34** *Let the assumptions of Theorem 4 hold. The solution  $q_{\lambda_n}^*$  to Eq. (4.7) satisfies*

$$\Xi(q_{\lambda_n}^*) \leq \frac{\text{tr} \left( \text{Cov}_{q_{\lambda_n}^*} (\mathbf{B}_{\text{off}} \theta) \right)}{4(\kappa_1 + \kappa_2)(\lambda_n + 1)^2}.$$

**Proof** [Proof of Lemma 34] Any constant shift in  $q_{\lambda_n}^*$  is preserved by its marginal distribution  $m_{\lambda_n}^*$ . Since the KL divergence is invariant to constant shift,  $\Xi(q_{\lambda_n}^*)$  is the same if we shift  $q_{\lambda_n}^*$  by a constant. WLOG, we can assume that  $\mathbb{E}_{q_{\lambda_n}^*} [\theta] = 0$ .

Let  $\tilde{m}_i(\theta_i) \propto \exp\left(-\frac{d_i}{2}\theta_i^2 + \mathbf{w}_i\theta_i\right)\pi(\theta_i)$ , and  $\tilde{m}(\theta) = \prod_{i=1}^D \tilde{m}_i(\theta_i)$ . By the variational representation of mutual information, we have

$$\Xi(\mathbf{q}_{\lambda_n}^*) = D_{\text{KL}}(\mathbf{q}_{\lambda_n}^* \parallel m_{\lambda_n}^*) \leq D_{\text{KL}}(\mathbf{q}_{\lambda_n}^* \parallel \tilde{m}).$$

By Lemma 32,  $\phi_{\lambda_n, i}^*$  is  $\kappa_2/(\lambda_n + 1)$ -convex. Since  $\pi$  is  $\kappa_1$ -log-concave,  $\tilde{m}$  is  $(\kappa_1 + \kappa_2)$ -log-concave. By the log-Sobolev inequality, we have:

$$\Xi(\mathbf{q}_{\lambda_n}^*) \leq D_{\text{KL}}(\mathbf{q}_{\lambda_n}^* \parallel \tilde{m}) \leq \frac{1}{\kappa_1 + \kappa_2} \int_{\Theta} \left\| \nabla_{\theta} \left( \frac{\theta^T \mathbf{B}_{\text{diag}} \theta - \theta^T \mathbf{B} \theta}{2(\lambda_n + 1)} \right) \right\|_2^2 \mathbf{q}_{\lambda_n}^*(\theta) d\theta = \frac{\mathbb{E}_{\mathbf{q}_{\lambda_n}^*} [\|\mathbf{B}_{\text{off}} \theta\|_2^2]}{4(\kappa_1 + \kappa_2)(\lambda_n + 1)^2}.$$

Under the assumed constraint  $\mathbb{E}_{\mathbf{q}_{\lambda_n}^*} [\theta] = 0$ , we conclude with the desired inequality:

$$\Xi(\mathbf{q}_{\lambda_n}^*) \leq \frac{\text{tr}(\text{Cov}_{\mathbf{q}_{\lambda_n}^*}(\mathbf{B}_{\text{off}} \theta))}{4(\kappa_1 + \kappa_2)(\lambda_n + 1)^2}.$$

■

### Proof [Proof of Theorem 4]

Define  $\tilde{\mathbf{B}}_{\text{off}} := \frac{\mathbf{B}_{\text{off}}}{\lambda_n + 1}$  and  $\tilde{\pi}_i := \pi_{i, \lambda_n \phi_{\lambda_n, i}^*, \frac{d_i}{\lambda_n + 1}}(\theta_i)$ .

We can write

$$\mathbf{q}_{\lambda_n}^*(\theta) = \frac{1}{\mathcal{Z}_D(\lambda_n)} \exp \left( -\frac{\theta^T \mathbf{B}_{\text{off}} \theta}{2(\lambda_n + 1)} + \mathbf{w}^T \theta + \sum_{i=1}^D c_{\pi_i} \left( \lambda_n \phi_{\lambda_n, i}^*, \frac{d_i}{\lambda_n + 1} \right) \right) \prod_{i=1}^D \tilde{\pi}_i(\theta_i).$$

By Lemma 32,  $\mathbf{q}_{\lambda_n}^*$  is a  $(\kappa_1 + \kappa_2)$ -log-concave. By Theorem 1 of Lacker et al. (2024), we have:

$$\log \mathcal{Z}_D(\lambda_n) - \sup_{m \in \mathbb{M}(\Theta)} \left[ -\frac{1}{2} \mathbb{E}_m[\theta]^T \tilde{\mathbf{B}}_{\text{off}} \mathbb{E}_m[\theta] + \mathbf{w}^T \mathbb{E}_m[\theta] - \sum_{i=1}^D D_{\text{KL}}(m_i \parallel \tilde{\pi}_i) \right] \leq \frac{\sum_{j=1}^D \sum_{i=1}^D [\mathbf{B}_{\text{off}}]_{ij}^2}{(\kappa_1 + \kappa_2)^2 (\lambda_n + 1)^2}, \quad (\text{E.32})$$

where  $\mathbb{E}_m[\theta]$  is the mean vector of  $m$ .

For any  $m \in \mathbb{M}(\Theta)$ , we have

$$\begin{aligned} D_{\text{KL}}(m \parallel \mathbf{q}_{\lambda_n}^*) &= \int_{\Theta} m(\theta) \left[ \log \mathcal{Z}_D(\lambda_n) + \frac{1}{2} \theta^T \tilde{\mathbf{B}}_{\text{off}} \theta - \mathbf{w}^T \theta + \log \frac{m(\theta)}{\prod_{i=1}^D \tilde{\pi}_i(\theta_i)} \right] d\theta. \\ &= \log \mathcal{Z}_D(\lambda_n) + \frac{1}{2} \mathbb{E}_m[\theta]^T \tilde{\mathbf{B}}_{\text{off}} \mathbb{E}_m[\theta] - \mathbf{w}^T \mathbb{E}_m[\theta] + \sum_{i=1}^D D_{\text{KL}}(m_i \parallel \tilde{\pi}_i). \end{aligned}$$

We invoke the upper bound on the log normalizer Eq. (E.32):

$$\inf_{m \in \mathbb{M}(\Theta)} D_{\text{KL}}(m \parallel \mathbf{q}_{\lambda_n}^*) \leq \frac{\sum_{j=1}^D \sum_{i=1}^D [\mathbf{B}_{\text{off}}]_{ij}^2}{(\kappa_1 + \kappa_2)^2 (\lambda_n + 1)^2}. \quad (\text{E.33})$$

By the  $T_2$ -transportation inequality (Theorem 1 and 2, Otto and Villani (2000)), we upper bound the Wasserstein metric with the square root of KL divergence:

$$\inf_{m \in \mathbb{M}(\Theta)} W_2(q_{\lambda_n}^*, m) \leq \sqrt{\frac{2}{\kappa_2 + \kappa_1} \inf_{m \in \mathbb{M}(\Theta)} D_{\text{KL}}(m \parallel q_{\lambda_n}^*)} \leq \sqrt{\frac{2 \sum_{j=1}^D \sum_{i=1}^D [\mathbf{B}_{\text{off}}]_{ij}^2}{(\kappa_1 + \kappa_2)^3 (\lambda_n + 1)^2}}. \quad (\text{E.34})$$

For  $\lambda_n \succ \sqrt{\text{tr}(\mathbf{B}_{\text{off}}^2)}$ , we have  $\inf_{m \in \mathbb{M}(\Theta)} W_2(q_{\lambda_n}^*, m) \xrightarrow{P_{\theta_0}} 0$ .

Consider the second regime  $\lambda_n \succ \sqrt{\text{tr}(\mathbf{B}_{\text{off}}^2)/D}$ . By the triangle inequality, we have

$$\begin{aligned} & \sup_{\mathbf{y} \in \mathbb{R}^n} \inf_{m \in \mathbb{M}(\Theta)} \mathbb{E}_{q_{\lambda_n}^*} \left[ \left( \frac{1}{D} \sum_{i=1}^D \psi(\theta_i) - \frac{1}{D} \sum_{i=1}^D \mathbb{E}_m[\psi(\theta_i)] \right)^2 \right]^{1/2} \\ & \leq \sup_{\mathbf{y} \in \mathbb{R}^n} \mathbb{E}_{q_{\lambda_n}^*} \left[ \left( \frac{1}{D} \sum_{i=1}^D \psi(\theta_i) - \frac{1}{D} \sum_{i=1}^D \mathbb{E}_{q_{\lambda_n}^*}[\psi(\theta_i)] \right)^2 \right]^{1/2} + \inf_{m \in \mathbb{M}(\Theta)} \left| \frac{1}{D} \sum_{i=1}^D \left( \mathbb{E}_{q_{\lambda_n}^*}[\psi(\theta_i)] - \mathbb{E}_m[\psi(\theta_i)] \right) \right|. \end{aligned} \quad (\text{E.35})$$

Since  $\psi$  is 1-Lipschitz, we apply Kantorovich duality to bound the second term.

$$\begin{aligned} & \left( \inf_{m \in \mathbb{M}(\Theta)} \left| \frac{1}{D} \sum_{i=1}^D \left( \mathbb{E}_{q_{\lambda_n}^*}[\psi(\theta_i)] - \mathbb{E}_m[\psi(\theta_i)] \right) \right| \right)^2 \\ & \leq \inf_{m_1, \dots, m_D} \frac{1}{D} \sum_{i=1}^D W_1^2(q_{\lambda_n, i}^*, m_i) \leq \inf_{m_1, \dots, m_D} \frac{1}{D} \sum_{i=1}^D W_2^2(q_{\lambda_n, i}^*, m_i). \end{aligned}$$

By the subadditivity inequality of Wasserstein distance and Eq. (E.34), we have

$$\inf_{m_1, \dots, m_D} \sum_{i=1}^D W_2^2(q_{\lambda_n, i}^*, m_i) \leq \inf_{m \in \mathbb{M}(\Theta)} W_2^2(q_{\lambda_n}^*, m) \leq \frac{2 \sum_{j=1}^D \sum_{i=1}^D [\mathbf{B}_{\text{off}}]_{ij}^2}{(\kappa_1 + \kappa_2)^3 (\lambda_n + 1)^2} \quad (\text{E.36})$$

Thus,

$$\inf_{m \in \mathbb{M}(\Theta)} \left| \frac{1}{D} \sum_{i=1}^D \left( \mathbb{E}_{q_{\lambda_n}^*}[\psi(\theta_i)] - \mathbb{E}_m[\psi(\theta_i)] \right) \right| \leq \sqrt{\frac{2 \sum_{j=1}^D \sum_{i=1}^D [\mathbf{B}_{\text{off}}]_{ij}^2}{D(\kappa_1 + \kappa_2)^3 (\lambda_n + 1)^2}}. \quad (\text{E.37})$$

The Lipschitzness implies  $\|\nabla \psi\|_2 \leq 1$ . To bound the first term, we apply Poincaré inequality to the function  $x \mapsto \frac{1}{D} \sum_{i=1}^D \psi(x_i)$ .

$$\begin{aligned} & \sup_{\mathbf{y} \in \mathbb{R}^n} \mathbb{E}_{q_{\lambda_n}^*} \left[ \left( \frac{1}{D} \sum_{i=1}^D \psi(\theta_i) - \frac{1}{D} \sum_{i=1}^D \mathbb{E}_{q_{\lambda_n}^*}[\psi(\theta_i)] \right)^2 \right] \leq \sup_{\mathbf{y} \in \mathbb{R}^n} \text{Var}_{q_{\lambda_n}^*} \left( \frac{1}{D} \sum_{i=1}^D \psi(\theta_i) \right) \\ & \leq \frac{1}{(\kappa_1 + \kappa_2) D^2} \sum_{i=1}^D \mathbb{E}_{q_{\lambda_n}^*} [\|\nabla \psi\|_2^2] \leq \frac{1}{D(\kappa_1 + \kappa_2)}. \end{aligned} \quad (\text{E.38})$$

Combining bounds Eq. (E.37) and Eq. (E.38), we have

$$\begin{aligned} & \sup_{\mathbf{y} \in \mathbb{R}^n} \inf_{m \in \mathbb{M}(\Theta)} \mathbb{E}_{\mathbf{q}_{\lambda_n}^*} \left[ \left( \frac{1}{D} \sum_{i=1}^D \psi(\theta_i) - \frac{1}{D} \sum_{i=1}^D \mathbb{E}_m[\psi(\theta_i)] \right)^2 \right] \\ & \leq \frac{(\kappa_1 + \kappa_2)^2 (\lambda_n + 1)^2 + 2 \sum_{j=1}^D \sum_{i=1}^D [\mathbf{B}_{\text{off}}]_{ij}^2}{D(\kappa_1 + \kappa_2)^3 (\lambda_n + 1)^2} \end{aligned} \quad (\text{E.39})$$

For  $\lambda_n \succ \sqrt{\text{tr}(\mathbf{B}_{\text{off}}^2)/D}$ , the bounds implies the Eq. (4.9).

Consider the third regime  $\lambda_n \prec (\kappa_1 + \kappa_2) [\text{tr}(\text{Cov}_{\mathbf{q}_0^*}(\mathbf{B}_{\text{off}}\theta))]^{-1}$ . Recall that  $\Xi$ -VI has a dual problem of the form  $\min_{\mathbf{q} \in \mathbb{M}(\Theta)} D_{\text{KL}}(\mathbf{q} \parallel \mathbf{q}_0^*)$  for some constant  $r(\lambda_n)$  depending on  $\lambda_n$ . If  $\Xi(\mathbf{q}_0^*) \leq r(\lambda_n)$ , then  $\mathbf{q}_{\lambda_n}^* = \mathbf{q}_0^*$ , hence  $\Xi(\mathbf{q}_0^*) \geq \Xi(\mathbf{q}_{\lambda_n}^*)$ . For  $t < \lambda_n$  and fixed  $n$ , we apply Lemma 34 to obtain an upper bound,

$$D_{\text{KL}}(\mathbf{q}_{\lambda_n}^* \parallel \mathbf{q}_0^*) - D_{\text{KL}}(\mathbf{q}_t^* \parallel \mathbf{q}_0^*) \leq \lambda_n (\Xi(\mathbf{q}_t^*) - \Xi(\mathbf{q}_{\lambda_n}^*)) \leq \lambda_n \Xi(\mathbf{q}_0^*) \xrightarrow{P_{\theta_0}} 0.$$

Finally, consider the fourth regime  $\lambda_n \prec D(\kappa_1 + \kappa_2) [\text{tr}(\text{Cov}_{\mathbf{q}_0^*}(\mathbf{B}_{\text{off}}\theta))]^{-1}$ . We follow an analogous derivation as the third regime:

$$D_{\text{KL}}(\mathbf{q}_{\lambda_n}^* \parallel \mathbf{q}_0^*) - D_{\text{KL}}(\mathbf{q}_t^* \parallel \mathbf{q}_0^*) \leq \lambda_n (\Xi(\mathbf{q}_t^*) - \Xi(\mathbf{q}_{\lambda_n}^*)) \leq \lambda_n \Xi(\mathbf{q}_0^*) = o(D).$$

Since  $\mathbf{q}_0^*$  is  $(\kappa_1 + \kappa_2)$ -log-concave, we invoke the  $T_2$ -transportation inequality:

$$W_2(\mathbf{q}_{\lambda_n}^*, \mathbf{q}_0^*) \leq \sqrt{\frac{2}{\kappa_2 + \kappa_1} D_{\text{KL}}(\mathbf{q}_{\lambda_n}^* \parallel \mathbf{q}_0^*)}.$$

The proofs for the third and fourth regimes are the same as the first two regimes, where we plug in the upper bounds for the KL divergence to upper-bound the Wasserstein distance. We skip repeating the details.  $\blacksquare$

**Proof** [Proof of Corollary 5] Given  $\text{tr}(\mathbf{B}_{\text{off}}^2) = \sum_{i=1}^D \eta_i^2$ , Theorem 4 ensures that the convergence of  $W_2(\mathbf{q}_{\lambda_n}^*, \mathbf{m}_{\lambda_n}^*)$  holds for  $\lambda_n \succ \sqrt{\sum_{i=1}^D \eta_i^2}$ . Since  $\sum_{i=1}^D \eta_i^2 \lesssim D^2$ , we have  $W_2(\mathbf{q}_{\lambda_n}^*, \mathbf{m}_{\lambda_n}^*)$  converges in probability to zero, for any choice of  $\lambda_n \succ D$ .  $\blacksquare$

### Proofs of Section A.3

We first state an auxiliary lemma to Theorem 13.

**Lemma 35** *Let  $\mathbf{q}_{\lambda}^*$  be the  $\Xi$ -variational posterior. Then  $\text{ELBO}(\mathbf{q}_{\lambda}^*)$  and  $\mathcal{C}_{\lambda}$  are monotonically decreasing function of  $\lambda$ .*

**Proof** [Proof of Lemma 35] Since  $\mathbf{q}_{\lambda}^*$  is a maximizer of  $\text{ELBO}(\mathbf{q}) - \lambda \Xi(\mathbf{q})$ , we have

$$\mathcal{C}_{\lambda} = \text{ELBO}(\mathbf{q}_{\lambda}^*) - \lambda \Xi(\mathbf{q}_{\lambda}^*).$$

For  $\lambda_1 < \lambda_2$ , we have

$$\mathcal{C}_{\lambda_1} = \text{ELBO}(q_{\lambda_1}^*) - \lambda_1 \Xi(q_{\lambda_1}^*) \geq \text{ELBO}(q_{\lambda_2}^*) - \lambda_1 \Xi(q_{\lambda_2}^*) \geq \text{ELBO}(q_{\lambda_2}^*) - \lambda_2 \Xi(q_{\lambda_2}^*) = \mathcal{C}_{\lambda_2}.$$

By Lagrangian duality, we have  $\text{ELBO}(q_{\lambda}^*) = \max_{\Xi(q) \leq t(\lambda)} \text{ELBO}(q)$  for  $t(\lambda)$  monotonically decreasing in  $\lambda$ .

For  $\lambda_1 < \lambda_2$ ,  $t(\lambda_1) \geq t(\lambda_2)$  hence

$$\text{ELBO}(q_{\lambda_1}^*) = \max_{\Xi(q) \leq t(\lambda_1)} \text{ELBO}(q) \geq \max_{\Xi(q) \leq t(\lambda_2)} \text{ELBO}(q) = \text{ELBO}(q_{\lambda_2}^*).$$

■

**Proof** [Proof of Theorem 13] Let  $(P_2(\Theta), W_2)$  be the metric space. We want to show that the functionals

$$F_{\lambda}(q) := D_{\text{KL}}(q \parallel q_0^*) + \lambda \Xi(q).$$

$\Gamma$ -converge to

$$F_{\infty}(q) := D_{\text{KL}}(q \parallel q_0^*) + \infty \Xi(q),$$

as  $\lambda \rightarrow \infty$ .

To verify  $\Gamma$  convergence, we make use of the property that the KL divergence functional  $D_{\text{KL}}(\cdot \parallel q_0^*)$  and  $\Xi(\cdot)$  functional are lower semicontinuous (l.s.c.) in Wasserstein metric. This is provided in Lemma 21.

Let  $q \in \mathbb{P}_2(\Theta)$  and  $W_2(q_{\lambda}, q) \rightarrow 0$ . If  $q$  is a product measure, then

$$F_{\infty}(q) = D_{\text{KL}}(q \parallel q_0^*) \leq \liminf_{\lambda \rightarrow \infty} D_{\text{KL}}(q_{\lambda} \parallel q_0^*) \leq \liminf_{\lambda \rightarrow \infty} F_{\lambda}(q_{\lambda}).$$

The first inequality holds because  $D_{\text{KL}}(\cdot \parallel q_0^*)$  is l.s.c.

If  $q$  is not a product measure, we have  $\liminf_{n \rightarrow \infty} \Xi(q_n) \geq \Xi(q) > 0$  by the lower semicontinuity of  $\Xi$ . Since the KL term is nonnegative, we have

$$F_{\infty}(q) = \infty = \liminf_{\lambda \rightarrow \infty} F_{\lambda}(q_{\lambda}).$$

Thus the liminf inequality is verified.

Next we show the existence of a recovery sequence. For any  $q \in \mathbb{P}_2(\Theta)$ , we take  $q_{\lambda} = q$ . If  $q$  is a product measure, then

$$F_{\infty}(q) = D_{\text{KL}}(q \parallel q_0^*) \geq D_{\text{KL}}(q \parallel q_0^*).$$

Otherwise,

$$F_{\infty}(q) = \infty \geq \limsup_{\lambda \rightarrow \infty} F_{\lambda}(q_{\lambda}).$$

This verifies the limsup inequality. Combining the liminf and limsup inequalities, we obtain that  $F_{\infty} = \Gamma - \lim_{\lambda \rightarrow \infty} F_{\lambda}$ .

Next we prove that the sequence  $F_{\lambda}$  is eqi-coercive. Take  $\lambda_j \rightarrow \infty$  and  $q_{\lambda_j}$  such that  $F_{\lambda_j}(q_{\lambda_j}) \leq t$  for all  $j$ . Then  $\Xi(q_{\lambda_j}) = o(1)$  because  $\lambda_j \Xi(q_{\lambda_j})$  is bounded as  $\lambda_j \rightarrow \infty$ . Moreover,  $D_{\text{KL}}(q_{\lambda_j} \parallel q_0^*)$  is upper bounded by  $t$ . Since  $D_{\text{KL}}(\cdot \parallel q_0^*)$  is Wasserstein

(geodesically) convex, it is coercive by Lemma 2.4.8 of Ambrosio et al. (2005). Thus, there exists a converging sequence  $q'_{\lambda_j}$  such that  $D_{\text{KL}}(q'_{\lambda_j} \parallel q_0^*) \leq D_{\text{KL}}(q_{\lambda_j} \parallel q_0^*) + o(1)$ . Since  $\Xi(q_{\lambda_j}) = o(1)$ , we obtain that  $F_{\lambda_j}(q'_j) \leq F_{\lambda_j}(q_j) + o(1)$ . This verifies the equi-coercivity of  $F_{\lambda}$ .

Finally, by the fundamental theorem of  $\Gamma$  convergence (Theorem 29), we conclude that

$$W_2(q_\infty^*, q_\lambda^*) \rightarrow 0, \quad \text{as } \lambda \rightarrow \infty,$$

and

$$|\mathcal{C}_\lambda - \mathcal{C}_\infty| \rightarrow 0, \quad \text{as } \lambda \rightarrow \infty.$$

By Corollary 2.1 of Braides (2014), every minimizer of  $F_\infty$  is the limit of some converging minimizing sequences of  $F_\lambda$ . For any  $q_\infty^* \in \mathbb{Q}_\infty$ , this implies the existence of a sequence  $q_\lambda^* \in \mathbb{Q}_\lambda$  such that

$$W_2(q_\infty^*, q_\lambda^*) \rightarrow 0, \quad \text{as } \lambda \rightarrow \infty.$$

■

**Proof** [Proof of Theorem 14] We define  $\mathbb{P}'_2(\Theta)$  as  $P'_2(\Theta) = \{q \in \mathbb{P}_2(\Theta) : \Xi(q) < \infty\}$ . The space  $(\mathbb{P}'_2(\Theta), W_2)$  is a metric space. We want to show that the sequence of functionals

$$F_\lambda(q) := D_{\text{KL}}(q \parallel q_0^*) + \lambda \Xi(q).$$

$\Gamma$ -converge to

$$F_0(q) := D_{\text{KL}}(q \parallel q_0^*),$$

as  $\lambda \rightarrow 0$ . Both  $F_\lambda(q)$  and  $F_0(q)$  are defined on  $(\mathbb{P}'_2(\Theta), W_2)$ .

We make use of Lemma 21 which shows that the KL divergence functional  $D_{\text{KL}}(\cdot \parallel q_0^*)$  and  $\Xi(\cdot)$  functional are lower semicontinuous (l.s.c.) in Wasserstein metric.

Let  $q \in \mathbb{P}'_2(\Theta)$  and  $W_2(q_\lambda, q) \rightarrow 0$ . We have

$$F_0(q) = D_{\text{KL}}(q \parallel q_0^*) \leq \liminf_{\lambda \rightarrow 0} D_{\text{KL}}(q_\lambda \parallel q_0^*) \leq \liminf_{\lambda \rightarrow 0} F_\lambda(q_\lambda).$$

The first inequality holds because  $D_{\text{KL}}(\cdot \parallel q_0^*)$  is l.s.c. The second inequality holds because  $\Xi(\cdot)$  is nonnegative.

Next we show that the existence of a recovery sequence. For any  $q \in \mathbb{P}_2(\Theta)$ , we take  $q_\lambda = q$ . Since  $\Xi(q) < \infty$ , we have

$$F_0(q) = D_{\text{KL}}(q \parallel q_0^*) \geq \limsup_{\lambda \rightarrow 0} D_{\text{KL}}(q \parallel q_0^*) + \lambda \Xi(q).$$

This verifies the limsup inequality. Combining the liminf and limsup inequalities, we obtain that  $F = \Gamma - \lim_{\lambda \rightarrow \infty} F_\lambda$ .

We proceed to establish equi-coercivity of the sequence  $F_\lambda$ . Consider a sequence  $\lambda_j \rightarrow 0$  and  $q_{\lambda_j} \in \mathbb{P}'_2(\Theta)$  for which  $F_{\lambda_j}(q_{\lambda_j}) \leq t$  holds for all  $j$ . Given that  $\Xi(q_{\lambda_j}) \geq 0$ , it follows that  $D_{\text{KL}}(q \parallel q_0^*) \leq t$ . Owing to the geodesic convexity of the Kullback-Leibler divergence  $D_{\text{KL}}(\cdot \parallel q_0^*)$  in the Wasserstein space, Lemma 2.4.8 from Ambrosio et al. (2005) ensures that it is coercive, implying that the set  $\{q \in \mathbb{P}'_2(\Theta) \mid D_{\text{KL}}(q \parallel q_0^*) \leq t\}$  is compact in the

metric space  $(\mathbb{P}'_2(\Theta), W_2)$ . Sequential compactness guarantees the existence of a convergent subsequence of  $q_{\lambda_j}$ , which converges to some  $q_0$  in  $\mathbb{P}'_2(\Theta)$ . Since  $D_{\text{KL}}(\cdot \parallel q_0^*)$  is lower semicontinuous (l.s.c.), we conclude that:

$$\begin{aligned} F_{\lambda_j}(q_0) &= D_{\text{KL}}(q_0 \parallel q_0^*) + \lambda_j \Xi(q_0) \leq D_{\text{KL}}(q_{\lambda_j} \parallel q_0^*) + \lambda_j \Xi(q_0) \\ &\leq F_{\lambda_j}(q_{\lambda_j}) + \lambda_j \Xi(q_0) = F_{\lambda_j}(q_{\lambda_j}) + o(1), \end{aligned}$$

Finally, by the fundamental theorem of  $\Gamma$  convergence, we conclude that

$$W_2(q_0^*, q_{\lambda}^*) \rightarrow 0, \quad \text{as } \lambda \rightarrow 0,$$

where  $q_0$  is a minimizer of  $F_0$ , and

$$|\mathcal{C}_{\lambda} - \mathcal{C}_0| \rightarrow 0, \quad \text{as } \lambda \rightarrow 0.$$

Since  $q_0^*$  is the unique minimizer of  $F_0$ , we conclude that  $q_0 = q_0^*$ .

To prove the convergence of optimal cost, we note that

$$F_{\lambda}(q_{\lambda}^*) \leq F_{\lambda}(q_0^*) = F_0(q_0^*) + \lambda \Xi(q_0^*).$$

Thus,

$$|\mathcal{C}_{\lambda} - \mathcal{C}_0| = |F_{\lambda}(q_{\lambda}^*) - F_0(q_0^*)| \leq \lambda \Xi(q_0^*).$$

■

Define a functional  $\Phi_{\lambda}$  that combines the objective functional of the inner variational objective problem and Assumption 5:

$$\Phi_{\lambda}(q) := \mathbb{E}_q \left[ -\ell(\mathbf{x}^{(n)}; \theta) + \sum_{i=1}^D \phi_i(\theta_i) \right] + (\lambda + 1) \Xi(q), \quad (\text{E.40})$$

where  $\phi_i : \Theta_i \mapsto \mathbb{R}$  are the one-dimensional function in the Lipschitz cost assumption (Assumption 5). Since  $\phi_i$  are tensorized, minimizing  $\Phi_{\lambda}$  over  $\mathcal{C}(m)$  is equivalent to solving the inner variational problem over  $\mathcal{C}(m)$ .

For proving Theorem 15, we introduce a Pythagorean theorem for the inner variational problem.

**Lemma 36** *Let  $q_{\lambda} \in \mathcal{C}(m)$  be a optimizer of  $\Phi_{\lambda}$  over  $\mathcal{C}(m)$ . Then*

$$D_{\text{KL}}(q, q_{\lambda}) \leq \Phi_{\lambda}(q) - \Phi_{\lambda}(q_{\lambda}), \quad \text{for all } q \in \mathcal{C}(m).$$

**Proof** [Proof of Lemma 36] We recall definition of the auxiliary measure  $q_{\text{aux}}$  in the proof of Theorem 25,  $q_{\text{aux}}(\theta) = \alpha^{-1} \exp \left( \frac{\ell(\mathbf{x}^{(n)}; \theta) - \sum_{i=1}^D \phi_i(\theta_i)}{\lambda + 1} \right) m(\theta)$ , where  $\alpha$  is the normalizing constant. Then

$$\Phi_{\lambda}(q) = D_{\text{KL}}(q \parallel q_{\text{aux}}) - \log \alpha, \quad (\text{E.41})$$

so that the entropic optimal transport problem is equivalent to minimizing  $D_{\text{KL}}(\cdot \parallel q_{\text{aux}})$ . In particular,  $q_\lambda = \arg \min_{\mathcal{C}(m)} D_{\text{KL}}(q \parallel q_{\text{aux}})$  and the Pythagorean theorem for relative entropy (Theorem 2.2, (Csiszár, 1975)) yields

$$D_{\text{KL}}(q \parallel q_{\text{aux}}) \geq D_{\text{KL}}(q_\lambda \parallel q_{\text{aux}}) + D_{\text{KL}}(q \parallel q_\lambda) \quad \text{for all } q \in \mathcal{C}(m).$$

In view of Eq. (E.41), the desired claim holds.  $\blacksquare$

The next Lemma is also auxiliary to the proof of Theorem 15.

**Lemma 37** *Let  $q_\lambda^* \in \mathcal{C}(m^*)$  be a optimizer of  $\Phi_\lambda$  over  $\mathcal{C}(m^*)$ , and  $q_\lambda^s \in \mathcal{C}(\tilde{m})$  be its shadow. Then*

$$|\Phi_\lambda(q_\lambda^*) - \Phi_\lambda(q_\lambda^s)| \leq LW_2(q_\lambda^*, q_\lambda^s).$$

**Proof** [Proof of Lemma 37] Using the Lipschitz cost assumption and Lemma 23, we have

$$\begin{aligned} \Phi_\lambda(q_\lambda^*) &= \mathbb{E}_{q_\lambda^*} \left[ -\ell(\mathbf{x}^{(n)}; \theta) + \sum_{i=1}^D \phi_i(\theta_i) \right] + (\lambda + 1)\Xi(q_\lambda^*). \\ &\geq \mathbb{E}_{q_\lambda^s} \left[ -\ell(\mathbf{x}^{(n)}; \theta) + \sum_{i=1}^D \phi_i(\theta_i) \right] - LW_2(q_\lambda^*, q_\lambda^s) + (\lambda + 1)\Xi(q_\lambda^s) \\ &= \Phi_\lambda(q_\lambda^s) - LW_2(q_\lambda^*, q_\lambda^s). \end{aligned}$$

The claim follows by a symmetric argument.  $\blacksquare$

**Proof** [Proof of Theorem 15] Consider the optimizers  $\tilde{q}_\lambda \in \mathcal{C}(\tilde{m})$  and  $q_\lambda^* \in \mathcal{C}(m_\lambda^*)$ . Let  $q_\lambda^s \in \mathcal{C}(\tilde{m})$  be the shadow of  $q_\lambda^*$ . By Lemma 23 and the Lipschitz cost assumption, we have:

$$\begin{aligned} \Phi_\lambda(q_\lambda^s) - \Phi_\lambda(q_\lambda^*) &\leq \int_{\Theta} \left( \ell(\mathbf{x}^{(n)}; \theta) - \sum_{i=1}^D \phi_i(\theta_i) \right) (q_\lambda^s(\theta) - q_\lambda^*(\theta)) d\theta \\ &\leq LW_2(q_\lambda^s, q_\lambda^*) \leq LW_2(m_\lambda^*, \tilde{m}). \end{aligned}$$

Lemma 37 implies  $\Phi_\lambda(\tilde{q}_\lambda) - \Phi_\lambda(q_\lambda^*) \leq LW_2(m_\lambda^*, \tilde{m})$ . Adding the inequalities shows:

$$|\Phi_\lambda(\tilde{q}_\lambda) - \Phi_\lambda(q_\lambda^s)| \leq 2LW_2(m_\lambda^*, \tilde{m}).$$

By Lemma 36, we have that  $D_{\text{KL}}(\tilde{\pi}, \pi^*) \leq 2LW_2(m_\lambda^*, \tilde{m})$ , and the transport inequality assumption implies:

$$W_\rho(q_\lambda^s, \tilde{q}_\lambda) \leq C_\rho(2LW_2(m_\lambda^*, \tilde{m}))^{\frac{1}{2\rho}}.$$

By Lemma 23, we get  $W_2(q_\lambda^*, q_\lambda^s) = W_2(m_\lambda^*, \tilde{m})$ . We conclude the proof via the triangle inequality,

$$W_2(q_\lambda^*, \tilde{q}_\lambda) \leq W_2(q_\lambda^*, q_\lambda^s) + W_2(q_\lambda^s, \tilde{q}_\lambda) \leq W_2(m_\lambda^*, \tilde{m}) + C_q(2LW_2(m_\lambda^*, \tilde{m}))^{\frac{1}{2q}}.$$

### Details of Section 3.1

**Proof** [Proof of Proposition 1] Let the true precision matrix be  $\Lambda_0$ . Let  $f(\mu, \Sigma)$  be the objective function Eq. (3.1) parameterized by the variational mean  $\mu$  and covariance  $\Sigma$ . A direct calculation shows

$$f(\mu, \Sigma) = \frac{1}{2} \left[ (\mu_0 - \mu)^T \Lambda_0 (\mu_0 - \mu) + \lambda \sum_{K=1}^D \log \Sigma_{KK} + \text{tr} \{ \Lambda_0 \Sigma \} - (\lambda + 1) \log |\Sigma| \right].$$

First, we confirm that the optimal  $\mu^*$  is equal to the true  $\mu_0$ . The first-order optimality condition of  $f$  with respect to  $\mu$  yields

$$\partial_\mu f(\mu, \Sigma) = 0 \implies \Lambda_0(\mu - \mu_0) = 0.$$

Since  $\Sigma_0$  is full-rank, its inverse  $\Lambda_0$  is full-rank hence the equality above yields  $\mu^* = \mu_0$ .

Now we turn to  $\Sigma^*$ . The first-order optimality yields the following characterization of the optimal precision matrix  $\Lambda^*$ :

$$\partial_\Sigma f(\mu, \Sigma) = 0 \implies \lambda(\Sigma_{\text{diag}}^*)^{-1} + \Lambda_0 - (\lambda + 1)\Lambda = 0 \implies \Lambda^* = \frac{1}{\lambda + 1}\Lambda_0 + \frac{\lambda}{\lambda + 1}(\Sigma_{\text{diag}}^*)^{-1}, \quad (\text{E.42})$$

where  $\Sigma_{\text{diag}}^*$ ,  $\Sigma_{\text{off}}^*$  denote the diagonal and off-diagonal minor of  $\Sigma^*$ , respectively. By Eq. (E.42), we have

$$\Lambda_{\text{diag}}^* = \frac{1}{\lambda + 1}\Lambda_{0,\text{diag}} + \frac{\lambda}{\lambda + 1}(\Sigma_{\text{diag}}^*)^{-1} \implies (\lambda + 1)\Lambda_{\text{diag}}^* - \lambda(\Sigma_{\text{diag}}^*)^{-1} = \Lambda_{0,\text{diag}}$$

By Lemma 18, we have the inequality

$$\Sigma_{\text{diag}}^{*-1} \leq \Lambda_{\text{diag}}^* \leq \Lambda_{0,\text{diag}}. \quad (\text{E.43})$$

By the Woodbury identity (Eq.(156) in Petersen et al. (2008)) with  $C = I_D$ , we have

$$\Sigma^* = \left[ \frac{1}{\lambda + 1}\Lambda_0 + \frac{\lambda}{\lambda + 1}\Sigma_{\text{diag}}^{*-1} \right]^{-1} = \frac{\lambda + 1}{\lambda}\Sigma_{\text{diag}}^* - \left[ \frac{\lambda}{\lambda + 1}\Sigma_{\text{diag}}^{*-1} + \frac{\lambda^2}{\lambda + 1}\Sigma_{\text{diag}}^{*-1}\Sigma_0\Sigma_{\text{diag}}^{*-1} \right]^{-1}.$$

Taking the diagonal elements on both sides, we have

$$\begin{aligned} \frac{1}{\lambda}\Sigma_{\text{diag}}^* &= \left( \left[ \frac{\lambda}{\lambda + 1}\Sigma_{\text{diag}}^{*-1} + \frac{\lambda^2}{\lambda + 1}\Sigma_{\text{diag}}^{*-1}\Sigma_0\Sigma_{\text{diag}}^{*-1} \right]^{-1} \right)_{\text{diag}} \\ &= \left( \left[ \underbrace{\frac{\lambda}{\lambda + 1}\Sigma_{\text{diag}}^{*-1} + \frac{\lambda^2}{\lambda + 1}\Sigma_{\text{diag}}^{*-1}\Sigma_0\Sigma_{\text{diag}}^{*-1}}_A + \underbrace{\frac{\lambda^2}{\lambda + 1}\Sigma_{\text{diag}}^{*-1}\Sigma_{0,\text{off}}\Sigma_{\text{diag}}^{*-1}}_B \right]^{-1} \right)_{\text{diag}}. \end{aligned}$$

Note that  $B$  is a matrix with zero diagonal entries. By Lemma 18, we have

$$\frac{1}{\lambda}\Sigma_{\text{diag}}^* = ([A + B]^{-1})_{\text{diag}} \geq [A + B]_{\text{diag}}^{-1} = A^{-1}.$$

This implies that

$$\frac{1}{\lambda} \Sigma_{\text{diag}}^* \geq \Sigma_{\text{diag}}^* \left[ \frac{\lambda}{\lambda+1} \Sigma_{\text{diag}}^* + \frac{\lambda^2}{\lambda+1} \Sigma_{0,\text{diag}} \right]^{-1} \Sigma_{\text{diag}}^*,$$

which after simplification yields

$$\Sigma_{\text{diag}}^* \leq \Sigma_{0,\text{diag}}. \quad (\text{E.44})$$

By Hua's identity, we have

$$\Sigma^* = \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{\lambda}{\lambda+1} \Sigma_{\text{diag}}^{*-1} \right]^{-1} = (\lambda+1) \Sigma_0 - \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{1}{\lambda(\lambda+1)} \Lambda_0 \Sigma_{\text{diag}}^* \Lambda_0 \right]^{-1}.$$

It follows that

$$\Sigma^* - \Sigma_0 = \lambda \Sigma_0 - \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{1}{\lambda(\lambda+1)} \Lambda_0 \Sigma_{\text{diag}}^* \Lambda_0 \right]^{-1}.$$

By Eq. (E.44), the matrix  $\Sigma^* - \Sigma_0$  is negative semidefinite. By Eq. (E.43), we have  $\Sigma_{\text{diag}}^* \geq \Lambda_{0,\text{diag}}^{-1}$ . Then

$$\|\Sigma^* - \Sigma_0\| \leq \left\| \lambda \Sigma_0 - \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{1}{\lambda(\lambda+1)} \Lambda_0 \Lambda_{0,\text{diag}}^{-1} \Lambda_0 \right]^{-1} \right\|.$$

Since  $\Sigma_{\text{diag}}^* \leq \Sigma_{0,\text{diag}}$ , we obtain a lower bound with analogous techniques.

$$\|\Sigma^* - \Sigma_0\| \geq \left\| \lambda \Sigma_0 - \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{1}{\lambda(\lambda+1)} \Lambda_0 \Sigma_{0,\text{diag}} \Lambda_0 \right]^{-1} \right\|.$$

This lower bound holds when the matrix on the right hand side is negative semidefinite. To see that, we have

$$\lambda \Sigma_0 - \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{1}{\lambda(\lambda+1)} \Lambda_0 \Sigma_{0,\text{diag}} \Lambda_0 \right]^{-1} = -\Sigma_0 + \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{\lambda}{\lambda+1} \Sigma_{0,\text{diag}}^{-1} \right]^{-1},$$

Since  $\Lambda_{0,\text{diag}} \geq \Sigma_{0,\text{diag}}^{-1}$ , we have

$$\left( -\Sigma_0 + \left[ \frac{1}{\lambda+1} \Lambda_0 + \frac{\lambda}{\lambda+1} \Sigma_{0,\text{diag}}^{-1} \right]^{-1} \right)_{\text{diag}} \leq -\Sigma_{0,\text{diag}} + \Sigma_{0,\text{diag}} = 0.$$

This completes the proof. ■

Next, we provide the explicit formula for the  $\Xi$ -VI solution when the exact posterior is bivariate Gaussian.

**Proposition 38** *Let the exact posterior  $q_0^*$  be a bivariate Gaussian distribution with mean  $\mu_0$  and precision matrix  $\Lambda_0 = \begin{pmatrix} a_0 & b_0 \\ b_0 & c_0 \end{pmatrix}$ . Then the  $\Xi$ -variational solution in Eq. (3.1) is*

a bivariate Gaussian distribution with mean  $\mu_0$  and the following precision and covariance matrices:

$$\Lambda^* = \begin{pmatrix} \frac{a_0}{2} + \sqrt{\frac{a_0^2}{4} - \frac{\lambda}{(\lambda+1)^2} \frac{a_0 b_0^2}{c_0}} & \frac{1}{\lambda+1} b_0 \\ \frac{1}{\lambda+1} b_0 & \frac{c_0}{2} + \sqrt{\frac{c_0^2}{4} - \frac{\lambda}{(\lambda+1)^2} \frac{c_0 b_0^2}{a_0}} \end{pmatrix},$$

$$\Sigma^* = \frac{1}{a_0 c_0 - b_0^2} \begin{pmatrix} c_0 & -\frac{b_0}{\psi(\lambda)} \\ -\frac{b_0}{\psi(\lambda)} & a_0 \end{pmatrix},$$

$$\text{where } \psi(\lambda) = \frac{1}{2} \left( \lambda + 1 + \sqrt{\left( \lambda - \frac{2b_0^2 - a_0 c_0}{a_0 c_0} \right)^2 + \frac{4b_0^2}{a_0^2 c_0^2} (a_0 c_0 - b_0^2)} \right).$$

Compared to the exact covariance, the variational covariance matrix is adjusted by a factor depending on the regularizer  $\lambda$ . The adjusting function  $\psi : [0, \infty) \mapsto [0, \infty)$  is strictly increasing. Thus, as  $\lambda$  increases, we have element-wise strictly decreasing convergence to the mean-field covariance, i.e.  $\lim_{\lambda \rightarrow \infty} \Sigma^* = \begin{pmatrix} a_0^{-1} & 0 \\ 0 & c_0^{-1} \end{pmatrix}$ .

**Proof** [Proof of Proposition 38] Denote  $\Lambda^* = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ . The inverse is

$$\Sigma^* = \frac{1}{ac - b^2} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix}.$$

As shown in Section 3.1, we have

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} = \frac{1}{\lambda + 1} \begin{pmatrix} a_0 & b_0 \\ b_0 & c_0 \end{pmatrix} + \frac{\lambda}{\lambda + 1} \begin{pmatrix} a - \frac{b^2}{c} & 0 \\ 0 & d - \frac{b^2}{c} \end{pmatrix}.$$

This implies

$$\begin{pmatrix} a + \lambda \frac{b^2}{c} & (\lambda + 1)b \\ (\lambda + 1)b & c + \lambda \frac{b^2}{a} \end{pmatrix} = \begin{pmatrix} a_0 & b_0 \\ b_0 & c_0 \end{pmatrix}.$$

This translates to a system of equations

$$\begin{aligned} (\lambda + 1)b &= b_0 \\ a + \lambda \frac{b^2}{c} &= a_0 \\ c + \lambda \frac{b^2}{a} &= c_0. \end{aligned}$$

The first Equation yields  $b = \frac{b_0}{\lambda+1}$ . The other two equations yield

$$\frac{a_0 c}{c_0 a} = \frac{ac + b^2}{ac + b^2} = 1 \implies c = \frac{c_0}{a_0} a.$$

Substituting  $c$  gives us

$$a + \lambda \frac{a_0 b^2}{c_0 a} = a_0.$$

which yields  $a = \frac{a_0}{2} \pm \sqrt{\frac{a_0^2}{4} - \frac{\lambda}{(\lambda+1)^2} \frac{a_0 b_0^2}{c_0}}$ .

Similarly, substituting  $a$  with  $c$  yields  $c = \frac{c_0}{2} \pm \sqrt{\frac{c_0^2}{4} - \frac{\lambda}{(\lambda+1)^2} \frac{c_0 b_0^2}{a_0}}$ .

Finally, use the fact that  $\Lambda^* = \Lambda_0$  when  $\lambda = 0$  to obtain the solution set

$$a = \frac{a_0}{2} + \sqrt{\frac{a_0^2}{4} - \frac{\lambda}{(\lambda+1)^2} \frac{a_0 b_0^2}{c_0}}, \quad b = \frac{b_0}{\lambda+1}, \quad c = \frac{c_0}{2} + \sqrt{\frac{c_0^2}{4} - \frac{\lambda}{(\lambda+1)^2} \frac{c_0 b_0^2}{a_0}}.$$

To obtain the covariance matrix, note that

$$|\Lambda^*| = ac - b^2 = \frac{a_0 c_0}{2} \left( 1 + \sqrt{1 - \frac{\lambda}{(\lambda+1)^2} \frac{4b_0^2}{a_0 c_0}} \right) - \frac{1}{\lambda+1} b_0^2.$$

By the matrix inversion formula,

$$\begin{aligned} \Sigma^* &= \frac{1}{|\Lambda^*|} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix} = \begin{pmatrix} \frac{c_0}{2} + \sqrt{\frac{c_0^2}{4} - \frac{\lambda}{(\lambda+1)^2} \frac{c_0 b_0^2}{a_0}} & b_0 \\ \frac{a_0 c_0}{2} \left( 1 + \sqrt{1 - \frac{\lambda}{(\lambda+1)^2} \frac{4b_0^2}{a_0 c_0}} \right) - \frac{1}{\lambda+1} b_0^2 & -(\lambda+1) \left( \frac{a_0 c_0}{2} \left( 1 + \sqrt{1 - \frac{\lambda}{(\lambda+1)^2} \frac{4b_0^2}{a_0 c_0}} \right) - \frac{1}{\lambda+1} b_0^2 \right) \\ -b_0 & \frac{a_0}{2} + \sqrt{\frac{a_0^2}{4} - \frac{\lambda}{(\lambda+1)^2} \frac{a_0 b_0^2}{c_0}} \\ -(\lambda+1) \left( \frac{a_0 c_0}{2} \left( 1 + \sqrt{1 - \frac{\lambda}{(\lambda+1)^2} \frac{4b_0^2}{a_0 c_0}} \right) - \frac{1}{\lambda+1} b_0^2 \right) & \frac{a_0 c_0}{2} \left( 1 + \sqrt{1 - \frac{\lambda}{(\lambda+1)^2} \frac{4b_0^2}{a_0 c_0}} \right) - \frac{1}{\lambda+1} b_0^2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{c_0}{2} & b_0 \\ \frac{a_0 c_0}{2} - \frac{b_0^2}{\lambda+1 + \sqrt{(\lambda+1)^2 - \lambda \frac{4b_0^2}{a_0 c_0}}} & -\frac{a_0 c_0}{2} \left( \lambda+1 + \sqrt{(\lambda+1)^2 - \lambda \frac{4b_0^2}{a_0 c_0}} \right) - b_0^2 \\ -\frac{a_0 c_0}{2} \left( \lambda+1 + \sqrt{(\lambda+1)^2 - \lambda \frac{4b_0^2}{a_0 c_0}} \right) - b_0^2 & \frac{a_0}{2} - \frac{b_0^2}{\lambda+1 + \sqrt{(\lambda+1)^2 - \lambda \frac{4b_0^2}{a_0 c_0}}} \end{pmatrix} \\ &= \frac{1}{a_0 c_0 - \psi^{-1}(\lambda) b_0^2} \begin{pmatrix} c_0 & -\frac{b_0}{\psi(\lambda)} \\ -\frac{b_0}{\psi(\lambda)} & a_0 \end{pmatrix}. \end{aligned}$$

where  $\psi(\lambda) = \frac{1}{2} \left( \lambda + 1 + \sqrt{\left( \lambda - \frac{2b_0^2 - a_0 c_0}{a_0 c_0} \right)^2 + \frac{4b_0^2}{a_0^2 c_0^2} (a_0 c_0 - b_0^2)} \right)$ . ■

**Proof** [Proof of Corollary 16] We use the well known property of the 2-Wasserstein metric that for any  $q_0, q_1 \in \mathbb{P}_2(\Theta)$ ,

$$W_2^2 \left( \prod_{i=1}^D q_{0,i}, \prod_{i=1}^D q_{1,i} \right) \leq \sum_{i=1}^D W_2^2(q_{0,i}, q_{1,i}) \leq W_2^2(q_0, q_1). \quad (\text{E.45})$$

By Theorem 13,  $W_2(q_\lambda^*, q_\infty^*) \rightarrow 0$  as  $\lambda \rightarrow \infty$ , hence  $W_2(m_\lambda^*, q_\infty^*) \rightarrow 0$  by the identity Eq. (E.45) which implies  $W_2(q_\lambda^{*(\infty)}, q_\lambda^*) \rightarrow 0$  by Theorem 15. An analogous derivation holds for  $\lambda \rightarrow 0$ . ■

## Appendix F. Additional Simulation Results

$N$	$\frac{\ \mathbb{E}_{\hat{q}_{10,N}}(\theta) - \mathbb{E}_{q_0^*}(\theta)\ _2}{\ \mathbb{E}_{q_0^*}(\theta)\ _2}$	$\frac{\ \text{Cov}_{\hat{q}_{10,N}}(\theta) - \text{Cov}_{q_0^*}(\theta)\ _F}{\ \text{Cov}_{q_0^*}(\theta)\ _F}$	$W_2(\hat{q}_{10,N}, q_0^*)$	Runtime (sec)
5	0.069 [0.059, 0.078]	0.813 [0.767, 0.884]	0.321 [0.307, 0.336]	0.017 [0.017, 0.017]
10	0.049 [0.041, 0.052]	0.672 [0.633, 0.724]	0.243 [0.231, 0.263]	0.222 [0.220, 0.225]
15	0.040 [0.035, 0.044]	0.639 [0.611, 0.664]	0.221 [0.215, 0.231]	1.076 [1.074, 1.081]
20	0.031 [0.029, 0.035]	0.588 [0.570, 0.605]	0.196 [0.191, 0.204]	3.458 [3.428, 3.494]
25	0.032 [0.027, 0.035]	0.583 [0.559, 0.604]	0.192 [0.184, 0.201]	8.528 [8.475, 8.602]
30	0.026 [0.024, 0.032]	0.582 [0.538, 0.596]	0.188 [0.176, 0.195]	17.848 [17.723, 18.021]
35	0.026 [0.024, 0.028]	0.545 [0.538, 0.565]	0.178 [0.173, 0.182]	32.286 [32.016, 32.531]

Table 1: Median and interquantile range of the approximation errors of the  $\Xi$ -VI posterior  $\hat{q}_{10,N}$  for varying Monte Carlo support sizes  $N$  in Laplace linear regression. For each  $N$ , we run Algorithm 1 with  $\lambda = 10$  and compare  $\hat{q}_{10,N}$  to the exact posterior  $q_0^*$  obtained via MCMC. Across 20 experimental replicates, we report the median and interquantile range of the relative mean error, relative covariance error, the  $W_2$  distance, and the runtime. Although larger  $N$  improves posterior accuracy, the computational cost increases sharply and the accuracy gains beyond  $N \approx 20$  are modest, thus  $N = 20$  is a reasonable choice for accurate and scalable inference in this problem.

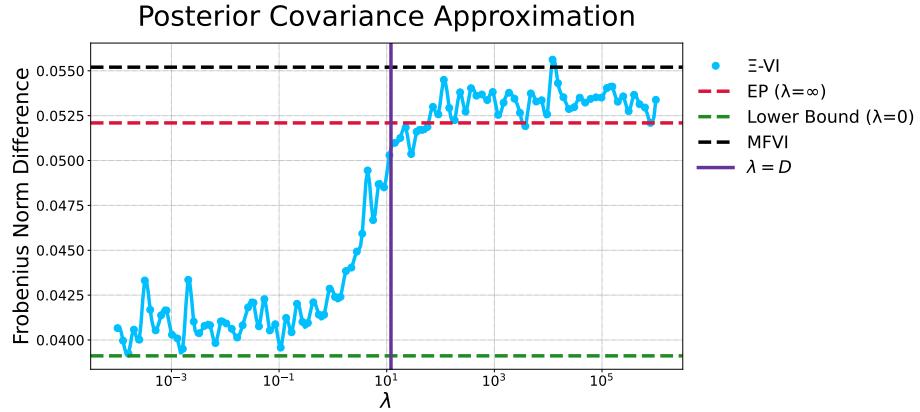


Figure 6: Approximation errors for posterior covariance for Laplace linear regression. The experiment implements Algorithm 1 with expectation propagation as the first step. Errors are quantified using the Frobenius norm and contrasted across a spectrum of  $\lambda$  values, including the theoretical lower bound at  $\lambda = 0$  and the diagonal EP approximation at  $\lambda = \infty$ . The vertical line marks the regularization parameter  $\lambda = D$ .

## Appendix G. Additional Analysis of the 8-Schools Model

In this section, we present additional results for the eight school example in Section 3.3.

Here the goal is to compare treatment effects between schools. We compute the posterior credible intervals for the differences in treatment effects  $\theta_i - \theta_j$  between schools  $i$  and  $j$ . Table 2 shows credible intervals for  $\theta_i - \theta_j$  across ten randomly chosen school pairs, calculated under the exact posterior, MFVI, and  $\Xi$ -VI with  $\lambda \in \{0, 1, 10, 1000\}$ . The results show  $\Xi$ -

Table 2: 95% posterior credible intervals for  $\theta_i - \theta_j$  for 10 randomly selected pairs of schools.

Method	$\theta_2 - \theta_5$	$\theta_6 - \theta_7$	$\theta_2 - \theta_4$	$\theta_4 - \theta_8$	$\theta_1 - \theta_2$
MFVI	[-12.08, 18.48]	[-20.12, 11.37]	[-13.99, 15.79]	[-15.37, 14.97]	[-13.47, 17.10]
True	[-8.50, 14.90]	[-17.74, 7.30]	[-11.28, 12.40]	[-13.02, 12.52]	[-9.21, 16.55]
$\lambda = 0$	[-7.81, 13.77]	[-16.17, 6.49]	[-10.42, 12.06]	[-11.19, 12.70]	[-8.12, 15.81]
$\lambda = 1$	[-9.02, 13.54]	[-16.71, 7.44]	[-11.57, 11.87]	[-11.73, 13.33]	[-9.09, 15.97]
$\lambda = 10$	[-10.43, 13.01]	[-15.77, 8.75]	[-12.57, 12.82]	[-12.21, 14.50]	[-10.54, 15.41]
$\lambda = 1000$	[-10.51, 12.93]	[-15.86, 9.53]	[-12.55, 13.19]	[-12.71, 14.47]	[-10.79, 15.64]

Method	$\theta_2 - \theta_8$	$\theta_3 - \theta_8$	$\theta_5 - \theta_6$	$\theta_2 - \theta_7$	$\theta_3 - \theta_4$
MFVI	[-14.83, 17.07]	[-17.71, 13.88]	[-16.77, 14.19]	[-17.30, 13.41]	[-17.62, 13.49]
True	[-12.09, 12.73]	[-16.08, 11.05]	[-12.62, 10.31]	[-14.97, 9.20]	[-14.79, 10.45]
$\lambda = 0$	[-10.15, 13.41]	[-14.54, 11.56]	[-11.12, 9.70]	[-13.66, 8.31]	[-14.33, 10.41]
$\lambda = 1$	[-11.26, 13.57]	[-14.31, 12.75]	[-11.49, 10.27]	[-15.28, 8.95]	[-14.95, 11.25]
$\lambda = 10$	[-11.58, 14.47]	[-14.69, 13.74]	[-11.15, 10.96]	[-15.09, 10.55]	[-15.18, 12.35]
$\lambda = 1000$	[-12.14, 14.34]	[-14.48, 14.32]	[-12.00, 11.09]	[-14.75, 10.63]	[-15.21, 13.12]

VI, especially with lower  $\lambda$  values, yields intervals that more accurately reflect those derived from the exact posterior, while MFVI produces the most inaccurate intervals.