Validated Variational Inference via Practical Posterior Error Bounds

Jonathan H. Huggins
Boston University

Mikołaj Kasprzak
University of Luxembourg

Trevor Campbell
University of British Columbia

Tamara Broderick
MIT

Abstract

Variational inference has become an increasingly attractive fast alternative to Markov chain Monte Carlo methods for approximate Bayesian inference. However, a major obstacle to the widespread use of variational methods is the lack of post-hoc accuracy measures that are both theoretically justified and computationally efficient. In this paper, we provide rigorous bounds on the error of posterior mean and uncertainty estimates that arise from full-distribution approximations, as in variational inference. Our bounds are widely applicable, as they require only that the approximating and exact posteriors have polynomial moments. Our bounds are also computationally efficient for variational inference because they require only standard values from variational objectives, straightforward analytic calculations, and simple Monte Carlo estimates. We show that our analysis naturally leads to a new and improved workflow for validated variational inference. Finally, we demonstrate the utility of our proposed workflow and error bounds on a robust regression problem and on a real-data example with a widely used multilevel hierarchical model.

1 Introduction

Exact Bayesian statistical inference is known for providing point estimates with desirable decision-theoretic properties as well as coherent uncertainties. Using Bayesian methods in practice, though, typically requires approximating these quantities. Therefore, it is crucial to quantify the error introduced by any approximation. There are two, essentially complementary, options: (1) rigorous *a priori* characterization of accuracy for finite data and (2) tools for evaluating approximation accuracy *a posteriori*. First, consider option #1. Markov chain Monte Carlo (MCMC) methods are the gold standard for sound approximate Bayesian inference in part due to their flexibility and strong *a priori* theoretical guarantees on quality for finite data. However, these guarantees are typically asymptotic in running time, and computational concerns have motivated a spate of alternative Bayesian approximations. Within the machine learning community, variational approaches (Blei et al., 2017; Wainwright et al., 2008) such as black-box and automatic differentiation variational inference (Kingma and Welling, 2014; Kucukelbir et al., 2015; Ranganath et al., 2014) are perhaps the most widely used. While these methods have empirically demonstrated computational gains on problems of interest, they do not come equipped with guarantees on the approximation accuracy of point estimates and uncertainties. There has been some limited but ongoing work in developing relevant *a priori* guarantees for common variational approaches (Alquier and Ridgway, 2017; Alquier et al., 2016; Chérief-Abdellatif and Alquier, 2018; Pati et al., 2018; Wang and Blei, 2018, 2019). There has also been work in developing (boosting) variational algorithms for which it may be possible to obtain *a priori* guarantees on convergence of the approximating distribution to arbitrary accuracy (Campbell and Li, 2019; Guo et al., 2016; Locatello et al., 2018a,b; Miller et al., 2017; Wang, 2016).

The examples above typically either have no guarantees or purely asymptotic guarantees – or require non-convex optimization. Thus, in every case, reliable evaluation tools (option #2) would provide an important bulwark for data analysis (as demonstrated by the widespread use of convergence diagnostics for MCMC (Gelman et al., 2013)). In any particular data analysis, such tools could determine if the approximate point estimates and uncertainties are to be trusted. Gorham and Mackey (2015, 2017); Gorham et al. (2019); Yao et al. (2018) have pioneered initial work in developing evaluation tools applicable to variational inference. However, current methods are either heuristic or cannot be applied in an automated way.

In this paper, we provide the first rigorous, automated, *and* computationally efficient error bounds on the quality of posterior point and uncertainty estimates for variational approximations. We highlight three practical aspects of our bounds here: (A) computational efficiency, (B) weak tail restrictions, and (C) relevant targets. For A, we use only

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

standard values computed in the course of variational inference, straightforward analytic calculations, and simple Monte Carlo (not MCMC) estimates. For B, we require only that the approximating and exact posteriors have polynomial moments – though we show even tighter bounds when exponential moments exist. For C, note that practitioners typically report posterior means for point estimates – and they report posterior variance, standard deviation, or mean absolute deviation for uncertainties (Gelman et al., 2013; Robert, 1994). So we directly bound the error in these quantities. We demonstrate the importance of bounding error in these output quantities directly, rather than bounding divergences between distributions, with illustrative counterexamples; namely, we show that common variational objectives such as the Kullback–Leibler (KL) divergence and α-divergences can be very small at the same time that mean and variance estimates are *arbitrarily* wrong.

To obtain our bounds, we make three main technical contributions, which may all be of independent interest beyond Bayesian methods. First, we show how to bound mean and uncertainty differences in terms of Wasserstein distance. Second, we develop novel bounds on the Wasserstein distance in terms of α-divergences – including the KL divergence – and moment bounds on the variational approximation. The moment conditions allow us to relate (scale-free) α-divergences to (scale-sensitive) Wasserstein distances. Finally, we derive efficiently computable bounds on α-divergences in terms of the objectives already widely used for variational inference – in particular, the evidence lower bound (ELBO) and χ upper bound (CUBO) (Dieng et al., 2017). By combining all three contributions, we obtain efficiently computed bounds on means and uncertainties in terms of the ELBO, CUBO, and certain polynomial or exponential moments of the variational approximation.

Our methods give rise to a new and improved workflow for validated variational inference. We illustrate the usefulness of our bounds as well as the practicality of our new workflow on a toy robust regression problem and a real-data example with a widely used multilevel hierarchical model. A `python` package for carrying out our workflow – including doing black-box variational inference and computing the bounds we develop in this paper – is available at `https://github.com/jhuggins/viabel`. The same repository also contains code for reproducing all of our experiments. Proofs of all our results are in Appendix D.

2 Preliminaries

Bayesian inference. Let $\theta \in \mathbb{R}^d$ denote a parameter vector of interest, and let z denote observed data. A Bayesian model consists of a prior measure $\pi_0(\mathrm{d}\theta)$ and a likelihood $\ell(z;\theta)$. Together, the prior and likelihood define a joint distribution over the data and parameters. The Bayesian posterior distribution π is the conditional in θ with fixed data z.[1] To write this conditional, we define the unnormalized posterior measure $\pi^*(\mathrm{d}\theta) := \ell(z;\theta)\pi_0(\mathrm{d}\theta)$ and the marginal likelihood, or evidence, $M := \int \mathrm{d}\pi^*$. Then the posterior is $\pi := \pi^*/M$.

Typically, practitioners report *summaries* – e.g., point estimates and uncertainties – of the posterior rather than the full posterior. Such summaries include the mean m_π, covariance Σ_π, ith component marginal standard deviation $\sigma_{\pi,i}$, and mean absolute deviation $\mathrm{MAD}_{\pi,i}$: for $\vartheta \sim \pi$,

$$m_\pi := \mathbb{E}(\vartheta), \quad \mathrm{MAD}_{\pi,i} := \mathbb{E}(|\vartheta_i - m_{\pi,i}|),$$
$$\sigma_{\pi,i} := \Sigma_{\pi,ii}^{1/2}, \qquad \Sigma_\pi := \mathbb{E}\{(\vartheta - m_\pi)(\vartheta - m_\pi)^\top\}.$$

Variational inference. In most applications of interest, it is infeasible to efficiently compute these summaries with respect to the posterior distribution in closed form or via simple Monte Carlo. Therefore, one must use an approximate inference method, which produces an approximation $\hat{\pi}$ to the posterior π. The summaries of $\hat{\pi}$ may in turn be used as approximations to the summaries of π. One approach, *variational inference*, is widely used in machine learning. Variational inference aims to minimize some measure of discrepancy $\mathcal{D}_\pi(\cdot)$ over a tractable family $\mathcal{Q}$ of potential approximation distributions (Blei et al., 2017; Wainwright et al., 2008):

$$\hat{\pi} = \operatorname*{arg\,min}_{\xi \in \mathcal{Q}} \mathcal{D}_\pi(\xi).$$

The variational family $\mathcal{Q}$ is chosen to be tractable in the sense that, for any $\xi \in \mathcal{Q}$, we are able to efficiently calculate relevant summaries either analytically or using independent and identically distributed samples from ξ.

KL divergence. The classical choice for the discrepancy in variational inference is the *Kullback–Leibler (KL) divergence* (or *relative entropy*) (Bishop, 2006):

$$\mathrm{KL}(\xi \mid \pi) := \int \log\left(\frac{\mathrm{d}\xi}{\mathrm{d}\pi}\right) \mathrm{d}\xi.$$

Note that the KL divergence is asymmetric in its arguments. The direction $\mathcal{D}_\pi(\xi) = \mathrm{KL}(\xi \mid \pi)$ is most typical in variational inference, largely out of convenience; the unknown marginal likelihood M appears in an additive constant that does not influence the optimization, and computing gradients requires estimating expectations only with respect to $\xi \in \mathcal{Q}$, which is chosen to be tractable. Minimizing $\mathrm{KL}(\xi \mid \pi)$ is equivalent to maximizing the *evidence lower bound* (ELBO; Bishop, 2006):

$$\mathsf{ELBO}(\xi) := \int \log\left(\frac{\mathrm{d}\pi^*}{\mathrm{d}\xi}\right) \mathrm{d}\xi.$$

[1] Since the data z are always fixed throughout this work, we have suppressed the dependence on z in the notation.

Rényi's α-divergence. Another choice of discrepancy for variational inference (Bui et al., 2017; Dieng et al., 2017; Hernández-Lobato et al., 2016; Li and Turner, 2016) is *Rényi's α-divergence*, which for $\alpha \in (0,1) \cup (1,\infty)$ is defined as

$$\mathrm{D}_\alpha(\pi \mid \xi) := \frac{1}{\alpha - 1} \log \int \left(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}\right)^{\alpha-1} \mathrm{d}\pi.$$

The α-divergence is typically used in variational inference with $\mathcal{D}_\pi(\xi) = \mathrm{D}_\alpha(\pi \mid \xi)$ for $\alpha > 1$; again, the unknown marginal likelihood M does not influence the optimization, and estimating gradients is tractable. Variational inference with the α-divergence is equivalent to minimizing a quantity known as the χ *upper bound* (CUBO; Dieng et al., 2017):

$$\mathsf{CUBO}_\alpha(\xi) := (1 - \alpha^{-1})\mathrm{D}_\alpha(\pi \mid \xi) - \log M.$$

The ELBO and CUBO are so-named since they respectively provide a lower and upper bound for $\log M$; see Appendix D.8. The α-divergence generalizes the KL divergence since $\mathrm{D}_\alpha(\pi \mid \xi) := \lim_{\alpha \to 1} \mathrm{D}_\alpha(\pi \mid \xi) = \mathrm{KL}(\pi \mid \xi)$ (Cichocki and Amari, 2010). Note, however, that here the KL divergence has the order of its arguments switched when compared to how it is used for variational inference.

Wasserstein distance. The Wasserstein distance is a measure of discrepancy that, unlike the previous two divergences, is influenced by a metric on the space on which the distributions are defined. It is widely used in the analysis of MCMC and large-scale data asymptotics (e.g., Durmus and Moulines, 2019; Durmus et al., 2019; Eberle and Majka, 2019; Joulin and Ollivier, 2010; Madras and Sezer, 2010; Rudolf and Schweizer, 2018; Vollmer et al., 2016). The p-Wasserstein distance between ξ and π is given by

$$\mathcal{W}_p(\xi, \pi) := \inf_{\gamma \in \Gamma(\xi,\pi)} \left\{ \int \|\theta - \theta'\|_2^p \gamma(\mathrm{d}\theta, \mathrm{d}\theta') \right\}^{1/p},$$

where $\Gamma(\xi, \pi)$ is the set of *couplings* between ξ and π, i.e., Borel measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\xi = \gamma(\cdot, \mathbb{R}^d)$ and $\pi = \gamma(\mathbb{R}^d, \cdot)$ (Villani, 2009, Defs. 6.1 & 1.1). The Wasserstein distance is difficult to use as a variational objective due to the (generally intractable) infimum over couplings, although there is recent work in this direction (Claici et al., 2018; Cuturi and Doucet, 2014; Srivastava et al., 2018).

2.1 Previous work on validating variational approximations

Stein discrepancies. Computable Stein discrepancies provide one approach for evaluating variational approximations (Gorham and Mackey, 2015, 2017; Gorham et al., 2019). That is, the Stein discrepancy between the posterior and variational approximation could be approximated using samples from the variational approximation. However, Stein discrepancy-based bounds on the Wasserstein distance require knowledge of certain properties of the posterior (e.g., one-sided Lipschitz constants) that are usually unavailable without additional analytic effort. Thus, there is not yet an automated way to apply an appropriate Stein operator that guarantees control of the Wasserstein distance (Erdogdu et al., 2018; Gorham et al., 2019).

Pareto-smoothed importance sampling and $\hat{k}$. Pareto-smoothed importance sampling (PSIS; Vehtari et al., 2019) is a method for reducing the variance of importance sampling estimators. The key quantity computed in PSIS is $\hat{k}$, which is an estimate of $k := \inf\{k' \mid \mathrm{D}_{1/k'}(\pi \mid \hat{\pi}) < \infty\}$. Yao et al. (2018) suggest using $\hat{k}$ as a measure of the quality of $\hat{\pi}$. Based on the empirical results and informal arguments of Vehtari et al. (2019), they propose that $\hat{k} \leq 0.5$ indicates a good variational approximation and $\hat{k} \in [0.5, 0.7]$ indicates minimal acceptability. In all cases the authors suggest using PSIS to improve estimates of posterior expectations. However, the link between a small $\hat{k}$ value and a high-quality posterior approximation is only heuristic. We find empirically in Section 5 and Section 5.2 that poor posterior approximations can have small $\hat{k}$ values.

3 Error bounds via posterior discrepancies

Given the concern with posterior summaries, a meaningful measure of posterior approximation quality should control the error in each of these summaries, i.e., $\|m_{\hat{\pi}} - m_\pi\|_2$, $|\,\mathrm{MAD}_{\hat{\pi},i} - \mathrm{MAD}_{\pi,i}\,|$, $\|\Sigma_{\hat{\pi}} - \Sigma_\pi\|_2$, and $|\sigma_{\hat{\pi},i} - \sigma_{\pi,i}|$. To be practical, this measure should also be computationally efficient. We start by focusing on the former challenge: finding a discrepancy that controls the error of these summaries. In particular, we (1) provide counterexamples to show that $\mathrm{KL}(\hat{\pi} \mid \pi)$ and $\mathrm{D}_\alpha(\pi \mid \hat{\pi})$ by themselves cannot be relied upon to control these errors, and (2) prove that the Wasserstein distance does provide the desired control. We address the latter challenge, computational efficiency, in Section 4.

KL divergence. Unfortunately, as we show in the following examples, even when $\mathrm{KL}(\hat{\pi} \mid \pi)$ is small, posterior summary approximations provided by $\hat{\pi}$ can be arbitrarily poor. To get a sense of scale for the KL divergence, we note that the KL divergence from a variational approximation to the exact posterior can easily range from 1 to nearly 500. See Appendix A for further discussion. First we note that the exact posterior standard deviation σ_π is a natural scale for the posterior mean error since changing the posterior mean by σ_π or more could fundamentally change practical decisions made based on the posterior. Our first example shows that even when $\mathrm{KL}(\hat{\pi} \mid \pi)$ is small, the mean error can be arbitrarily large, whether measured relative to σ_π or $\sigma_{\hat{\pi}}$.

Proposition 3.1 (Arbitrarily poor mean approximation)**.** *For any $t > 0$, there exist (A) one-dimensional, unimodal distributions $\hat{\pi}$ and π such that $\mathrm{KL}(\hat{\pi} \mid \pi) < 0.9$ and*

$(m_{\hat{\pi}} - m_\pi)^2 > t\sigma_\pi^2$, *and (B) one-dimensional, unimodal distributions* $\hat{\pi}$ *and* π *such that* $\mathrm{KL}(\hat{\pi} \mid \pi) < 0.3$ *and* $(m_{\hat{\pi}} - m_\pi)^2 > t\sigma_{\hat{\pi}}^2$.

In more detail, let $\mathsf{Weibull}(k, 1)$ denote the Weibull distribution with shape $k > 0$ and scale 1. For (A), for any $t > 0$, we can choose $k = k(t)$, $\pi = \mathsf{Weibull}(k, 1)$, and $\hat{\pi} = \mathsf{Weibull}(k/2, 1)$, where $k(t) \searrow 0$ as $t \to \infty$. We exchange the two distributions for (B).

Our second example shows that $\mathrm{KL}(\hat{\pi} \mid \pi)$ can remain small even when the variance difference is arbitrarily large.

Proposition 3.2 (Arbitrarily poor variance approximation)**.** *For any* $t \in (1, \infty]$*, there exist one-dimensional, mean-zero, unimodal distributions* $\hat{\pi}$ *and* π *such that* $\mathrm{KL}(\hat{\pi} \mid \pi) < 0.12$ *but* $\sigma_\pi^2 \geq t\sigma_{\hat{\pi}}^2$.

Here, for any $t > 0$ we let $h = h(t)$, $\pi = \mathcal{T}_h$ (standard t-distribution with h degrees of freedom), and $\hat{\pi} = \mathcal{N}(0, 1)$ (standard Gaussian), where $h(t) \searrow 2$ as $t \to \infty$.

Rényi's α-divergence. We similarly demonstrate that small $\mathrm{D}_\alpha(\pi \mid \hat{\pi})$ does not imply accurate mean or variance estimates. We focus on the canonical case $\alpha = 2$, which will also play a key role in our analyses below.

Proposition 3.3 (Arbitrarily poor mean and variance approximation)**.** *For any* $t > 0$*, there exist one-dimensional, unimodal distributions* $\hat{\pi}$ *and* π *with* $\mathrm{D}_2(\pi \mid \hat{\pi}) < 0.4$ *such that* $\sigma_{\hat{\pi}}^2 \geq t\sigma_\pi^2$ *and* $(m_{\hat{\pi}} - m_\pi)^2 \geq t\sigma_\pi^2$.

We again take $\pi = \mathsf{Weibull}(k, 1)$ and $\hat{\pi} = \mathsf{Weibull}(k/2, 1)$ with $k = k(t) \searrow 0$ as $t \to \infty$.

Wasserstein distance. In contrast to both the KL and α-divergences, the Wasserstein distance accounts for the metric on the underlying space. Intuitively, the Wasserstein distance is large when the mass of two distributions is "far apart." Thus, it is a natural choice of discrepancy for bounding the error in the approximate posterior mean and uncertainty, since these quantities also depend on the underlying metric. Our next result confirms that the Wasserstein distance controls the error in these quantities.

Theorem 3.4. *If* $\mathcal{W}_1(\hat{\pi}, \pi) \leq \varepsilon$ *or* $\mathcal{W}_2(\hat{\pi}, \pi) \leq \varepsilon$*, then*

$$\|m_{\hat{\pi}} - m_\pi\|_2 \leq \varepsilon \text{ and } \max_i |\,\mathrm{MAD}_{\hat{\pi},i} - \mathrm{MAD}_{\pi,i}\,| \leq 2\varepsilon.$$

If $\mathcal{W}_2(\hat{\pi}, \pi) \leq \varepsilon$*, then, for* $S := \sqrt{\min\{\|\Sigma_{\hat{\pi}}\|_2, \|\Sigma_\pi\|_2\}}$,

$$\max_i |\sigma_{\hat{\pi},i} - \sigma_{\pi,i}| \leq \varepsilon \text{ and } \|\Sigma_{\hat{\pi}} - \Sigma_\pi\|_2 < 2\varepsilon(S + \varepsilon).$$

Remark 3.5. The Wasserstein distance can also be used to bound the difference between expectations of any smooth function. More precisely, if the function ϕ satisfies $|\phi'| \leq L$ and $\mathcal{W}_1(\hat{\pi}, \pi) \leq \varepsilon$, then $|\int \phi \,\mathrm{d}\hat{\pi} - \int \phi \,\mathrm{d}\pi| \leq \varepsilon L$.

While our focus is on parameter inference, there are many cases when we are interested in *predictive accuracy*, including in Bayesian deep learning. In such cases, Wasserstein bounds on the posterior remain useful. Assuming $f(z_{\text{new}} \mid \theta)$ is the distribution for new data given parameter θ, the posterior predictive distribution is $\mu(z_{\text{new}}) := \int f(z_{\text{new}} \mid \theta)\pi(\mathrm{d}\theta)$, with the approximate posterior predictive $\hat{\mu}$ defined analogously.

Proposition 3.6. *If* $\mathcal{W}_p(f(\cdot \mid \theta), f(\cdot \mid \theta')) \leq C\,\|\theta - \theta'\|_2$ *for some* $C \geq 0$ *and* $\mathcal{W}_p(\hat{\pi}, \pi) \leq \varepsilon$*, then* $\mathcal{W}_p(\hat{\mu}, \mu) \leq \varepsilon C$.

Remark 3.7. The assumption on $f(\cdot \mid \theta)$ in Proposition 3.6 holds for many commonly used distributions including Gaussian distributions with fixed variance and the Bernoulli distribution with softmax parameterization.

4 A complete workflow for validated variational inference

In this section, we develop a comprehensive approach to variational inference with rigorously validated output. First we prove a number of new results which let us bound the Wasserstein distance between the variational and true posteriors in terms of quantities that can be efficiently computed or upper-bounded. These Wasserstein bounds, when combined with ideas from importance sampling, provide the tools for formulating our proposed workflow.

4.1 Computationally efficient error bounds

In this section we return to the question of computationally efficient posterior error bounds. In particular, although we have shown that the Wasserstein distance provides direct control of the error in approximate posterior summaries of interest, it itself is not tractable to compute or estimate. Our general strategy in this section is use standard variational objectives – namely, the ELBO and CUBO – to bound the Wasserstein distance. We thereby achieve bounds on the error of posterior summaries by Theorem 3.4. More detail about our results in this section and related work can be found in, respectively, Appendices B and C.

Our process consists of two steps. First, we use tail properties of the distribution $\xi \in \mathcal{Q}$ to arrive at bounds on the Wasserstein distance via the KL or α-divergence. Second, we use ELBO and CUBO to bound the KL and α-divergences.

A key challenge in realizing our goal is bounding a scale-dependent distance (the Wasserstein distance) with a scale-invariant divergence (the KL or α-divergence). To see the scale-invariance, we note a broader result: these divergences are invariant to reparameterization. For a transformation $T : \mathbb{R}^d \to \mathbb{R}^d$, let $T\#\eta$ denote the pushforward measure of η, which is the distribution of the random variable $T(\vartheta)$ for $\vartheta \sim \eta$.

Lemma 4.1. *The KL and* α*-divergence are invariant under a smooth, invertible transformation* T*, i.e.,* $\mathrm{D}_\alpha(\eta \mid \nu) = \mathrm{D}_\alpha(T\#\eta \mid T\#\nu)$ *and* $\mathrm{KL}(\eta \mid \nu) = \mathrm{KL}(T\#\eta \mid T\#\nu)$.

Yao et al. (2018) make a similar observation. A simple

example illustrates that the Wasserstein distance is not invariant to reparameterization. For $\sigma > 0$, let $\nu_\sigma(\theta) := \sigma^{-d}\nu(\theta/\sigma)$ define the rescaled version of ν (with η_σ defined analogously). Then $\mathcal{W}_p(\eta_\sigma, \nu_\sigma) = \sigma\mathcal{W}_p(\eta, \nu)$. It follows that any bound of a scale-dependent distance such as the Wasserstein distance using a scale-invariant divergence must incorporate some notion of scale. Toward that end, we start by defining the moment constants $C_p^{\mathrm{PI}}(\xi)$ and $C_p^{\mathrm{EI}}(\xi)$ and associated tail behaviors. For $p \geq 1$, we say that ξ is *p-polynomially integrable* if

$$C_p^{\mathrm{PI}}(\xi) := 2 \inf_{\theta_0} \left\{ \int \|\theta - \theta_0\|_2^p \xi(\mathrm{d}\theta) \right\}^{\frac{1}{p}} < \infty$$

and that ξ is *p-exponentially integrable* if

$$C_p^{\mathrm{EI}}(\xi) := 2 \inf_{\theta_0, \epsilon > 0} \left[\frac{1}{\epsilon} \left\{ \frac{3}{2} + \log \int e^{\epsilon \|\theta - \theta_0\|_2^p} \xi(\mathrm{d}\theta) \right\} \right]^{\frac{1}{p}} < \infty.$$

Assuming the variational approximation $\hat{\pi}$ has polynomial (respectively, exponential) tails, our next result provides a bound on the p-Wasserstein distance using the 2-divergence (respectively, the KL divergence).

Proposition 4.2. *If $\pi \ll \hat{\pi}$,[2] then*

$$\mathcal{W}_p(\hat{\pi}, \pi) \leq C_{2p}^{\mathrm{PI}}(\hat{\pi}) \left[\exp\left\{ \mathrm{D}_2(\pi \mid \hat{\pi}) \right\} - 1 \right]^{\frac{1}{2p}}$$

and

$$\mathcal{W}_p(\hat{\pi}, \pi) \leq C_p^{\mathrm{EI}}(\hat{\pi}) \left[\mathrm{KL}(\pi \mid \hat{\pi})^{\frac{1}{p}} + \left\{ \mathrm{KL}(\pi \mid \hat{\pi})/2 \right\}^{\frac{1}{2p}} \right].$$

Our next results confirm that, even though Proposition 4.2 uses KL and α-divergences, our bounds capture the growth in Wasserstein distance, as desired, and thus do not suffer the pathologies observed in Propositions 3.1 to 3.3.

Proposition 4.3 (cf. Propositions 3.1 and 3.3)**.** *For a fixed $k \in (0, \infty)$, let $\eta = \mathsf{Weibull}(k/2, 1)$ and $\nu = \mathsf{Weibull}(k, 1)$. Then, for $\alpha > 1$, $\mathrm{D}_\alpha(\eta \mid \nu) = \infty$. On the other hand, $\mathrm{D}_\alpha(\nu \mid \eta) < \infty$; but, as $k \searrow 0$, the moment constant from Proposition B.2 satisfies $C_p^{\mathrm{PI}}(\eta) \nearrow \infty$.*

Proposition 4.4 (cf. Proposition 3.2)**.** *If η is a standard normal measure and $\nu = \mathcal{T}_h$ is a standard t-distribution with $h \geq 2$ degrees of freedom, then $\mathrm{D}_\alpha(\eta \mid \nu) < \infty$. However, as $h \searrow 2$, we have $C_p^{\mathrm{PI}}(\nu) \nearrow \infty$.*

Next, we turn to showing how we can use the ELBO and CUBO to bound the KL and α-divergences that appear in Proposition 4.2. For $\alpha > 1$ and any distribution η, define

$$\mathrm{H}_\alpha(\xi, \eta) := \tfrac{\alpha}{\alpha - 1} \left\{ \mathsf{CUBO}_\alpha(\xi) - \mathsf{ELBO}(\eta) \right\}.$$

Lemma 4.5. *For any distribution η such that $\pi \ll \eta$,*

$$\mathrm{KL}(\pi \mid \hat{\pi}) \leq \mathrm{D}_\alpha(\pi \mid \hat{\pi}) \leq \mathrm{H}_\alpha(\hat{\pi}, \eta).$$

[2] $\pi \ll \hat{\pi}$ denotes π is absolutely continuous with respect to $\hat{\pi}$.

Then, combining Proposition 4.2 and Lemma 4.5 yields the desired bounds on the p-Wasserstein distance given only the quantities $C_{2p}^{\mathrm{PI}}(\hat{\pi})$, $C_p^{\mathrm{EI}}(\hat{\pi})$, $\mathsf{CUBO}_\alpha(\hat{\pi})$, and $\mathsf{ELBO}(\eta)$, all of which can be either efficiently estimated or bounded (with high probability); we address these computational issues in detail in Section 4.3.

Theorem 4.6. *For any $p \geq 1$ and any distribution η, if $\pi \ll \hat{\pi}$, then*

$$\mathcal{W}_p(\hat{\pi}, \pi) \leq C_{2p}^{\mathrm{PI}}(\hat{\pi}) \left[\exp\left\{ \mathrm{H}_2(\hat{\pi}, \eta) \right\} - 1 \right]^{\frac{1}{2p}}$$

and

$$\mathcal{W}_p(\hat{\pi}, \pi) \leq C_p^{\mathrm{EI}}(\hat{\pi}) \left[\mathrm{H}_2(\hat{\pi}, \eta)^{\frac{1}{p}} + \left\{ \mathrm{H}_2(\hat{\pi}, \eta)/2 \right\}^{\frac{1}{2p}} \right].$$

4.2 Importance sampling

Before presenting our workflow for validated variational inference, we briefly discuss a final ingredient: importance sampling. Standard importance sampling with importance distribution $\hat{\pi}$ operates as follows. After obtaining samples $\theta_1, \ldots, \theta_T \sim \hat{\pi}$, we can define importance weights $w_t := \pi^*(\theta_t)/\hat{\pi}(\theta_t)$ and self-normalized weights $\tilde{w}_t := w_t / \sum_{t=1}^T w_t$. Then, the importance sampling estimator for $\int \phi \, \mathrm{d}\pi$ is $\sum_{t=1}^T \tilde{w}_t \phi(\theta_t)$. Importance sampling can decrease the bias at the cost of some additional variance relative to the simple Monte Carlo estimate $T^{-1} \sum_{t=1}^T \phi(\theta_t)$. Recall from Section 2.1 that Pareto-smoothed importance sampling (PSIS) can improve upon standard importance sampling by significantly reducing variance without much extra bias. In additional, PSIS provides a crucial diagnostic quantity, $\hat{k}$. When $\hat{k} > 0.7$, the importance weights are too high-variance to be reliable, even when using PSIS.

Our approach to bounding the Wasserstein distance in terms of the α-divergence has intriguing connections to the theory of importance sampling. As pointed out by Dieng et al. (2017), minimizing the 2-divergence is equivalent to minimizing the variance of the (normalized) importance weight $\pi(\theta_t)/\hat{\pi}(\theta_t)$, which is equal to $\exp\{\mathrm{D}_2(\pi \mid \hat{\pi})\} - 1$. Moreover, the estimation error of importance sampling can be bounded as a function of $\mathrm{KL}(\pi \mid \hat{\pi})$ (Chatterjee and Diaconis, 2018), which is upper bounded by $\mathrm{D}_2(\pi \mid \hat{\pi})$. Thus, minimizing the 2-divergence simultaneously leads to better importance distributions and smaller Wasserstein error – as long as the moments of the variational approximation do not increase disproportionately to the 2-divergence decrease. In practice such pathological behavior appears to be unusual; see Section 5 and Dieng et al. (2017, §3).

4.3 A workflow for black box variational inference

The usual approach to black box variational inference is (1) to choose $\mathcal{D}_\pi(\xi) = \mathrm{KL}(\xi \mid \pi)$ (i.e., to maximize $\mathsf{ELBO}(\xi)$), and (2) to use (products of) Gaussians as the variational family $\mathcal{Q}$ (Carpenter et al., 2017a; Kucukelbir

et al., 2015; Ranganath et al., 2014; Salvatier et al., 2016). Based on Theorem 4.6 and our discussion in Section 4.2, we suggest a number of deviations from the typical variational inference procedure, including integrating checks based on our novel bounds. We first provide an outline of our default workflow recommendation, then discuss each step in detail – along with some potential refinements. We show the workflow in action in Section 5. We write $\mathcal{Q}_h^{\mathcal{T}}$ to denote the mean-field variational family consisting of product of t-distributions with h degrees of freedom.

Algorithm 1: Validated variational workflow

1 Set $\mathcal{Q}$ to be $\mathcal{Q}_{40}^{\mathcal{T}}$
2 Find $\hat{\pi} \in \mathcal{Q}$ that minimizes $\mathsf{CUBO}_2(\xi)$
3 **if** $\hat{k} > 0.7$ **then**
  Refine choice of $\mathcal{Q}$ or reparameterize the model
  Return to step 2
4 Find $\eta \in \mathcal{Q}$ that maximizes $\mathsf{ELBO}(\xi)$
5 Estimate $\mathsf{ELBO}(\eta)$ and $\mathsf{CUBO}_2(\hat{\pi})$ via Monte Carlo
6 Use Lemma 4.5 to compute bound $\bar{\delta}_2 \geq \mathrm{D}_2(\pi \mid \hat{\pi})$
7 Use Theorem 4.6 to compute bound $\bar{w}_2 \geq \mathcal{W}_2(\pi, \hat{\pi})$
8 **if** $\bar{\delta}_2$ *and* $\bar{w}_2$ *are large* **then**
  Refine choice of $\mathcal{Q}$ or reparameterize the model
  Return to step 2
else if $\bar{\delta}_2$ *and* $\bar{w}_2$ *are very small* **then**
  approximate π with $\hat{\pi}$
else when $\bar{\delta}_2$ and $\bar{w}_2$ are moderately small
  Use PSIS to refine the posterior expectations produced by $\hat{\pi}$

For step 1, we choose a heavy-tailed variational family to ensure that the 2-divergence (and hence $\mathsf{CUBO}_2(\xi)$) is finite (that is, such that $\mathrm{D}_2(\pi \mid \xi) < \infty$ for all $\xi \in \mathcal{Q}$). The choice of 40 degrees of freedom is somewhat arbitrary. Slightly different choices should produce similar results. It is also possible to select a different variational family specific to the problem at hand as long as the 2-divergence is guaranteed to be finite. For step 2, we minimize the CUBO to obtain as tight a bound as possible when we apply Theorem 4.6 (though note that usually the CUBO objective – like the negative ELBO – is non-convex, so we may not be able to find the global minimum). Toward the same end, in step 4 we separately find the distribution η that results in largest ELBO. However, before going to the effort of finding η, in step 3 we check that $\hat{k} \leq 0.7$, since otherwise our estimate of $\mathsf{CUBO}_2(\hat{\pi})$ is not reliable and thus we should not trust any bounds on the 2-divergence or Wasserstein distance computed using Lemma 4.5 and Theorem 4.6. How precisely to refine the choice of $\mathcal{Q}$ or reparameterize the model is problem-dependent. One possibility is to use multivariate t-distributions with h degrees of freedom for $\mathcal{Q}$; unlike $\mathcal{Q}_h^{\mathcal{T}}$, the multivariate versions can capture correlations in the posterior.

For step 5, we can use simple Monte Carlo to compute high-accuracy estimates for $\mathsf{ELBO}(\eta)$ and $\mathsf{CUBO}_2(\hat{\pi})$:

$$\widehat{\mathsf{CUBO}}_2(\hat{\pi}) := \tfrac{1}{2}\log\left[\tfrac{1}{T}\sum_{t=1}^{T}\left\{\tfrac{\mathrm{d}\pi^*}{\mathrm{d}\hat{\pi}}(\theta_t^{\hat{\pi}})\right\}^2\right], \; (\theta_t^{\hat{\pi}})_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} \hat{\pi}$$
$$\widehat{\mathsf{ELBO}}(\eta) := \tfrac{1}{T}\sum_{t=1}^{T}\log\tfrac{\mathrm{d}\pi^*}{\mathrm{d}\eta}(\theta_t^{\eta}), \qquad (\theta_t^{\eta})_{t=1}^{T} \overset{\text{i.i.d.}}{\sim} \eta.$$

Ensuring the accuracy of $\widehat{\mathsf{CUBO}}_2(\hat{\pi})$ and $\widehat{\mathsf{ELBO}}(\eta)$ reduces to the well-studied problem of estimating the accuracy of a simple Monte Carlo approximation (e.g., Koehler et al., 2009). We can also convert these estimates into high-probability upper bounds using standard concentration inequalities (Boucheron et al., 2013). For step 6, we use Lemma 4.5 to obtain the estimated 2-divergence bound

$$\bar{\delta}_2 := \widehat{\mathrm{H}}_2(\hat{\pi}, \eta) := 2\left\{\widehat{\mathsf{CUBO}}_2(\hat{\pi}) - \widehat{\mathsf{ELBO}}(\eta)\right\}.$$

For step 7, to compute Wasserstein bounds using Theorem 4.6, we can bound $C_{2p}^{\mathrm{PI}}(\hat{\pi})$ using the central moments of the distribution: if $\hat{\pi} = \prod_{i=1}^{d}\mathcal{T}_h(\mu_i, \sigma_i)$ and $C_h := h/(h-2)$, then

$$C_2^{\mathrm{PI}}(\hat{\pi}) \leq 2C_h\textstyle\sum_{i=1}^{d}\sigma_i^2$$
$$C_2^{\mathrm{PI}}(\hat{\pi}) \leq 2C_h^2\left\{\tfrac{2(h-1)}{h-4}\textstyle\sum_{i=1}^{d}\sigma_i^2 + (\sum_{i=1}^{d}\sigma_i^2)^2\right\}.$$

Since $\widehat{C}_p^{\mathrm{EI}}(\hat{\pi}) = \infty$ for t-distribution variational families, we cannot use the second bound from Theorem 4.6. For variational families without analytically computable moments, we can bound the moment constants $C_p^{\mathrm{PI}}(\hat{\pi})$ and $C_p^{\mathrm{EI}}(\hat{\pi})$ by fixing any θ_0, ϵ and sampling from $\hat{\pi}$. We can intuitively think of θ_0 as the "center" of the distribution, so a natural choice is setting it equal to the mean of $\hat{\pi}$.

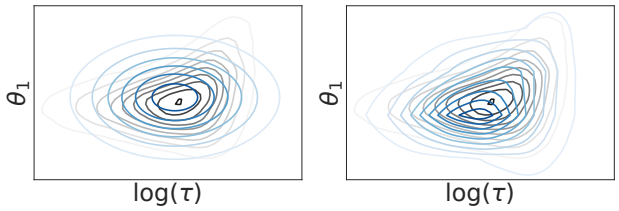
For step 8, what qualifies as a moderately or very small $\bar{w}_2$ value will depend on the desired accuracy and natural scale of the problem. $\bar{\delta}_2$ has a more universal scale; in particular, $\bar{\delta}_2 < 4.6$ could be treated as moderately small since the variance of the importance weights is $\exp\{\mathrm{D}_2(\hat{\pi} \mid \pi))\} - 1 < 100$, so PSIS with a reasonable number of samples should be effective; for some $\delta_* \ll 1$ (for example, $\delta_* = 0.01$), $\bar{\delta}_2 < \delta_*$ could be treated as very small, since the term multiplying $C_p^{\mathrm{PI}}(\xi)$ in Proposition 4.2 and Theorem 4.6 will be less than $\delta_*^{1/p}$.

5 Two case studies

Next we demonstrate our variational inference workflow and the usefulness of our bounds through two case studies.

5.1 Case study #1: the eight schools model

We apply our variational workflow to approximate the posterior for the eight schools data and model (Gelman et al., 2013, Sec. 5.5), a canonical example of a Bayesian hierarchical analysis. Yao et al. (2018) previously considered this model in the setting of evaluating variational inference. In



(a) centered (b) non-centered

Figure 1: Approximate posteriors for CHIVI variational approximations (blue) and for HMC (black; ground truth).

the eight schools data, we have observations corresponding to the mean y_n and standard deviation σ_n of a treatment effect at each of eight schools, indexed by $n \in \{1, \dots, 8\}$. The goal is to estimate the overall treatment effect μ, the standard deviation τ of school-level treatment effects, and the true school-level treatment effects θ_n. There are two standard ways to parameterize the model. The *centered parameterization* is

$$y_n \mid \theta_n \sim \mathcal{N}(\theta_n, \sigma_n), \ \theta_n \mid \mu, \tau \sim \mathcal{N}(\mu, \tau), \qquad (1)$$
$$\mu \sim \mathcal{N}(0, 5), \qquad \tau \sim \text{half-Cauchy}(0, 5).$$

The *non-centered parameterization* decouples θ and τ through the transformation $\tilde{\theta}_n = (\theta_n - \mu_n)/\tau$ and replaces Eq. (1) with

$$y_n \mid \tilde{\theta}_n \sim \mathcal{N}(\mu + \tau\tilde{\theta}_n), \qquad \tilde{\theta}_n \sim \mathcal{N}(0, 1).$$

The standard deviation of the y_n is 9.8 and the median σ_n value is 11, which suggests the overall scale of the problem is roughly 10.

Diagnosing a poorly parameterized model. We begin by considering the centered parameterization. Following steps 1 and 2, we use $\mathcal{Q} = \mathcal{Q}_{40}^{\mathcal{T}}$ and minimize the CUBO using CHIVI (Dieng et al., 2017) to obtain $\hat{\pi}$. The results appear in the first column of Table 1. Since $\hat{k} > 0.7$, according to step 3 we should either reparameterize the model or choose a different $\mathcal{Q}$. If we compute $\bar{\delta}_2$ and $\bar{w}_2$, then we reach the same conclusion. Thus, these diagnostics correctly determine that $\hat{\pi}$ is a poor approximation to the posterior and that PSIS does not provide an improvement.

Correctly validating an improved parameterization. Fig. 1a compares approximate posteriors from CHIVI and Hamiltonian Monte Carlo (HMC; Neal, 2011) – namely the dynamic HMC implementation in Stan (Carpenter et al., 2017b; Hoffman and Gelman, 2014). The HMC samples serve as ground truth. This comparison illustrates why the centered parameterization is not conducive to variational inference when using a mean-field variational family: the conditional variance of any θ_n is strongly dependent on τ. We can remedy this issue by using the non-centered parameterization. Repeating steps 1 and 2, we now find that $\hat{k} = 0.55 \leq 0.7$, suggesting the variational approximation

	centered df = 40	non-centered df = 40	non-centered df = 8
D_2 bound $\bar{\delta}_2$	14	1.6	3.8
$\hat{k}$	0.88	0.55	0.40
W_2 bound $\bar{w}_2$	983	15	29
mean error	0.10	0.14	0.18
with PSIS	0.40	0.04	0.03
std. dev. error	1.1	0.03	0.25
with PSIS	1.2	0.03	0.02
covariance error	8.4	0.95	1.3
with PSIS	5.4	0.45	0.32

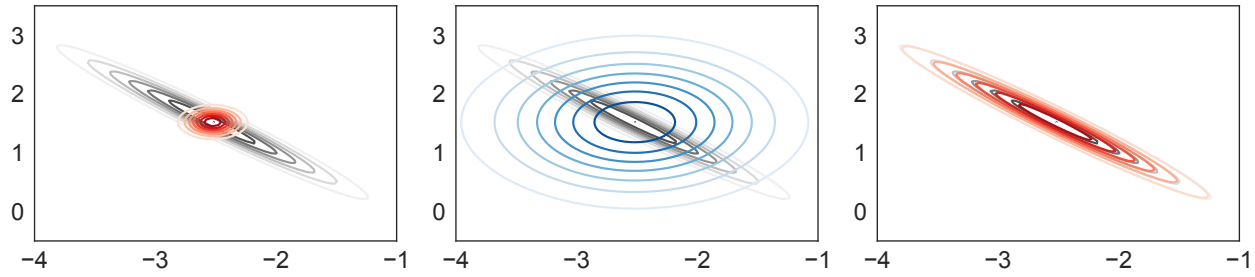
Table 1: Results for eight schools model for the parameter vector $(\mu, \log\tau, \theta_1, \dots, \theta_8)$. The mean and standard deviation errors are defined as, respectively, $\|m_\pi - m_{\hat{\pi}}\|_2$ and $\|\sigma_\pi - \sigma_{\hat{\pi}}\|_2$. The covariance error is defined as $\|\Sigma_\pi - \Sigma_{\hat{\pi}}\|_2^{1/2}$. We use the square root for the covariance error in order to place it on the same scale as the mean error, standard deviation error, and the 2-Wasserstein bound. To provide a sense of the overall scale, we note that $\|\Sigma_\pi\|_2^{1/2} = 9.7$.

is at least acceptable as an importance distribution (step 3). However, $\bar{\delta}_2$ and $\bar{w}_2$ remain at best moderately small, so the variational approximation should not be used directly (steps 4–8). These diagnostic results are confirmed graphically in Fig. 1b and quantitatively in the second column of Table 1. Furthermore, applying PSIS does reduce approximation error, as expected.

Limitations of $\hat{k}$. So far $\hat{k}$, $\bar{\delta}_2$, and $\bar{w}_2$ have all provided similar diagnostic information. To see how they can diverge, we repeat our workflow, but this time we use $\mathcal{Q} = \mathcal{Q}_8^{\mathcal{T}}$. The results appear in the final column of Table 1. Variational approximations with heavier tails should decrease k and $\hat{k}$ since the importance weights will have

	mean-field KLVI	mean-field CHIVI	full-rank KLVI
D_2 bound	8.7	4.9	6×10^{-3}
$\hat{k}$	0.92	0.34	-0.93
W_2 bound	4.4	8.4	0.39
mean error	0.01	< 0.01	0.01
with PSIS	0.06	0.01	0.01
std. dev. error	0.73	0.09	< 0.01
with PSIS	0.49	< 0.01	< 0.01
covariance error	0.92	0.72	$< .1$
with PSIS	0.82	0.11	$< .1$

Table 2: Results for robust regression. See Table 1 for further explanation. To provide a sense of the overall scale, we note that $\|\Sigma_\pi\|_2^{1/2} = 0.93$.



(a) mean-field KLVI (b) mean-field CHIVI (c) full-rank KLVI

Figure 2: For robust regression, approximate posteriors for variational approximations (KLVI red, CHIVI blue) and for the exact posterior (black).

more finite moments. While $\hat{\pi} \in \mathcal{Q}_8^{\mathcal{T}}$ has a small $\hat{k}$ value (0.40), its accuracy is worse than the non-centered $\hat{\pi} \in \mathcal{Q}_{40}^{\mathcal{T}}$ ($\hat{k} = 0.55$). The lower quality of $\hat{\pi} \in \mathcal{Q}_8^{\mathcal{T}}$ is, however, accurately reflected in the larger $\bar{\delta}_2$ and $\bar{w}_2$ values. The accuracy of PSIS using $\hat{\pi} \in \mathcal{Q}_8^{\mathcal{T}}$, however, is better than that of PSIS using the non-centered $\hat{\pi} \in \mathcal{Q}_{40}^{\mathcal{T}}$. In sum, while $\hat{k}$ does provide a useful diagnostic for when $\hat{\pi}$ will serve as a good importance distribution, it does not provide a reliable heuristic for the accuracy of $\hat{\pi}$ as an approximation to π. Hence, our workflow uses $\hat{k}$ only to validate the use of importance sampling, but not the quality of $\hat{\pi}$.

5.2 Case study #2: robust regression

A second example of our variational inference workflow confirms our findings in the eight schools example and further clarifies the differences between $\hat{k}$ and our bounds. Specifically, we consider the posterior of the toy robust regression model with coefficients $\theta \in \mathbb{R}^d$ and observed data $(x_1, y_1), \ldots, (x_N, y_N) \in \mathbb{R}^{d+1}$:

$$\theta_i \sim \mathcal{N}(0, 10), \qquad y_n \mid x_n, \theta \sim \mathcal{T}_{40}(\beta^\top x_n, 1). \tag{2}$$

We take $d = 2$ and $N = 25$. We generate data according to Eq. (2) with $\beta = (-2, 1)$ and each x_n Gaussian-distributed, with $\mathrm{var}(x_{ni}) = 1$ and $\mathrm{cov}(x_{n1}, x_{n2}) = 0.75$.

A poor quality approximation. For illustrative purposes, first we approximate the posterior with standard black-box variational inference (i.e., by maximizing the ELBO). We refer to this method as KLVI. As before, we use $\mathcal{Q} = \mathcal{Q}_{40}^{\mathcal{T}}$. The results appear in the first column of Table 2. The variational approximation is poor due to the strong posterior correlation between θ_1 and θ_2 (Fig. 2a), and PSIS is unable to correct for the underdispersion of the posterior approximation. The large values of $\bar{d}_2$, $\bar{w}_2$, and $\hat{k}$ accurately reflect these findings.

$\hat{k}$ does not detect a poor quality CHIVI approximation. Next, we instead use CHIVI with $\mathcal{Q}_{40}^{\mathcal{T}}$. The results appear in the second column of Table 2. The variational approximation is in some sense better because it no longer underestimates the marginal variances (Fig. 2b), which is reflected in the smaller standard deviation error. However, since the mean-field family cannot capture the posterior correlation structure, the covariance error remains large. The poor covariance approximation is reflected in large $\bar{d}_2$ and $\bar{w}_2$ values; however, it is not reflected in the small $\hat{k}$ value (0.34).

The complementary roles of $\bar{\delta}_2$ and $\bar{w}_2$ versus $\hat{k}$. We have just seen that the small $\hat{k}$ value when using CHIVI with a mean-field family does not reflect the quality of the variational approximation. However, the PSIS errors are small, so $\hat{k}$ *does* accurately capture the fact that the CHIVI approximation is a good importance sampling distribution. Thus, $\hat{k}$ provide complementary information to $\bar{\delta}_2$ and $\bar{w}_2$. This complementarity is further illustrated when using KLVI with a full-rank variational family (Fig. 2c and the final column of Table 2). Now the posterior approximation is very accurate, which is reflected in the very small $\bar{d}_2$ value and fairly small $\bar{w}_2$ value. Step 8 of our workflow suggests not using PSIS when $\bar{d}_2$ and $\bar{w}_2$ are small since it will be difficult to improve the posterior approximation accuracy. Applying PSIS confirms that importance sampling is not necessary: although the $\hat{k}$ value is extremely small (in fact, negative), PSIS and $\hat{\pi}$ provide nearly identical accuracy.

6 Conclusion

In conclusion, as we have shown through both theory and experiment, our workflow for validated variational inference potentially provides a framework for making variational methods more competitive with Markov chain Monte Carlo. We end by noting that our work complements recent proposals for making variational approximations arbitrarily accurate (Campbell and Li, 2019; Guo et al., 2016; Locatello et al., 2018a,b; Miller et al., 2017; Wang, 2016) since our bounds can provide a stopping criteria for when a variational approximation no longer needs to be improved.

Acknowledgements

The authors thank Sushrutha Reddy for pointing out some improvements to our Wasserstein bounds on the standard deviation and variance, and also Daniel Simpson, Lester Mackey, Arthur Gretton, and Pierre Jacob for valuable discussions and many useful references. M. Kasprzak was supported in part by an EPSRC studentship and FNR grant FoRGES (R-AGR-3376-10). T. Campbell was supported by a National Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and Discovery Launch Supplement. T. Broderick was supported in part by an NSF CAREER Award, an ARO YIP Award, the Office of Naval Research, a Sloan Research Fellowship, the CSAIL-MSR Trustworthy AI Initiative, and DARPA.

References

P. Alquier and J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *arXiv.org*, June 2017.

P. Alquier, J. Ridgway, and N. Chopin. On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research*, 17:1–41, 2016.

P. Baqué, F. Fleuret, and P. Fua. Multi-modal Mean-Fields via Cardinality-Based Clamping. *CVPR*, pages 4303–4312, 2017.

C. Bauckhage. Computing the Kullback-Leibler Divergence between two Weibull Distributions. *arXiv.org*, 2013.

C. M. Bishop. *Pattern recognition and machine learning*, chapter 10: Approximate Inference. Springer, 2006.

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

S. G. Bobkov and F. Götze. Exponential Integrability and Transportation Cost Related to Logarithmic Sobolev Inequalities. *Journal of Functional Analysis*, 163:1–28, 1999.

F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la faculte des sciences de Toulouse*, 13(3):331–352, 2005.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

T. D. Bui, J. Yan, and R. E. Turner. A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation. *Journal of Machine Learning Research*, 18:1–72, Oct. 2017.

T. Campbell and X. Li. Universal boosting variational inference. In *Advances in Neural Information Processing Systems*, 2019.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017a.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017b.

S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, Apr. 2018.

B.-E. Chérief-Abdellatif and P. Alquier. Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, 12 (2):2995–3035, 2018.

A. Cichocki and S.-I. Amari. Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities. *Entropy*, 12(6):1532–1568, June 2010.

S. Claici, E. Chien, and J. Solomon. Stochastic Wasserstein barycenters. In *International Conference on Machine Learning*, 2018.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. M. Blei. Variational Inference via χ Upper Bound Minimization. In *Advances in Neural Information Processing Systems*, 2017.

H. Djellout, A. Guillin, and L. Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *The Annals of Probability*, 32(3B):2702–2732, July 2004.

A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, Nov. 2019.

A. Durmus, S. Majewski, and B. Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20:1–46, 2019.

A. Eberle and M. B. Majka. Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability*, 24(0):1–36, 2019.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global Nonconvex Optimization with Discretized Diffusions. In *Advances in Neural Information Processing Systems*, 2018.

A. Gelman, J. Carlin, H. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2013.

J. Gorham and L. Mackey. Measuring Sample Quality with Stein's Method. In *Advances in Neural Information Processing Systems*, 2015.

J. Gorham and L. Mackey. Measuring Sample Quality with Kernels. In *International Conference on Machine Learning*, 2017.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, Oct. 2019.

N. Gozlan. A characterization of dimension free concentration in terms of transportation inequalities. *The Annals of Probability*, 37(6):2480–2498, Nov. 2009.

F. Guo, X. Wang, K. Fan, T. Broderick, and D. Dunson. Boosting variational inference. In *Advances in Neural Information Processing Systems*, 2016.

J. M. Hernández-Lobato, Y. Li, M. Rowland, T. D. Bui, D. Hernández-Lobato, and R. E. Turner. Black-Box Alpha Divergence Minimization. In *International Conference on Machine Learning*, 2016.

M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

A. Joulin and Y. Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *The Annals of Probability*, 38(6):2418–2442, 2010.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.

B. J. K. Kleijn and A. W. van der Vaart. The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

E. Koehler, E. Brown, and S. J. P. A. Haneuse. On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *The American Statistician*, 63(2): 155–162, May 2009.

A. Kucukelbir, R. Ranganath, A. Gelman, and D. M. Blei. Automatic Variational Inference in Stan. In *Advances in Neural Information Processing Systems*, June 2015.

Y. Li and R. E. Turner. Rényi Divergence Variational Inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

F. Locatello, G. Dresdner, R. Khanna, I. Valera, and G. Rätsch. Boosting black box variational inference. In *Advances in Neural Information Processing Systems*, 2018a.

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

N. Madras and D. Sezer. Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli*, 16(3):882–908, Aug. 2010.

A. Miller, N. Foti, and R. Adams. Variational boosting: iteratively refining posterior approximations. In *International Conference on Machine Learning*, 2017.

R. M. Neal. Mcmc using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

D. Pati, A. Bhattacharya, and Y. Yang. On Statistical Optimality of Variational Bayes. In *International Conference on Artificial Intelligence and Statistics*, 2018.

R. Ranganath, S. Gerrish, and D. M. Blei. Black Box Variational Inference. In *International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

C. P. Robert. *The Bayesian Choice*. Springer, New York, NY, 1994.

D. Rudolf and N. Schweizer. Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli*, 4A: 2610–2639, 2018.

J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

S. Srivastava, C. Li, and D. B. Dunson. Scalable Bayes via barycenter in Wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

A. W. van der Vaart. *Asymptotic Statistics*. University of Cambridge, 1998.

A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto Smoothed Importance Sampling. *arXiv.org*, July 2019.

C. Villani. *Topics in Optimal Transportation*. American Mathematical Soc., 2003.

C. Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. (Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research*, 17 (159):1–48, 2016.

M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

X. Wang. Boosting variational inference: theory and examples. Master's thesis, Duke University, 2016.

Y. Wang and D. M. Blei. Frequentist Consistency of Variational Bayes. *Journal of the American Statistical Association*, 17(239):1–86, June 2018.

Y. Wang and D. M. Blei. Variational Bayes under Model Misspecification. In *Advances in Neural Information Processing Systems*, 2019.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but Did It Work?: Evaluating Variational Inference. In *International Conference on Machine Learning*, 2018.