

2022-03-29

Robust, automated, and accurate black-box variational inference

M. Welandawe, M. Riis Anderson, A. Vehtari, J. Huggins. 2022. "Robust, Automated, and Accurate Black-box Variational Inference" <https://doi.org/10.48550/arXiv.2203.15945>
<https://hdl.handle.net/2144/46451>

Downloaded from DSpace Repository, DSpace Institution's institutional repository

Robust, Automated, and Accurate Black-box Variational Inference

Manushi Welandawe¹, Michael Riis Andersen²,
Aki Vehtari³, and Jonathan H. Huggins¹

¹*Department of Mathematics & Statistics, Boston University, e-mail: manushiw@bu.edu ; huggins@bu.edu*

²*DTU Compute, Technical University of Denmark, e-mail: michael.riis@gmail.com*

³*Department of Computer Science, Aalto University, e-mail: aki.vehtari@aalto.fi*

Abstract: Black-box variational inference (BBVI) now sees widespread use in machine learning and statistics as a fast yet flexible alternative to Markov chain Monte Carlo methods for approximate Bayesian inference. However, stochastic optimization methods for BBVI remain unreliable and require substantial expertise and hand-tuning to apply effectively. In this paper, we propose *Robust, Automated, and Accurate BBVI* (RAABBVI), a framework for reliable BBVI optimization. RAABBVI is based on rigorously justified automation techniques, includes just a small number of intuitive tuning parameters, and detects inaccurate estimates of the optimal variational approximation. RAABBVI adaptively decreases the learning rate by detecting convergence of the fixed-learning-rate iterates, then estimates the symmetrized Kullback–Leiber (KL) divergence between the current variational approximation and the optimal one. It also employs a novel optimization termination criterion that enables the user to balance desired accuracy against computational cost by comparing (i) the predicted relative decrease in the symmetrized KL divergence if a smaller learning were used and (ii) the predicted computation required to converge with the smaller learning rate. We validate the robustness and accuracy of RAABBVI through carefully designed simulation studies and on a diverse set of real-world model and data examples.

1. Introduction

A core strength of the Bayesian approach is that it is conceptually straightforward to carry out inference in *arbitrary* models, which enables the user to employ whatever model is most appropriate for the problem at hand. The flexibility and uncertainty quantification provided by Bayesian inference have led to its widespread use in statistics [23, 50] and machine learning [3, 42], including in deep learning [e.g., 8, 21, 29, 33, 38, 49, 53]. Using Bayesian methods in practice, however, typically requires using approximate inference algorithms to estimate quantities of interest such as posterior functionals (e.g., means, covariances, predictive distributions, and tail probabilities) and measures of model fit (e.g., marginal likelihoods and cross-validated predictive accuracy). Therefore, a core challenge in modern Bayesian statistics is the development of *general-purpose* (or *black-box*) algorithms that can accurately approximate these quantities for whatever model the user chooses. In machine learning, rather than using MCMC, *black-box variational inference* (BBVI) has become the method of choice because of its scalability and wide-applicability [4, 8, 33, 49, 61]. BBVI is implemented in many software packages for general-purpose inference such as Stan, Pyro, PyMC3, and TensorFlow Probability, which have seen widespread adoption by applied data analysts, statisticians, and data scientists.

To ensure reliability and wide applicability, a high-quality black-box approximate inference framework should be (1) accurate, (2) automated, (3) intuitively adjustable by the

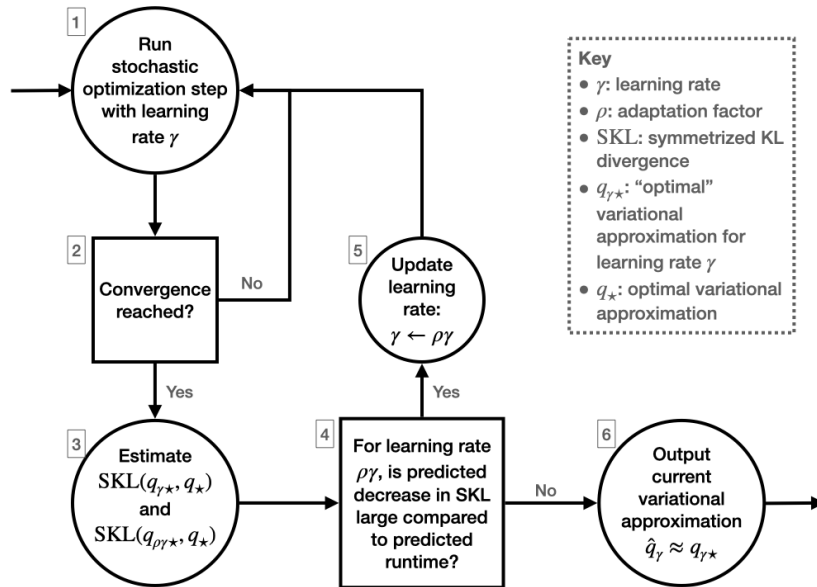


Fig 1: Schematic of the high-level logic of our proposed *Robust, Automated, and Accurate BBVI* (RAABBVI) algorithm.

user, and (4) robust to failure and tuning parameter selection. While some recent progress has been made in developing tools for assessing the accuracy of variational approximations [28, 65], stochastic optimization methods for BBVI remain unreliable and require substantial hand-tuning of the number of iterations and optimizer tuning parameters. Moreover, there are few tools available for determining whether the variational parameters estimated by these frameworks are close to optimal in any meaningful sense and, if not, how to address the problem – More iterations? A different learning rate schedule? A smaller initial or final learning rate? The work of Agrawal, Sheldon and Domke [1] exhibits the lack of a reliable, coherent BBVI optimization methodology: the authors synthesize and compare recent advances such as normalizing flows and gradient estimators, must rely on running the optimization for 30,000 iterations for 5 different step sizes for all 30 benchmarked models, despite these models varying greatly in the complexity and dimensionality of the posteriors.

In this paper we aim to provide a practical, cohesive, and theoretically well-grounded, optimization framework for BBVI. Our approach builds on a recent line of work inspired by Pflug [46], which uses a fixed learning rate γ that is adaptively decreased by a multiplicative factor ρ once the optimization iterates, which form a homogenous Markov chain, have converged [9, 10, 13, 45, 64, 66]. A benefit of this approach is that, for a given learning rate, a dramatically more accurate estimate of the optimal variational parameter can be obtained by using *iterate averaging* [16, 47, 52]. However, as we have shown in previous work, existing convergence checks can be unreliable and stop too early [13]. Since the learning rate is decreased by a constant multiplicative factor, decreasing it too early can slow down the optimization by an order of magnitude or more. Hence, it is crucial to develop methods that do not prematurely declare convergence. On the other hand, an optimization framework

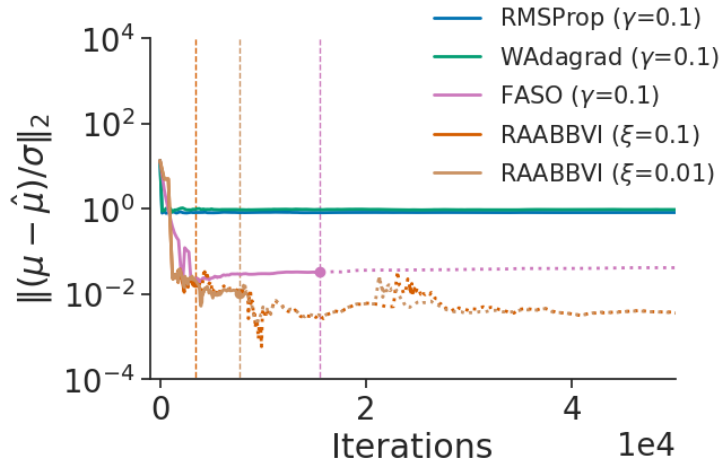


Fig 2: Accuracy comparison of variational inference algorithms using the *dogs* dataset/model from the `posteriordb` package (see Section 7.2 for details). Accuracy is measured in terms of relative mean error $\|(\mu - \hat{\mu})/\sigma\|_2$, where μ and σ are, respectively, the posterior mean and standard deviation vectors and $\hat{\mu}$ is the variational approximation to μ . The vertical lines indicate the termination points for RAABBVI and for *fixed-learning-rate automated stochastic optimization* (FASO), which is the same as RAABBVI except it always terminates (moves to step 6) after step 4. The fixed-learning rate algorithms has learning rate γ and RAABBVI has a user-specified accuracy threshold ξ .

must also provide a termination criterion that triggers when it is no longer worthwhile to decrease the learning rate further, either because the current variational approximation is sufficiently accurate or because further optimization would be too time-consuming.

The key idea that informs our solutions to these challenges is that we want q_{γ^*} , the target variational approximation for learning rate γ , to be close to the optimal variational approximation q_* . We measure closeness in terms of *symmetrized Kullback–Leibler (KL) divergence* $\text{SKL}(q_*, q_{\gamma^*})$ and show that closeness in symmetrized KL divergence can be translated into bounds on other widely used accuracy metrics like Wasserstein distance [5, 28]. Figure 1 summarizes our proposed algorithm, which we call *Robust, Automated, and Accurate BBVI* (RAABBVI). The primary contributions of this paper are in steps 2, 3, and 4. In step 2, to determine convergence at a fixed step size, we build upon our approach in Dhaka et al. [13], where we establish that the scale-reduction factor \hat{R} [22, 23, 58], which is widely used to determine convergence of Markov chain Monte Carlo algorithms, can be combined with a Monte Carlo standard error (MCSE) [24, 58] cutoff to construct a convergence criteria. We have improved upon our previous proposal by:

- (a) adaptively finding the size of the convergence window, which may need to be large for challenging or high-dimensional distributions over the model parameters, and
- (b) developing a new rigorous MCSE cutoff condition that guarantees the symmetrized KL divergence between $q_{\rho\gamma^*}$ and the estimate of $q_{\rho\gamma^*}$ obtained via iterate averaging will be small.

In step 3, we leverage recent results that characterize the bias of the stationary distribution of SGD with a fixed learning rate [16] to estimate $\text{SKL}(q_*, q_{\gamma*})$ and $\text{SKL}(q_*, q_{\rho\gamma*})$ without access to q_* . These estimates enable, in step 4, the use of a termination criterion that compares (i) the predicted relative decrease in the KL divergence if the smaller learning rate $\rho\gamma$ were used and (ii) the predicted computation required to converge with the learning rate $\rho\gamma$. By trading off between (i) and (ii), the criterion enables the user to balance desired accuracy against computational cost. Figure 2 provides an example of the high accuracy and reliability achievable using RAABBVI compared to alternative optimization algorithms and demonstrates how the user can trade off accuracy and computation by adjusting the accuracy threshold ξ .

In summary, by drawing on recent developments in theory and methods for fixed-learning-rate stochastic optimization, tools from MCMC methodology and results from functional analysis, RAABBVI delivers a number of benefits:

- it relies on rigorously justified automation techniques, including automatic learning rate adaptation;
- it has an interpretable, user-adjustable accuracy parameter along with a small number of additional intuitive tuning parameters;
- it detects inaccurate estimates of the optimal variational approximation; and
- it can flexibly incorporate additional or future methodological improvements related to variational inference and stochastic optimization.

We demonstrate through synthetic comparisons and real-world model and data examples that RAABBVI provides high-quality black-box approximate inferences. We make RAABBVI available as part of the open source Python package VIABEL.¹

2. Preliminaries and Background

2.1. Bayesian inference

Let $\theta \in \mathbb{R}^d$ denote a parameter vector of interest, and let x denote observed data. A Bayesian model consists of a prior density $\pi_0(d\theta)$ and a likelihood $\ell(x; \theta)$. Together, the prior and likelihood define a joint distribution over the data and parameters. The Bayesian posterior distribution π is the conditional distribution of θ given fixed data x , with x suppressed in the notation since it is always fixed throughout this work. To write this conditional, we define the unnormalized posterior density $\pi^u(\theta) := \ell(x; \theta)\pi_0(d\theta)$ and the marginal likelihood, or evidence, $Z := \int \pi^u(d\theta)$. Then the posterior is $\pi := \pi^u/Z$. Typically, practitioners report *posterior summaries*, such as point estimates and uncertainties, rather than the full posterior. For $\vartheta \sim \pi$, typical summaries include the mean $m_\pi := \mathbb{E}(\vartheta)$, the covariance $\Sigma_\pi := \mathbb{E}\{(\vartheta - m_\pi)(\vartheta - m_\pi)^\top\}$, and $[a, b]$ interval probability $I_{\pi, i, a, b} := \mathbb{P}(\vartheta_i \in [a, b]) = \mathbb{E}\{\mathbb{1}(\vartheta_i \in [a, b])\}$, where $\mathbb{1}(C)$ is equal to one when C is true and zero otherwise.

2.2. Variational inference

In most Bayesian models, it is infeasible to efficiently compute quantities of interest such as posterior means, variances, and quantiles. Therefore, one must use an approximate inference method, which produces an approximation q to the posterior π . The summaries of

¹<https://github.com/jhuggins/viabel>

q 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 $\mathcal{Q} = \{q_\lambda : \lambda \in \mathbb{R}^m\}$ of approximating distributions [4, 61]:

$$q_{\lambda_*} = \arg \min_{q_\lambda \in \mathcal{Q}} \mathcal{D}_\pi(q_\lambda).$$

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

In variational inference, the standard choice for the discrepancy $\mathcal{D}_\pi(\cdot)$ is the *Kullback–Leibler (KL) divergence* $\text{KL}(q \mid \pi) := \int \log(dq/d\pi) dq$ [3]. Note that the KL divergence is asymmetric in its arguments. The direction $\mathcal{D}_\pi(q) = \text{KL}(q \mid \pi)$ is most typical in variational inference, largely out of convenience; the unknown marginal likelihood Z appears in an additive constant that does not influence the optimization, and computing gradients requires estimating expectations only with respect to $q \in \mathcal{Q}$, which is chosen to be tractable. Minimizing $\text{KL}(q \mid \pi)$ is equivalent to maximizing the *evidence lower bound* [ELBO; 3]:

$$\text{ELBO}(q) := \int \log\left(\frac{d\pi^u}{dq}\right) dq.$$

While numerous other divergences have been used in the literature [e.g., 7, 15, 26, 36, 62, 63], we focus on $\text{KL}(q \mid \pi)$ since it is the most widely used, the default or only choice in widely used frameworks such as Stan, Pyro, and PyMC3, and easiest to estimate when using simple Monte Carlo sampling to approximate the gradient [14].

2.3. Black-box variational inference

Black-box variational inference (BBVI) methods have greatly extended the applicability of variational inference by removing the need for model-specific derivations [12, 34, 40, 48, 56] and enabling the use of more flexible approximation families [33, 44, 54]. This flexibility is a result of using simple Monte Carlo (and automatic differentiation) to approximate the (gradients of the) expectations that define common choices of the discrepancy objective [40, 44]. To estimate the optimal variational parameter λ_* , BBVI methods commonly uses stochastic optimization schemes of the form

$$\lambda^{(k+1)} \leftarrow \lambda^{(k)} - \gamma^{(k)} d^{(k)}, \quad (1)$$

where $d^{(k)} \in \mathbb{R}^m$ is the descent direction and $\gamma^{(k)} > 0$ is the learning rate (also called the “step size”) at iteration k . Standard stochastic gradient descent corresponds to taking $d^{(k)} = \hat{g}^{(k)}$, a (usually unbiased) stochastic estimate of the gradient $g(\lambda^{(k)}) := \nabla_{\lambda^{(k)}} \mathcal{D}_\pi(q_{\lambda^{(k)}})$. We are particularly interested in adaptive stochastic optimization methods [e.g., 17, 27, 32] that use a smoothed and/or rescaled version of $\hat{g}^{(k)}$ as the descent direction. For example, RMSProp [27] tracks an exponential moving average of the squared gradient, $\nu^{(k+1)} = \beta \nu^{(k)} + (1 - \beta) \hat{g}^{(k)} \odot \hat{g}^{(k)}$, which is used to rescale the current stochastic gradient: $d^{(k)} = \hat{g}^{(k+1)} / \sqrt{\nu^{(k)}}$. Mean while, Adam [32] tracks an exponential moving average of the gradient $m^{(k+1)} = \beta_1 m^{(k)} + (1 - \beta_1) \hat{g}^{(k)}$ as well as the squared gradient $\nu^{(k+1)} = \beta_2 \nu^{(k)} + (1 - \beta_2) \hat{g}^{(k)} \odot \hat{g}^{(k)}$ and used both to rescale the current stochastic gradient: $d^{(k)} = m^{(k)} / \sqrt{\nu^{(k)}}$. The benefits of adaptive algorithms include that they tend to be more stable and are scale invariant, so the learning rate can be set in a problem-independent manner.

2.4. Fixed-learning-rate stochastic optimization

If the learning rate is fixed so that $\gamma^{(k)} = \gamma$, then we can view the iterates $\lambda^{(1)}, \lambda^{(2)}, \dots$ produced by Eq. (1) as a homogenous Markov chain, which under certain conditions will have a stationary distribution μ_γ [16, 25]. Stochastic optimization with fixed learning rate exhibits two distinct phases: a transient (a.k.a. warm-up) phase during which iterates make rapid progress towards the optimum, followed by a stationary (a.k.a. mixing) phase during which iterates oscillate around the mean of the stationary distribution, $\bar{\lambda}_\gamma := \int \lambda \mu_\gamma(d\lambda)$.

The mean $\bar{\lambda}_\gamma$ is a natural target because even if the variance of the individual iterates $\lambda^{(k)}$ means they are far from λ_* , $\bar{\lambda}_\gamma$ can be a much more accurate approximation to λ_* . For example, the following result quantifies the bias of standard fixed-learning-rate SGD (Gitman et al. [25] provide similar results for momentum-based SGD algorithms):

Theorem 2.1 (Dieuleveut, Durmus and Bach [16, Theorem 4]). *Under regularity conditions on the objective function and the unbiased stochastic gradients, there exist constant vectors $A, B \in \mathbb{R}^m$ such that²*

$$\bar{\lambda}_\gamma - \lambda_* = A\gamma + B\gamma^2 + o(\gamma^2) \quad (2)$$

and a matrix $A' \in \mathbb{R}^{m \times m}$ such that

$$\int (\lambda - \lambda_*)(\lambda - \lambda_*)^\top \mu_\gamma(d\lambda) = A'\gamma + O(\gamma^2).$$

In other words, at stationarity, a single iterate will satisfy $\lambda^{(k)} - \lambda_* = O(\gamma^{1/2})$ (with high probability) while its expectation will satisfy $\bar{\lambda}_\gamma - \lambda_* = O(\gamma)$. Therefore, when the learning rate is small, $\bar{\lambda}_\gamma$ is substantially better than $\lambda^{(k)}$ as an estimator for λ_* .

In practice the *iterate average* (i.e., sample mean)

$$\hat{\lambda}_\gamma := \frac{1}{k^{\text{avg}}} \sum_{k=0}^{k^{\text{avg}}-1} \lambda^{(k^{\text{conv}}+k)} \quad (3)$$

provides an estimate of $\bar{\lambda}_\gamma$, where k^{conv} is the iteration at which the optimization has reached the stationary phase and k^{avg} is the number of iterations used to compute the average. Using $\hat{\lambda}_\gamma$ as an estimate of λ_* is known as Polyak–Ruppert averaging [2, 47, 52].

When using iterate averaging, it is crucial to ensure the iterate average accurately approximates $\bar{\lambda}_\gamma$. Considering a stationary Markov chain, we can compute the Monte Carlo estimate of the mean of the Markov Chain at stationarity. Then the notion of *effective sample size* (ESS) aids in quantifying the accuracy of this Monte Carlo estimate. Further, the effective sample size can also be used to define the *Monte Carlo standard error* (MCSE) when the Markov chain satisfies a central limit theorem. We can efficiently estimate the ESS and also approximate the MCSE (see Appendix B.1 for details). We denote the estimates of ESS and MCSE using iterates $\lambda_i^{(k^{\text{conv}}:k)}$ as $\widehat{\text{ESS}}(\lambda_i^{(k^{\text{conv}}:k)})$ and $\widehat{\text{MCSE}}(\lambda_i^{(k^{\text{conv}}:k)})$ respectively. Dhaka et al. [13] use the conditions $m^{-1} \sum_{i=1}^m \widehat{\text{MCSE}}(\lambda_i^{(k^{\text{conv}}:k)}) < 0.02$ and $\widehat{\text{ESS}}(\lambda_i^{(k^{\text{conv}}:k)}) > 20$ to determine when to stop iterate averaging. However, no rigorous justification is given for the MCSE threshold.

²As stated in Dieuleveut, Durmus and Bach [16], the $B\gamma^2$ term is written as $O(\gamma^2)$. However, the fact that this term is of the form $B\gamma^2 + o(\gamma^2)$ can be extracted from the proof.

2.5. Automatically scheduling learning rate decreases

A benefit of using a fixed learning rate is that it can be adaptively and automatically decreased once the iterates reach stationarity. For example, if the initial learning rate is γ_0 , after reaching stationarity the learning rate can decrease to $\gamma_1 := \rho\gamma_0$, where $\rho \in (0, 1)$ is a user-specified adaptation factor. The process can be repeated: when stationarity is reached at learning rate γ_t , the learning rate can decrease to $\gamma_{t+1} := \rho\gamma_t$. In this way the learning rate is not decreased too early (when the iterates are still making fast progress toward the optimum) or too late (when the accuracy of the iterates is no longer improving). Compare this adaptive approach to the canonical one of setting a schedule such as $\gamma^{(k)} = \Delta/(\text{O}+k)^\square$, which requires the choice of three tuning parameters. These tuning parameters can have a dramatic effect on the speed of convergence, particularly when $\lambda^{(0)}$ is far from λ_* .

The question of how to determine when the stationary phase has been reached has a long history with recent renewed attention [9, 10, 13, 35, 45, 46, 64, 66]. One line of work [35, 64, 66] is based on finding an invariant function that has expectation zero under the stationary distribution of the iterates, then using a test for whether the empirical mean of the invariant function is sufficiently close to zero. An alternative approach developed in Dhaka et al. [13] makes use of the potential scale reduction factor \widehat{R} , perhaps the most widely used MCMC diagnostic for detecting stationarity [22, 23, 58]. The standard approach to computing \widehat{R} is to use multiple Markov chains. If we have J chains and K iterates in each chain, then $\widehat{R} := (\widehat{V}/\widehat{W})^{1/2}$, where \widehat{V} and \widehat{W} are estimates of, respectively, the between-chain and within-chain variances. In the split- \widehat{R} version, each chain is split into two before carrying out the computation above, which helps with detecting non-stationarity [23, 58] and allows for use even when $J = 1$. Let $\widehat{R}_i(W)$ denote the split- \widehat{R} value computed from $\lambda_i^{(k-W+1)}, \dots, \lambda_i^{(k)}$, the i th dimension of the last W iterates. Dhaka et al. [13] uses the stationarity condition $\max_i \widehat{R}_i(100) < 1.1$,

3. Methodological Approach

We now summarize our approach to designing a robust, automatic, and accurate optimization framework for BBVI.

Robustness. A robust method should not be too sensitive to the choice of tuning parameters. It should also work well on a wide range of “typical” problems. To achieve this we design adaptively schemes for setting parameters such as the window size for detecting convergence that are problem-dependent.

Automation. An automatic method should require minimal input from the user. Any inputs that are required should be clearly necessary (e.g., the model and the data) or be intuitive to an applied user who is not an expert in variational inference and optimization. Therefore, we require the parameters of any adaptation scheme to either be intuitive or not require adjustment by the user. Examples of intuitive parameters include maximum number of iterations, maximum runtime, and, when defined appropriately, accuracy.

Accuracy. Since we are focused on the optimization perspective, accuracy corresponds to being close to the optimum. So, for variational inference, the key question is how to measure the accuracy of the variational approximation returned by the optimization algorithm, \hat{q}_* ,

as an estimate of the optimal variational approximation $q_* := q_{\lambda_*}$. But the answer to this question depends upon choosing an appropriate measure of the discrepancy between q_* and the posterior π . Based on the discussion in Section 2.1, the goal should be for quantities such as m_{q_*} , Σ_{q_*} , and $I_{q_*,i,a,b}$ to be close to, respectively, m_π , Σ_π , and $I_{\pi,i,a,b}$. The interval probabilities are already on an interpretable scale, so ensuring that $|I_{q_*,i,a,b} - I_{\pi,i,a,b}|$ is much less than 1 is an intuitive notion of accuracy. Since $\Sigma_\pi^{1/2}$ establishes the relevant scale of the problem for means and standard deviations, so it is appropriate to ensure that $\|\Sigma_\pi^{-1/2}(m_\pi - m_{q_*})\|_2$ and $\|\Sigma_\pi^{-1}(\Sigma_\pi - \Sigma_{q_*})\|_2 = \|I - \Sigma_\pi^{-1}\Sigma_{q_*}\|_2$ are much less than 1.

We apply our approach to robustness, accuracy, and automation to design the two core components of a BBVI stochastic optimization framework with automated learning rate scheduling:

1. A **termination rule** for stopping the optimization once the final approximation is close to the optimal approximation (Section 4).
2. A **learning rate scheduler**, which must detect stationarity and determine how many iterates to average before decreasing the learning rate (Section 5).

4. Termination Rule

The development of our termination rule will proceed in three steps. First, we will select an appropriate discrepancy measure between distributions. Next, we will design an idealized termination rule based on this discrepancy measure. Finally, we will develop an implementable version of the idealized termination rule that satisfies our criteria for being robust, automatic, and accurate.

4.1. Choice of Accuracy Measure

While we want to choose a discrepancy measure that guarantees the accuracy of mean, covariance, and interval probabilities, ideally it would also guarantee other plausible expectations of interest (e.g., predictive densities) are accurately approximated. The *Wasserstein distance* provides one convenient metric for accomplishing this goal, and is widely used in the analysis of MCMC algorithms and in large-scale data asymptotics [e.g., 18–20, 30, 39, 51, 60]. For $p \geq 1$ and a positive-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, we define the (p, Σ) -*Wasserstein distance* between distributions η and ζ as

$$\mathcal{W}_{p,\Sigma}(\eta, \zeta) := \inf_{\omega} \left\{ \int \|\Sigma^{-1/2}(\theta - \theta')\|_2^p \omega(d\theta, d\theta') \right\}^{1/p},$$

where the infimum is over the set of *couplings* between η and ζ ; that is, Borel measures ω on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\eta = \omega(\cdot, \mathbb{R}^d)$ and $\zeta = \omega(\mathbb{R}^d, \cdot)$ [59, Defs. 6.1 & 1.1]. Small (p, Σ) -Wasserstein distance implies many functionals of the two distributions are close relative to the scale determined by $\Sigma^{1/2}$.

Specifically, we have the following result, which is an immediate corollary of Huggins et al. [28, Theorem 3.4].

Proposition 4.1. *If $\mathcal{W}_{p,\Sigma}(\eta, \zeta) \leq \varepsilon$ for any $p \geq 1$, then*

$$\|\Sigma^{-1/2}(m_\eta - m_\zeta)\|_2 \leq \varepsilon$$

If $\mathcal{W}_{2,\Sigma}(\eta, \zeta) \leq \varepsilon$, then, for $\varrho := \min\{\|\Sigma^{-1}\Sigma_\eta\|_2^{1/2}, \|\Sigma^{-1}\Sigma_\zeta\|_2^{1/2}\}$,

$$\|\Sigma^{-1}(\Sigma_\eta - \Sigma_\zeta)\|_2 < 2\varepsilon(\varrho + \varepsilon).$$

More generally, small (p, Σ) -Wasserstein distance for any $p \geq 1$ guarantees the accuracy of expectations for any function f with small Lipschitz constant with respect to the metric $d_\Sigma(\theta, \theta') := \|\Sigma^{-1/2}(\theta - \theta')\|_2$ – that is, when $\sup_{\theta \neq \theta'} |f(\theta) - f(\theta')|/d_\Sigma(\theta, \theta')$ is small.

While the Wasserstein distance controls the error in mean and covariance estimates, it does not provide strong control on interval probability estimates. The KL divergence, however, does, since for distributions η and ζ , $|I_{\eta,i,a,b} - I_{\zeta,i,a,b}| \leq \sqrt{\text{KL}(\eta | \zeta)}/2$ for all $a < b$ and i (see Appendix B.2). As we show next, in many scenarios we can bound the Wasserstein distance by the KL divergence and therefore enjoy the benefits of both. Our result is based on the following definition, which makes the notion of the scale of a distribution precise:

Definition 4.2. For $p \geq 1$ and a positive-definite matrix $\Sigma \in \mathbb{R}^{d \times d}$, the distribution η is said to be (p, Σ) -exponentially controlled if

$$\inf_{\theta'} \log \int e^{\|\Sigma^{-1/2}(\theta - \theta')\|_2^p} \eta(d\theta) \leq d/2. \quad (4)$$

Specially, $\Sigma^{1/2}$ establishes the appropriate scale for uncertainty with respect to η . For example, if $\eta = \mathcal{N}(m, V)$, then it is a straightforward exercise to confirm that η is $(2, 1.78^2 V)$ -exponentially controlled.

The following result establishes the relevant link between the KL divergence and the Wasserstein distance via Definition 4.2.

Proposition 4.3. If η is (p, Σ) -exponentially controlled, then for all ζ absolutely continuous with respect to η ,

$$\mathcal{W}_{p,\Sigma}(\eta, \zeta) \leq (3 + d)\mathcal{K}_p(\eta | \zeta),$$

where $\mathcal{K}_p(\eta | \zeta) := \text{KL}(\zeta | \eta)^{\frac{1}{p}} + \{\text{KL}(\zeta | \eta)/2\}^{\frac{1}{2p}}$.

Proof. The result follows from Bolley and Villani [5, Corollary 2.3] after the change-of-variable $\theta \mapsto \Sigma^{-1/2}\theta$ and using the fact that the KL divergence is invariant under diffeomorphisms, then applying Eq. (4). \square

If ζ and η could operate over different scales, then we can use the *symmetrized KL divergence* $\text{SKL}(\zeta, \eta) := \text{KL}(\zeta | \eta) + \text{KL}(\eta | \zeta)$. Indeed, it follows from Proposition 4.3 that if $\text{SKL}(\zeta, \eta)$ is small, then the (p, Σ) -Wasserstein distance is small whenever *either* η or ζ is (p, Σ) -exponentially controlled.

4.2. An Idealized Termination Rule

Based on our developments in Section 4.1, we will define our termination rule in terms of the symmetrized KL divergence. Recall that q_* denotes the optimal variational approximation to π and \hat{q}_* denotes an estimate of q_* . Since total variation and $(1, \Sigma)$ -Wasserstein distances are controlled by the square root of the KL divergence, we focused on the square root of the symmetrized KL divergence. For the current learning rate $\gamma > 0$, let $q_{\gamma*} := q_{\bar{\lambda}_\gamma}$ denote the target γ -learning-rate variational approximation. The termination rule we propose is based

on the trade-off between the improved accuracy of the approximation if the learning rate were reduced to $\rho\gamma$ and the time required to reach that improved accuracy. To quantify the improved accuracy, we introduce a user-chosen target accuracy target ξ for $\text{SKL}(q_*, \hat{q}_*)^{1/2}$. If the user expects $\text{KL}(\pi | q_*)$ to be large, then setting ξ to a moderate value such as 1 or 10 could give acceptable performance. If the user expects $\text{KL}(\pi | q_*)$ to be small, then setting ξ to a value such as 0.1 or 0.01 might be more appropriate. Using ξ , define the *relative SKL improvement* RSKL as

$$\text{RSKL} := \frac{\text{SKL}(q_*, q_{\rho\gamma_*})^{1/2} + \xi}{\text{SKL}(q_*, q_{\gamma_*})^{1/2}},$$

where the first term measures the relative improvement of the approximation if the learning rate were reduced and the second term measures the current accuracy relative to the desired accuracy. To quantify the time to obtain the relative accuracy improvement, we use the number of iterations to reach convergence for the fixed learning rate $\rho\gamma$. Letting K_{γ_*} denote the number of iterations required to reach the target γ -learning-rate variational approximation, we define the relative iteration increase

$$\text{RI} := \frac{K_{\rho\gamma_*}}{K_{\gamma_*} + K_0},$$

where K_0 denotes a number of iterations the user would consider “small”. Combining RSKL and RI, we obtain the *inefficiency index* $\mathcal{I} = \text{RSKL} \times \text{RI}$, the relative improvement in accuracy times the relative increase in runtime. Our *idealized SKL inefficiency termination rule* triggers when $\mathcal{I} > \tau$, where τ is a user-specified inefficiency threshold which allows the user to trade off accuracy with computation, but only up to the point where $\text{SKL}(q_*, q_{\gamma_*})^{1/2} \approx \xi$.

4.3. An Implementable Termination Rule

The idealized SKL inefficiency termination rule cannot be directly implemented since q_* is unknown – and if it were known, it would be unnecessary to run the optimization algorithm. However, we will show that it is possible to obtain a good estimate of the symmetrized KL divergence between the approximation obtained with a given learning rate γ' and the optimal approximation without access to q_* .

Recall that q_{γ_*} denotes the target γ -learning-rate variational approximation. With a slight abuse of notation, we let the optimal zero-learning-rate approximation refer to the optimal approximation: $q_{0_*} := q_*$. Our approach is motivated by Theorem 2.1 and in particular the form of the bias $\bar{\lambda}_\gamma - \lambda_*$ in Eq. (2). Our analysis focuses on the still-common setting when \mathcal{Q} is the family of mean-field Gaussian distributions, where the parameter $\lambda = (\tau, \psi) \in \mathbb{R}^{2d}$ corresponds to the distribution $q_\lambda = \mathcal{N}(\tau, \text{diag } e^{2\psi})$.

Proposition 4.4. *Let \mathcal{Q} be the family of mean-field Gaussian distributions. If Eq. (2) holds and $\gamma' = O(\gamma)$, then there is a constant $C > 0$ depending only on A and λ_* such that*

$$\text{SKL}(q_{\gamma_*}, q_{\gamma'_*}) = C(\gamma - \gamma')^2 + o(\gamma^2).$$

See Appendix A.1 for the proof. Assuming that the current learning rate is γ , then the previous learning rate was γ/ρ . Let $\delta_\gamma := \text{SKL}(q_{\gamma_*}, q_{\gamma/\rho_*})$ denote the symmetrized

KL divergence between the optimal variational approximations obtained at each of these learning rates. In principle we can use Proposition 4.4 to estimate C by

$$\hat{C} = \delta_\gamma \rho^2 / \{\gamma^2(1 - \rho)^2\},$$

and then estimate that

$$\text{SKL}(q_{\gamma_*}, q_*) \approx \hat{C} \gamma^2 = \delta_\gamma \rho^2 / (1 - \rho)^2.$$

There are, however, two problems with the tentative approach just outlined. The first problem is that Theorem 2.1 only holds for standard SGD; however, adaptive SGD algorithms are widely used in practice. Indeed, we observe empirically that for RMSProp $\text{SKL}(q_{\gamma_*}, q_{\gamma'^*}) = \Theta(|\gamma - \gamma'|)$ (see Fig. C.6). Our solution is to modify an adaptive gradient method to behave asymptotically like SGD. In the cases of RMSProp and Adam, we propose Averaged RMSProp (avgRMSProp) and Averaged Adam (avgAdam), where

$$\nu^{(k+1)} = \beta_k \nu^{(k)} + (1 - \beta_k) \hat{g}^{(k)} \odot \hat{g}^{(k)},$$

for $\beta_k = 1 - 1/k$. Hence, $\nu^{(k+1)} = (k+1)^{-1} \sum_{k'=0}^k \hat{g}^{(k')} \odot \hat{g}^{(k')}$ is the averaged squared gradient over all iterations [41, §4]. As long as the SGD Markov chain is ergodic and $\mathbb{E}[\hat{g}^{(k)} \odot \hat{g}^{(k)}] < \infty$ at stationarity, $\nu^{(k)}$ converges almost surely to a constant and hence the SGD bias analysis also applies to avgRMSProp and avgAdam.

The second problem is that Proposition 4.4 only holds for the mean-field Gaussian variational family. However, other variational families such as normalizing flows are of substantial practical interest. Therefore, we consider the weaker assumption that either Eq. (2) holds or there exist constant vectors $\Lambda, A \in \mathbb{R}^m$ and a constant $\kappa \in [1/2, 1)$ such that

$$\bar{\lambda}_\gamma = \lambda_* + \Lambda \gamma^\kappa + A \gamma + o(\gamma^{2\kappa}). \quad (5)$$

Adding the latter assumption, we have following generalization of Proposition 4.4:

Proposition 4.5. *Let \mathcal{Q} be the family of mean-field Gaussian distributions. If (i) Eq. (5) holds for some $\kappa \in [1/2, 1)$ or Eq. (2) holds (in which case let $\kappa = 1$), and (ii) $\gamma' = O(\gamma)$, then there is a constant $C > 0$ such that*

$$\text{SKL}(q_{\gamma_*}, q_{\gamma'^*}) = C \{\gamma^\kappa - (\gamma')^\kappa\}^2 + o(\gamma^{2\kappa}).$$

Moreover, C depends on only λ_* and either Λ (if $\kappa \in [1/2, 1)$) or A (if $\kappa = 1$).

See Appendix A.2 for the proof. Using Proposition 4.5 and omitting $o(\gamma^{2\kappa})$ terms, we have $\delta_\gamma \approx C \gamma^{2\kappa} (1/\rho^\kappa - 1)^2$. To improve the reliability of the estimates based on Proposition 4.5, we propose to use the symmetrized KL estimates between the variational approximations obtained at successive fixed learning rates. Let γ_0 denote the initial learning rate, so that after t learning rate decreases, the learning rate is $\gamma_t := \gamma_0 \rho^t$. Let $\delta_t := \text{SKL}(q_{\gamma_t}, q_{\gamma_{t-1}})$ and assume the current learning rate is γ_t .

Depending on the optimization algorithm, we can estimate κ (or set $\kappa = 1$ if using modified adaptive SGD algorithm with a mean-field Gaussian variation family) and C using a regression model of the form

$$\log \delta_t = \log C + 2 \log(1/\rho^\kappa - 1) + 2\kappa \log \gamma_t + \eta_t, \quad t = 1, \dots, T, \quad (6)$$

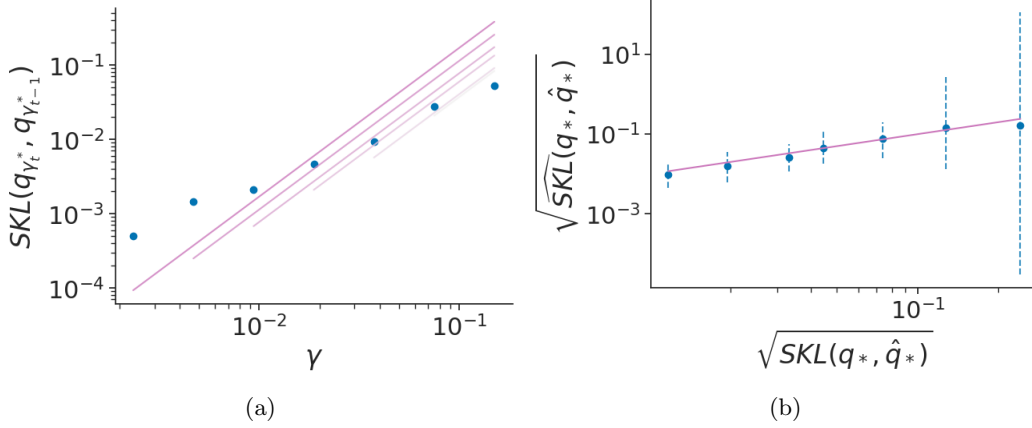


Fig 3: Results for estimating the symmetrized KL divergence with avgAdam in the case of a Gaussian distribution $\mathcal{N}(0, V)$ with $d = 100$ and $V_{ij} = j\mathbf{1}[i = j]$ (diagonal non-identity covariance). **(a)** Learning rate versus symmetrized KL divergence of adjacent iterate averaged estimates of the optimal variational distribution. The lines indicate the linear regression fits, with setting $\kappa = 1$. **(b)** Square root of true symmetrized KL divergence versus the estimated value with 95% credible interval. The uncertainty of the estimates decreases and remains well-calibrated as the learning rate decreases.

where $\eta_t \sim \mathcal{N}(0, \sigma^2)$. Given the estimate \hat{C} for C and the estimate $\hat{\kappa}$ for κ (or $\hat{\kappa} = 1$), we obtain the estimated relative SKL,

$$\widehat{\text{RSKL}} = \rho^{\hat{\kappa}} + \frac{\xi}{\hat{C}^{1/2} \gamma_t^{\hat{\kappa}}}.$$

Because we use the regression model in Eq. (6) in a low-data setting, we place priors on $\log C$ and σ :

$$\log C \sim \text{Cauchy}(0, 10), \quad \sigma \sim \text{Cauchy}^+(0, 10),$$

where Cauchy^+ is the Cauchy distribution truncated to nonnegative values. If we use adaptive SGD algorithm then we also place a prior on κ :

$$\kappa \sim \text{Unif}(0, 1),$$

for a further generalized version of Proposition 4.5. Also, because we expect early SKL estimates to be less informative about C (and κ due to the influence of $o(\gamma^{2\kappa})$ terms), we use a weighted regression with the likelihood term for (δ_t, γ_t) having weight

$$w_t = \{1 + (T - t)^2/9\}^{-1/4}. \quad (7)$$

Given above details, we use dynamic Hamiltonian Monte Carlo (HMC) algorithm and posterior mean(s) to estimate \hat{C} (and $\hat{\kappa}$). Figure 3 validates that, in the case of avgAdam, the log of the learning rate and symmetrized KL divergence have approximately a linear relationship and that our regression approach to estimating C leads to reasonable estimates of $\text{SKL}(q_{\gamma_*}, q_*)$. See Fig. C.2 for similar results with other target distributions with avgAdam.

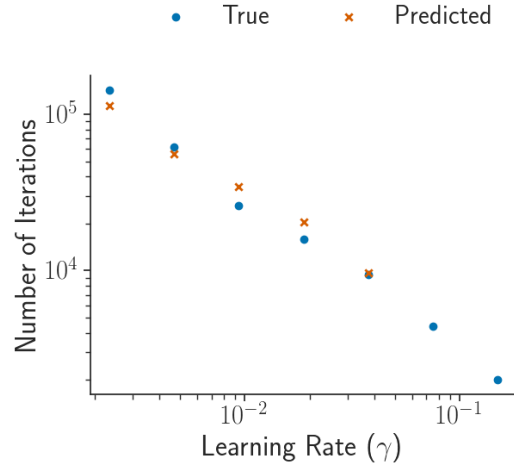


Fig 4: Results for predicting the number of iterations needed to reach convergence at each learning rate decrease in the case of Gaussian distribution $\mathcal{N}(0, V)$ with $d = 100$ and $V_{ij} = j\mathbb{1}[i = j]$ (diagonal non-identity covariance). The blue points (orange crosses) represent the true (predicted) number of iterations needed to reach convergence.

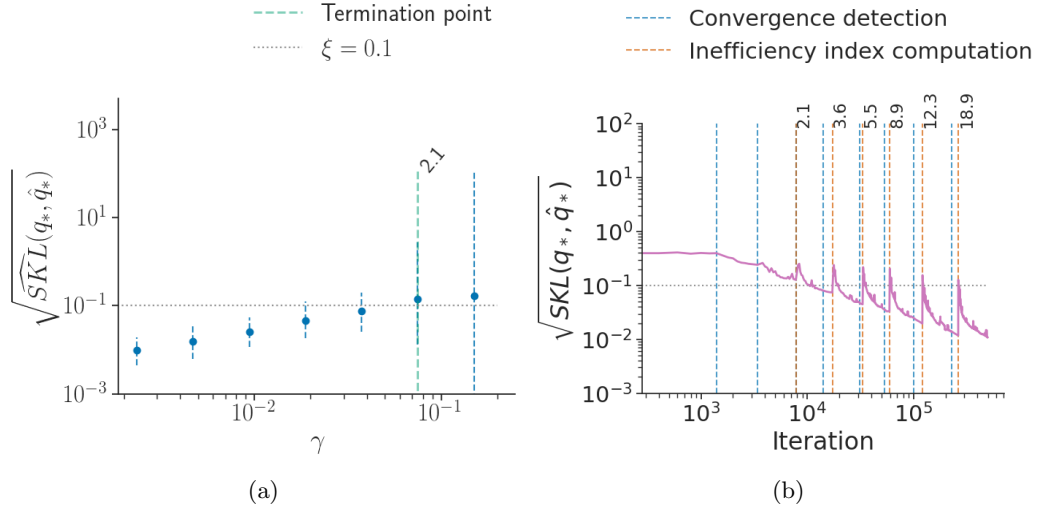


Fig 5: Results for the termination rule trigger point in the case of a Gaussian distribution $\mathcal{N}(0, V)$ with $d = 100$ and $V_{ij} = j\mathbb{1}[i = j]$ (diagonal non-identity covariance). **(a)** Learning rate versus square root of estimated symmetrized KL divergence with 95% credible interval (dashed blue line). The green vertical line indicates the termination rule trigger point with corresponding $\hat{\mathcal{I}}$ value. **(b)** Iterations versus square root of symmetrized KL divergence between iterate average and optimal variational approximation. The vertical lines indicate the convergence detection points using \hat{R} (blue) and inefficiency index computation ($\hat{\mathcal{I}}$) points (orange) with corresponding values.

To estimate RI we need to estimate the number of iterations to reach convergence at the next learning rate γ_{t+1} . It is reasonable to assume that there is an exponential growth in the convergence iterations as the learning rate decreases since stochastic gradients take time to converge at a polynomial rate [6]. Recall that K_{γ_t} is the number of iterations to reach convergence at the current learning rate. We fit a weighted least square regression model of the form

$$\log K_{\gamma_t} = \alpha \log \gamma_t + \beta + \nu_t, \quad t = 1, \dots, T, \quad (8)$$

where $\nu_t \sim \mathcal{N}(0, \sigma_t^2)$. We then use the coefficient estimates $\hat{\alpha}$ and $\hat{\beta}$ to predict the number of iterations required for convergence at the next learning rate to be $\hat{K}_{\gamma_{t+1}} := \gamma_{t+1}^{\hat{\alpha}} e^{\hat{\beta}}$. We use the same weights given in Eq. (7) for observations of the regression model due to the non-linear behavior of the earlier convergence iterate estimates. Figure 4 demonstrates that linear relationship in Eq. (8) does in fact hold and that our weighted least square regression model predicts the number of convergence iterations K_{γ_t} quite accurately. The estimated relative iterations is then $\widehat{\text{RI}} = \hat{K}_{\gamma_{t+1}} / (K_{\gamma_t} + K_0)$.

Using the estimates $\widehat{\text{RSKL}}$ and $\widehat{\text{RI}}$ we obtain the termination rule $\widehat{\mathcal{I}} = \widehat{\text{RSKL}} \cdot \widehat{\text{RI}} > \tau$. Figure 5 shows that, when $\xi = 0.1$ (the user chosen target-accuracy), the termination rule triggers when the square root of the symmetrized KL divergence is approximately equal to ξ . Figures C.3 to C.5 shows similar results of other Gaussian targets and posterior models and datasets (see Section 7.2 for details).

5. Learning Rate Scheduler

For a fixed learning rate, computing the iterate average $\hat{\lambda}_\gamma$ defined in Eq. (3) requires determining the iteration k^{conv} at which stationarity has been reached and the number of iterations k^{avg} to use for computing the average. We address each of these in turn.

5.1. Detecting convergence to stationarity

We investigated two approaches to detecting stationarity: the SASA+ algorithm of Zhang et al. [66] and the \widehat{R} -based criterion from Dhaka et al. [13]. We made several adjustments to both approaches to reduce the number of tuning parameters and to make the remaining ones more intuitive. We found empirically that the \widehat{R} criterion outperformed the SASA+ criterion, so we describe the former here and the latter in Appendix B.3.

Let $\widehat{R}(\lambda_i^{(k-W+1)}, \dots, \lambda_i^{(k)})$ denote the split- \widehat{R} of the i th component of the last W iterates and define

$$\widehat{R}_{\max}(W) := \max_{1 \leq i \leq m} \widehat{R}(\lambda_i^{(k-W+1)}, \dots, \lambda_i^{(k)}).$$

An $\widehat{R}_{\max}(W)$ value close to 1 indicates the last W iterates are close to stationarity. In MCMC applications having $\widehat{R}_{\max}(W) \leq 1.01$ is desirable [57, 58]. Dhaka et al. [13] uses the weaker condition $\widehat{R}_{\max}(W) \leq 1.1$ since iterate averaging does not require the same level of precision as MCMC. Dhaka et al. [13] took the window size $W = 100$, but in more challenging and high-dimensional problems a fixed W is insufficient. Therefore, we instead search over window sizes between a minimum window size W_{\min} and $0.95k$ to find the one that minimizes $\widehat{R}_{\max}(W)$. The minimum window size is necessary to ensure the \widehat{R} values are reliable. We use

the upper bound $0.95k$ to always allow a small amount of “warm-up” without sacrificing more than 5% efficiency. Therefore, we estimate $W_{\text{opt}} = \arg \min_{W_{\min} \leq W \leq 0.95k} \widehat{R}_{\max}(W)$ using a grid search over 5 equally spaced values ranging from W_{\min} to $0.95k$ and require $\widehat{R}_{\max}(W_{\text{opt}}) \leq 1.1$ as the stationarity condition.

Figure 6 compares our adaptive SASA+ and adaptive \widehat{R} criteria to the Δ ELBO rule from Kucukelbir et al. [34], which is used Stan’s ADVI implementation (cf. the results of Dhaka et al. [13]). Additionally, it compares to another convergence detection approach proposed by Pesme, Dieuleveut and Flammarion [45], where they used a distance based statistics to detect the convergence, described in Appendix B.4. While adaptive SASA+, Δ ELBO, and the distance-based statistic approach sometimes trigger too early or too late, adaptive \widehat{R} consistently triggers when the full window suggests convergence has been reached. See Fig. C.1 for additional Gaussian target examples.

5.2. Determining the number of iterates for averaging

After detecting convergence to stationarity, we need to find k^{avg} large enough to ensure the iterative average is sufficiently close to the mean $\bar{\lambda}_{\gamma_t}$. But what is close enough? Building on our discussion in Section 3, we aim to ensure the error in the variational parameter estimates are small relative to the scale of uncertainty. For mean-field Gaussian distributions, the following result allows us to make such a guarantee precise.

Proposition 5.1. *Let \mathcal{Q} be the family of mean-field Gaussian distributions. Let $\hat{\lambda} = (\hat{\tau}, \hat{\psi})$ denote an approximation to $\bar{\lambda} = (\bar{\tau}, \bar{\psi})$. Define $\hat{\sigma} := \exp(\hat{\psi})$ and $\bar{\sigma} := \exp(\bar{\psi})$. If there exists $\varepsilon \in (0, 1/2)$ such that $|\hat{\tau}_i - \bar{\tau}_i| \leq \varepsilon \bar{\sigma}_i$ and $|\hat{\psi}_i - \bar{\psi}_i| \leq \varepsilon$, then*

$$\frac{|\hat{\sigma}_i - \bar{\sigma}_i|}{\bar{\sigma}_i} \leq 1.5\varepsilon \quad \text{and} \quad \frac{|\hat{\tau}_i - \bar{\tau}_i|}{\bar{\sigma}_i} \leq 1.75\varepsilon.$$

See Appendix A.4 for the proof. Based on Proposition 5.1, for mean-field Gaussian variational family we use the iterate average once the mean MCSEs $d^{-1} \sum_{i=1}^d \text{MCSE}(\hat{\tau}_{\gamma,i})/\hat{\sigma}_{\gamma,i}$ and $d^{-1} \sum_{i=1}^d \text{MCSE}(\hat{\psi}_{\gamma,i})$ are less than ε_{t-1} . For other variational families we rely on the less rigorous condition that $m^{-1} \sum_{i=1}^m \text{MCSE}(\hat{\lambda}_{\gamma,i})$ is less than ε_{t-1} . We also require the effective sample sizes of all parameters to be at least 50 to ensure the MCSE estimates are reliable.

Because the MCSE check requires computing d ESS values, for high-dimensional models the computational cost can be significant. Therefore, it is important to optimize when the checks are carried out. A well-known approach in such situations is the “doubling trick.” Let W_{conv} denote the window size when convergence is detected, and let W_{opt} denote the minimal window size that would result in the MCSE check being satisfied. The doubling trick would suggest checking at iteration numbers $k^{\text{conv}} + 2^j W_{\text{conv}}$ for $j = 0, 1, \dots$, in which case the total computational cost is within a factor of $4 \log W_{\text{opt}}$ of the optimal scenario in in which the check is only done at $k^{\text{conv}} + W_{\text{conv}}$ and $k^{\text{conv}} + W_{\text{opt}}$. However, we can potentially do substantially better by accounting for the different computational cost of the optimization versus the MCSE check.

Proposition 5.2. *Assume that the cost of the MCSE check using K iterates is $C_E K$ and the cost of K iterations of optimization is $C_O K$. Let $r := C_O/C_E$, $\chi(r) := 1 + (1+r)^{-1/2}$, and $g(r) := (2+r+2r^{1/2})/(1+r)$. If the MCSE check is done on iteration numbers $k^{\text{conv}} + \chi(r)^j W_{\text{conv}}$ for $j = 0, 1, \dots$, then the total computational cost will be within a factor of $g(r)$ of optimal.*

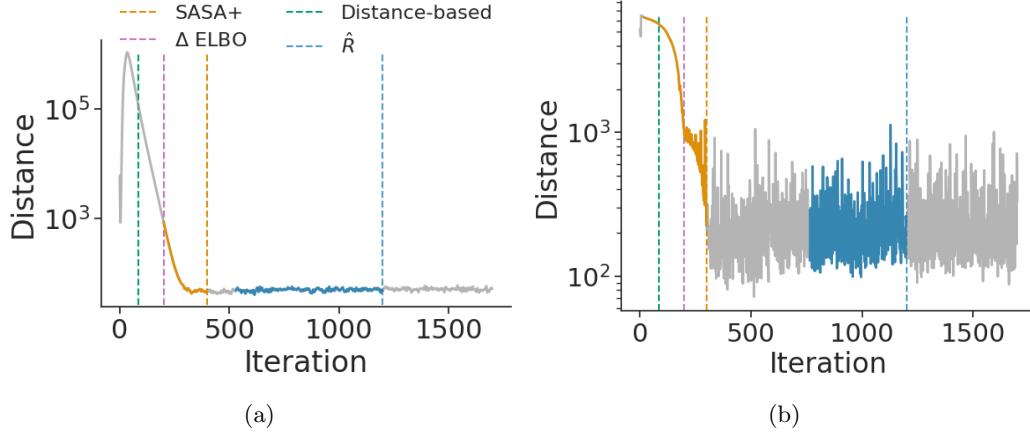


Fig 6: Iteration number versus distance between iterate average and current iterate. The vertical lines indicate convergence detection trigger points and (for SASA+ and \hat{R}) the colored portion of the accuracy values indicate they are part of the window used for convergence detection. (a) An uncorrelated Gaussian distribution $\mathcal{N}(0, V)$ with $d = 500$ and $V = I$. (b) A posteriordb dataset/model *mcycle_gp* with $d = 66$.

See Appendix A.5 for the proof. Since $g(0) = 4$ and $g(r)$ is monotonically decreasing in r , when $r \approx 0$ – that is, C_O is negligible compared to C_E – we recover the doubling rule since $\chi(0) = 2$. However, as long as r is significantly greater than zero, the worst-case additional cost factor can be substantially less than 4. Therefore we carry out the MCSE check on iteration numbers $k^{\text{conv}} + \chi(r)^j W_{\text{conv}}$ with r estimated based on the actual runtimes of the optimization so far and the first MCSE check.

6. Complete Framework

Combining our innovations from Sections 4 and 5 leads to our complete framework. When γ is fixed, our proposal from Section 5 is summarized in Algorithm 1, which we call *Fixed-learning-rate Automated Stochastic Optimization* (FASO). Combining the termination rule from Section 4 with FASO, we get our complete framework, *Robust, Automated, and Accurate Black-box Variational Inference* (RAABBVI), which we summarize in Algorithm 2. We will verify the robustness of RAABBVI through numerical experiments. Accuracy is guaranteed by the symmetrized KL divergence termination rule and the MCSE criterion for iterate averaging. RAABBVI is automatic since the user is only required to provide a target distribution and the only tuning parameter we recommend changing from their defaults are defined on interpretable, intuitive scales:

- **accuracy threshold ξ :** The symmetrized KL divergence accuracy threshold can be set based on the expected accuracy of the variational approximation. If the user expects $\text{KL}(\pi | q_*)$ to be large, then we recommend choosing $\xi \in [1, 10]$. If the user expects $\text{KL}(\pi | q_*)$ to be fairly small, then we recommend choosing $\xi \in [.01, 1]$. Our experiments suggest $\xi = 0.1$ is a good default value.

- **inefficiency threshold τ** : We recommend setting inefficiency threshold $\tau = 1$, as this weights accuracy and computation equally. A larger value (e.g., 2) could be chosen if accuracy is more important while a smaller value (e.g., 1/2) would be appropriate if computation is more of a concern.
- **maximum number of iterations K_{\max}** : The maximum number of iterations can be set by the user based on their computational budget. RAABBVI will warn the user if the maximum number of iterations is reached without convergence, so the user can either increase K_{\max} or accept the estimated level of accuracy that has been reached.

We expect the remaining tuning parameters will typically not be adjusted by the user. We summarize our recommendations:

- **initial learning rate γ_0** : When using adaptive methods such as RMSProp or Adam, the initial learning rate can essentially be set in a problem-independent manner. We use $\gamma_0 = 0.3$ in all of our experiments. If using non-adaptive methods, a line search rule such as the one proposed in Zhang et al. [66] could be used to find a good initial learning rate.
- **minimum window size W_{\min}** : We recommend taking $W_{\min} = 200$ so that that each of the split- \hat{R} values are based on at least 100 samples.
- **small iteration number K_0** : The value of K_0 should represent a number of iterations the user considers to be fairly small (that is, not requiring too much computational effort). We use $K_0 = 5W_{\min} = 1000$ for our experiments, but it could be adjusted by the user.
- **initial iterate average relative error threshold ε_0** : We recommend scaling ε_0 with ξ since more accurate iterate averages are required for sufficiently accurate symmetrized KL estimates. Therefore, we take $\varepsilon_0 = \xi$ by default.
- **adaptation factor ρ** : We recommend taking $\rho = 0.5$ because having smaller ρ value could lead to a few δ_t values to estimate C and κ and having larger ρ value will make algorithm too slow.
- **Monte Carlo samples M** : We find that $M = 10$ provides a good balance between gradient accuracy and computational burden but the performance is fairly robust to the choice of M as long as it is not too small.

7. Experiments

Unless stated otherwise, all experiments use avgAdam to compute the descent direction, mean-field Gaussian distributions as the variational family, and the tuning parameters values recommended in Section 6. We fit the regression model for C (and κ) in Stan, which resulted in extremely small computational overhead of less than 0.5%. We compare RAABBVI to FASO and fixed-learning rate versions of RMSProp and a windowed version of Adagrad (WAdagrad), which the default optimizer in PyMC3. We run all the algorithms that do not have a termination criterion for $K_{\max} = 100,000$ iterations and for the fixed-learning-rate algorithms we use learning rate $\gamma = 0.1$. We use symmetrized KL divergence as the accuracy measure when we can compute the ground-truth optimal variational approximation. Otherwise, we use relative mean error $\|(\mu - \hat{\mu})/\sigma\|_2$ and relative standard deviation error $\|\hat{\sigma}/\sigma - 1\|_2$, where μ and σ are, respectively, the posterior mean and standard deviation vectors and $\hat{\mu}$ and $\hat{\sigma}$ are the variational approximations to, respectively, μ and σ .

Algorithm 1: Fixed-learning-rate Automated Stochastic Optimization (FASO)

Input: initial variational parameter $\lambda^{(0)}$,
learning rate γ ,
minimum window size W_{\min} ,
iterate average relative error threshold ε ,
maximum iterations K_{\max}

```

1  $k^{\text{conv}} \leftarrow \text{null}$  // iteration when stationarity reached
2  $\text{success} \leftarrow \text{false}$ 
3 for  $k = 1, \dots, K_{\max}$  do
4   compute stochastic gradient  $\hat{g}^{(k)}$ 
5   compute descent direction  $d_k$ 
6   // step in descent direction
7    $\lambda^{(k+1)} \leftarrow \lambda^{(k)} - \gamma d_k$ 
8   if  $k^{\text{conv}} = \text{null}$  and  $k \bmod k_{\text{check}} = 0$  then check for convergence
9     // define window-based ESS
10     $\hat{R}_{\max}(W) := \max_{1 \leq i \leq m} \hat{R}_i(\lambda_i^{(k-W+1)}, \dots, \lambda_i^{(k)})$ 
11    // compute optimal window
12     $W_{\text{opt}} \leftarrow \arg \min_{W_{\min} \leq W \leq 0.95k} \hat{R}_{\max}(W)$ 
13    if  $\hat{R}_{\max}(W_{\text{opt}}) \leq 1.1$  then
14       $k^{\text{conv}} \leftarrow k - W_{\text{opt}}$ 
15      // window size at which to check MCSE
16       $W_{\text{check}} \leftarrow W_{\text{opt}}$ 
17       $\chi_* \leftarrow \chi(r)$  // see Prop. 5.2
18
19  if  $k^{\text{conv}} \neq \text{null}$  and  $k - k^{\text{conv}} = W_{\text{check}}$  then check for accuracy of iterate average
20     $W \leftarrow W_{\text{check}}$ 
21     $\hat{\lambda} \leftarrow W^{-1} \sum_{i=k-W+1}^k \lambda^{(i)}$ 
22     $e \leftarrow \text{MCSE}(\lambda^{(k-W)}, \dots, \lambda^{(k)})$ 
23     $\text{ESS}_{\min} \leftarrow \min_i \text{ESS}(\lambda_i^{(k-W)}, \dots, \lambda_i^{(k)})$ 
24    if mean-field Gaussian family then  $e_i \leftarrow e_i / \exp(\hat{\psi}_i)$  for  $i = 1, \dots, d$ 
25    if mean  $e_i < \varepsilon$  and  $\text{ESS}_{\min} \geq 50$  then
26       $\text{success} \leftarrow \text{true}$ 
27      break
28    else
29       $W_{\text{check}} \leftarrow \chi_* W$ 
30
31   $\hat{\lambda} \leftarrow W^{-1} \sum_{i=k-W+1}^k \lambda^{(i)}$ 
32  return  $(k, \hat{\lambda}, \text{success})$ 

```

Algorithm 2: Robust, automated, and accurate black-box variational inference (RAABBVI)

Input: initial variational parameter $\lambda^{(0)}$,
 maximum number of iterations K_{\max} ,
 initial learning rate γ_0 (default: 0.3),
 minimum window size W_{\min} (default: 200),
 accuracy threshold ξ (default: 0.1),
 inefficiency threshold τ (default: 1.1),
 iterate average error threshold ε (default: 0.1),
 adaptation factor ρ (default: 0.5)
 expected convergence iterations K_0 (default: 1000)

```

1  $\bar{\lambda}_{\text{curr}} \leftarrow \lambda^{(0)}$  // current iterate average
2  $\gamma \leftarrow \gamma_0$  // learning rate
3  $k \leftarrow 0$  // total iterations
4  $t \leftarrow 0$  // total epochs
5 while  $k < K_{\max}$  do
6    $\bar{\lambda}_{\text{prev}} \leftarrow \bar{\lambda}_{\text{curr}}$  // record previous iterate average
7    $k_{\text{new}}, \bar{\lambda}_{\text{curr}}, \text{success} \leftarrow \text{FASO}(\bar{\lambda}_{\text{curr}}, \gamma, W_{\min}, \varepsilon, K_{\max} - k)$ 
8   if not success then
9     print "Warning: failed to converge. Estimated error is error"
10    break
11    $k \leftarrow k + k_{\text{new}}$  // update total iterations
12   if  $t \geq 1$  then
13      $\delta_t \leftarrow \text{SKL}(q_{\bar{\lambda}_{\text{prev}}}, q_{\bar{\lambda}_{\text{curr}}})$ 
14     compute estimates  $\hat{C}$  and  $\hat{\kappa}$  using weighted linear regression
15      $\widehat{\text{RSKL}} \leftarrow \rho^{\hat{\kappa}} + \xi / (\hat{C}^{1/2} \gamma^{\hat{\kappa}})$ 
16      $\hat{K}_{\gamma} \leftarrow k_{\text{new}}^{(t)} - k_{\text{new}}^{(t-1)}$ 
17     if  $t \geq 2$  then
18       // remove the converged iterations of initial variation parameter
19       compute estimates  $\hat{\alpha}$  and  $\hat{\beta}$  using weighted least squares
20       if  $\hat{\beta} < 0$  then
21          $\hat{K}_{\rho\gamma} \leftarrow (\rho\gamma)^{\hat{\alpha}} e^{\hat{\beta}}$ 
22       else
23          $\hat{K}_{\rho\gamma} \leftarrow \hat{K}_{\gamma}$ 
24          $\widehat{\text{RI}} \leftarrow \widehat{\text{RSKL}} / (\hat{K}_{\gamma} + K_0)$ 
25         if  $\widehat{\text{RSKL}} \cdot \widehat{\text{RI}} > \tau$  then
26           break
26    $\gamma \leftarrow \rho\gamma$  // decrease learning rate
27    $t \leftarrow t + 1$  // increment epoch counter
28 return  $\bar{\lambda}_{\text{curr}}$ 

```

7.1. Accuracy with Gaussian Targets

First, to explore optimization accuracy relative to the optimal variational approximation, we consider Gaussian targets of the form $\pi = \mathcal{N}(0, V)$. In such cases we can compute the ground-truth optimal variational approximation either analytically (because the distribution belongs to the mean-field variational family and hence $q_* = \pi$) or numerically using deterministic optimization (since the KL divergence between Gaussians is available in closed-form). Specifically, we consider the covariances $V = I$ (identity covariance), $V_{ij} = j\mathbb{1}[i = j]$ (diagonal non-identity covariance), $V_{ij} = \mathbb{1}[i = j] + 0.8\mathbb{1}[i \neq j]$ (uniform covariance with correlation 0.8), and $V_{ij} = \mathbb{1}[i = j] + 0.8^{|i-j|}\mathbb{1}[i \neq j]$ (banded covariance with maximum correlation 0.8).

Figure 7 shows the comparison between RAABBVI and fixed-learning rate versions of RMSProp, Windowed Adagrad (WAdagrad), and FASO. The results indicate that RAABBVI performs well when compared with fixed-learning rate approaches. Moreover, by varying the accuracy threshold ξ , the quality of the final approximation \hat{q}_* also varies such that $\text{SKL}(q_*, \hat{q}_*)^{1/2} \approx \xi$.

To demonstrate the flexibility of our framework, we used RAABBVI with a variety of optimization methods: RMSProp, avgRMSProp, avgAdam, natural gradient descent (NGD), and stochastic quasi-Newton (SQN). See Appendices B.5 and B.6 for details. Figure C.7 shows that avgAdam and avgRMSProp optimization methods have a similar improvement in symmetrized KL divergence between optimal and estimated variational approximation for all cases. NGD is not stable for the diagonal non-identity covariance structure and SQN does not perform well with uniform covariance structure. Even though RMSProp shows an improvement in accuracy for large step sizes, accuracy does not improve as the step size decreases.

7.2. Reliability Across Applications

To validate the robustness and reliability of RAABBVI across a realistic use cases, we consider 18 diverse dataset/model pairs found in the `posteriordb` package³ (see Table C.1 for details). The accuracy was computed based on ground-truth estimates obtained using the posterior draws included in `posteriordb` package if available. Otherwise, we ran Stan’s dynamic HMC algorithm [55] to obtain the ground truth (4 chains for 50,000 iterations each). To stabilize the optimization, we initialize the variational parameter estimates using RMSProp for the initial learning rate only. A comparison across optimization methods validates our choice of avgAdam over alternatives (Fig. C.8).

Accuracy and robustness. First we investigate the accuracy and algorithmic robustness of RAABBVI. In terms of robustness, Figures 8, C.10 and C.11 validate our termination criteria since after reaching the termination point there is no considerable improvement of the accuracy for most of the models and datasets. While in many cases the mean estimates are quite accuracy, the standard deviation estimates tended to be poor, which is consistent with typical behavior of mean-field approximations. To examine whether RAABBVI can achieve more accurate results with a more flexible variational family, we conduct the same experiment using multivariate Gaussian approximation family. In some cases the accuracy of the mean and/or standard deviations estimates improve (*bball_0*, *dogs_log*, *8schools_c*,

³<https://github.com/stan-dev/posteriordb>

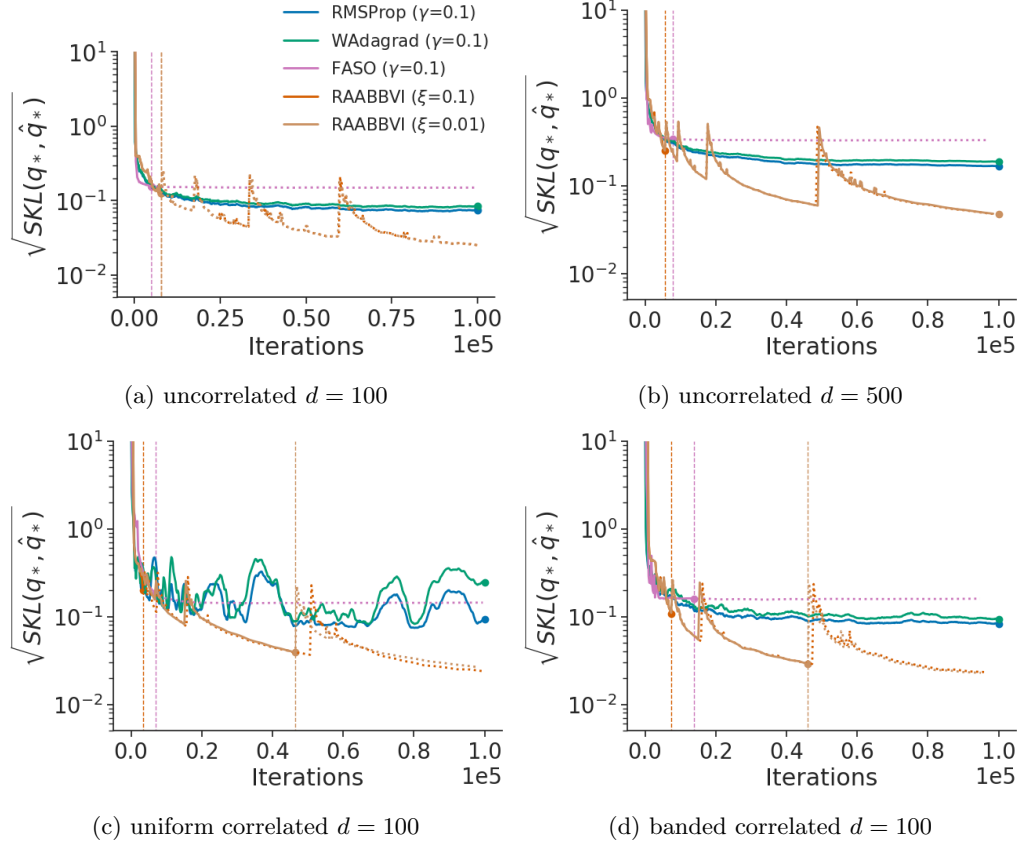


Fig 7: Accuracy comparison of variational inference algorithms using Gaussian targets, where accuracy is measured in terms of square root of symmetrized KL divergence between iterate average and optimal variational approximation. The vertical lines indicate the termination rule trigger points of FASO and RAABBVI. Iterate averages for RMSProp and WAdagrad computed at every 200th iteration using a window size of 20% of iterations.

hudson_lynx, *hmm_example*, *nes2000*, and *sblrc*). However, the results are inconsistent and sometimes worse due to the higher-dimensional, more challenging optimization problem. Overall, the results highlight the need to complement improved an optimization framework such as RAABBVI with posterior approximation accuracy diagnostics.

Comparison to MCMC. We additionally benchmarked the runtime and accuracy of RAABBVI to `Stan`'s dynamic HMC algorithm, for which we ran 1 chain for 25,000 iterations including 5,000 warmup iterations. We measure runtime in terms of number of likelihood evaluations. Figures 9, C.12 and C.13 show that RAABBVI tend to result similar or better posterior mean estimates (the exceptions are *gp_pois_regr*, *hudson_lynx*, and *sblrc*). However, the RAABBVI standard deviation estimates tend to be less accurate even when using the full-rank Gaussian variational family.

7.3. Robustness to Tuning Parameters: Ablation Study

To validate the robustness of RAABBVI to different choices of algorithm tuning parameters, we consider the Gaussian targets and the *dogs* dataset/model. We vary one tuning parameter while keeping the recommended default values for all others. We consider the following values for each parameter (default in bold):

- initial learning rate γ_0 : 0.01, 0.1, **0.3**, 0.5
- minimum window size W_{\min} : 100, **200**, 300, 500
- initial iterate average relative error threshold ε_0 : 0.05, **0.1**, 0.2, 0.5
- inefficiency threshold τ : 0.1, 0.5, **1.0**, 1.2
- Monte Carlo samples M : 1, 5, **10**, 15, 25.

We repeat each experimental condition 10 times to verify robustness to different initializations of the variational parameters. Figures 10, 11 and C.14 to C.18 suggest that overall the accuracy and runtime of RAABBVI is not too sensitive to the choice of the tuning parameters. However, extreme tuning parameter choices (e.g., $\gamma = 0.01$ or $M \leq 5$) can lead to longer runtimes.

8. Discussion

As we have shown through both theory and experiments, RAABBVI, our stochastic optimization framework for black-box variational inference, delivers a number of benefits compared to existing approaches:

- The user only needs to, at most, adjust a small number of tuning parameters which empower the user to intuitively control and trade off computational cost and accuracy. Moreover, RAABBVI is robust, both in terms of accuracy and computational cost, to small changes in all tuning parameters.
- Our framework can easily incorporate different stochastic optimization methods such as adaptive versions, natural gradient descent, and stochastic quasi-Newton methods. In practice we found that the averaged versions of RMSProp and Adam we propose perform particularly well. But the performance of RAABBVI will benefit from future innovations in stochastic optimization methodology.
- RAABBVI allows for any choice of tractable variational family and stochastic gradient estimator. For example, in many cases we find accuracy improves when using the full-rank Gaussian variational family rather than the mean-field one.

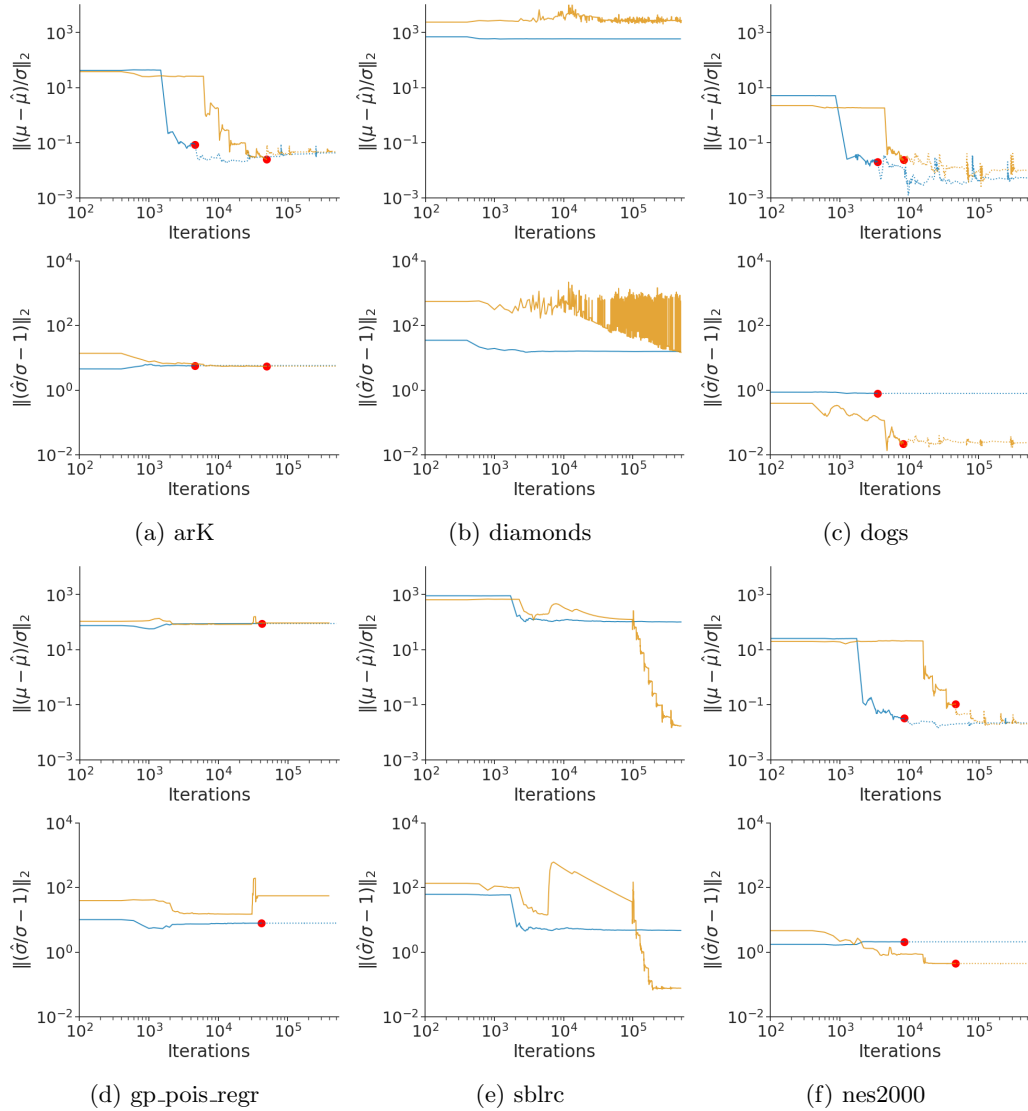


Fig 8: Accuracy of mean-field (blue) and full-rank (orange) Gaussian family approximations for selected `posteriordb` data/models, where accuracy is measured in terms of relative mean error (top) and relative standard deviation error (bottom). The red dots indicate where the termination rule triggers.

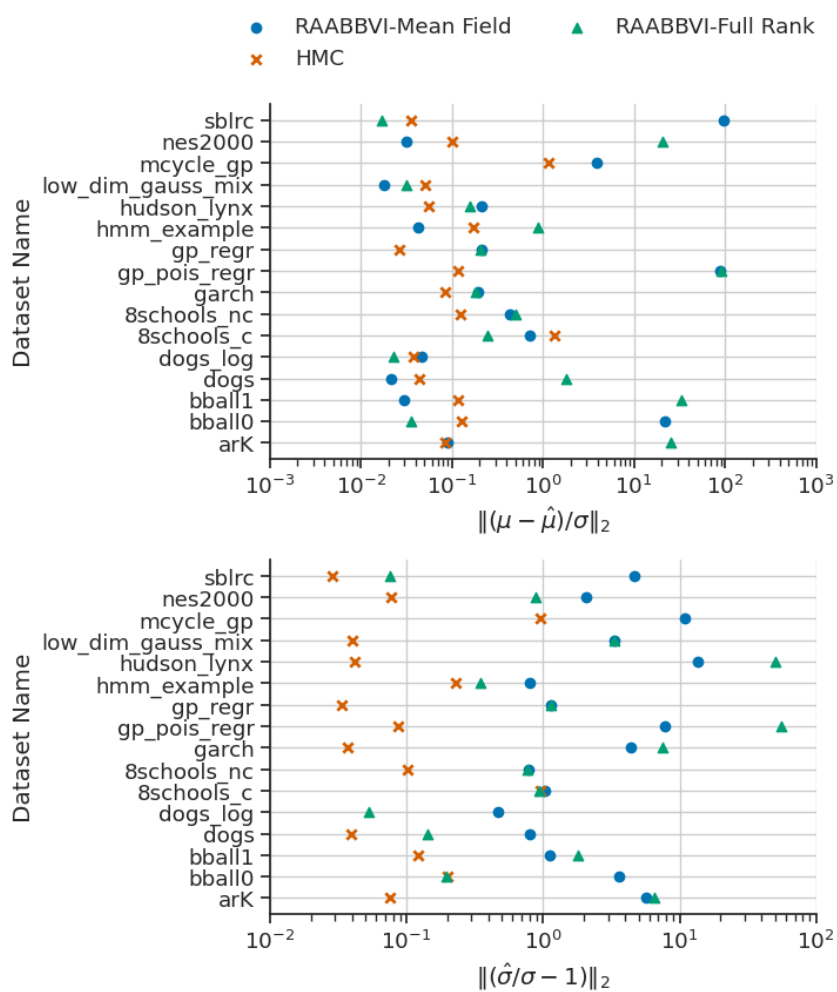


Fig 9: Results of RAABBVI with mean-field Gaussian and full-rank Gaussian family comparison to dynamic HMC in terms of relative mean error (top) and relative standard deviation error (bottom).

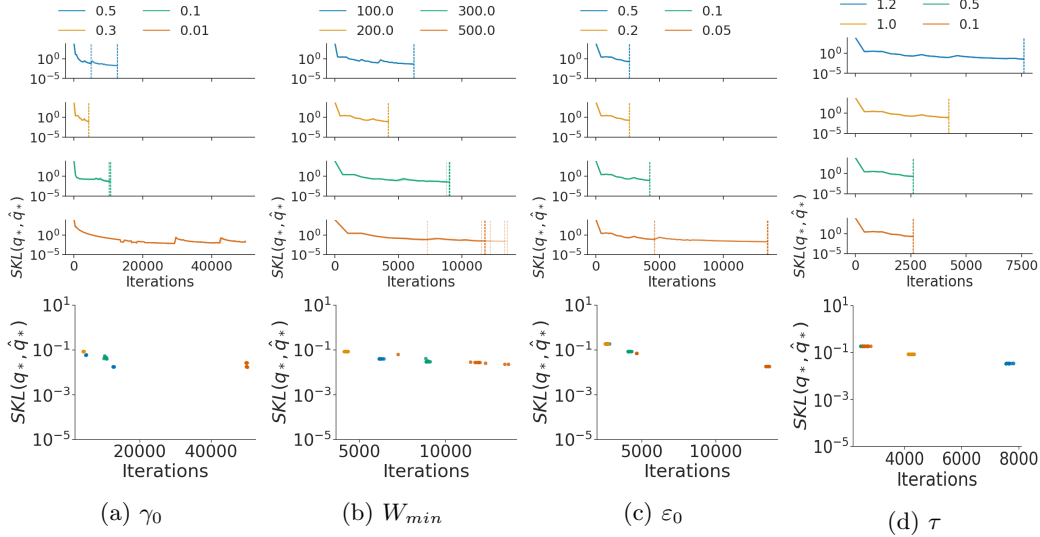


Fig 10: Robustness to tuning parameters (a) initial learning rate γ_0 , (b) minimum window size W_{min} , (c) initial iterate average relative error threshold ϵ_0 , and (d) inefficiency threshold τ . Results use Gaussian target $\mathcal{N}(0, V)$ with $d = 100$ and $V = I$ (identity covariance). **(top)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation. The distinct lines represent repeated experiments and the vertical lines indicate the termination rule trigger points. **(bottom)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation at the termination rule trigger point.

Our empirical results also highlight some of the limitations of BBVI, which sometimes is less accurate than dynamic HMC when given equal computational budgets. However, BBVI can be further sped up using, for example, data subsampling when the dataset size is large (which was not the case for the `posteriodb` datasets from our experiments).

Acknowledgments

M. Welandawe and J. H. Huggins were supported by the National Institute of General Medical Sciences of the National Institutes of Health under grant number R01GM144963 as part of the Joint NSF/NIGMS Mathematical Biology Program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. A. Vehtari was supported by Academy of Finland Flagship programme: Finnish Center for Artificial Intelligence, FCAI and M. R. Andersen was supported by Innovation Fund Denmark (grant number 8057-00036A).

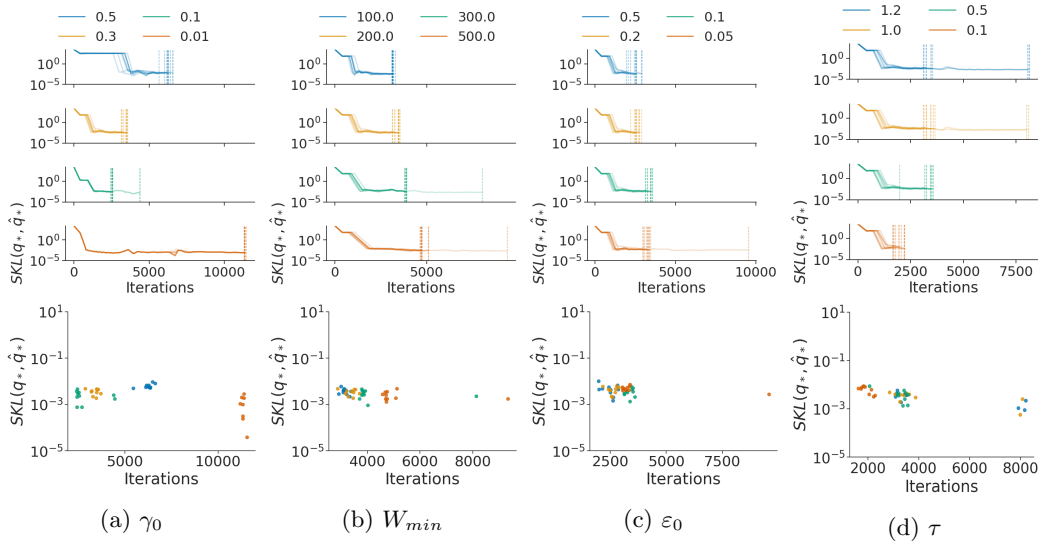


Fig 11: Robustness to tuning parameters (a) initial learning rate γ_0 , (b) minimum window size W_{min} , (c) initial iterate average relative error threshold ϵ_0 , and (d) inefficiency threshold τ using *dogs* dataset from `posteriordb` package. **(top)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation. The transparent lines represent repeated experiments and the vertical lines indicate the termination rule trigger points. **(bottom)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation at the termination rule trigger point.

References

- [1] AGRAWAL, A., SHELDON, D. and DOMKE, J. (2020). Advances in Black-Box VI: Normalizing Flows, Importance Weighting, and Optimization. In *Advances in Neural Information Processing Systems*.
- [2] BACH, F. and MOULINES, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems* 1–9.
- [3] BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- [4] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **112** 859–877.
- [5] BOLLEY, F. and VILLANI, C. (2005). Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la faculte des sciences de Toulouse* **13** 331–352.
- [6] BUBECK, S., ELDAN, R. and LEHEC, J. (2015). Finite-Time Analysis of Projected Langevin Monte Carlo. In *Advances in Neural Information Processing Systems*.
- [7] BUI, T. D., YAN, J. and TURNER, R. E. (2017). A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation. *Journal of Machine Learning Research* **18** 1–72.
- [8] BURDA, Y., GROSSE, R. B. and SALAKHUTDINOV, R. (2016). Importance Weighted Autoencoders. In *International Conference on Learning Representations*.
- [9] CHEE, J. and LI, P. (2020). Understanding and Detecting Convergence for Stochastic Gradient Descent with Momentum. *arXiv.org arXiv:2008.12224 [cs.LG]*.
- [10] CHEE, J. and TOULIS, P. (2018). Convergence diagnostics for stochastic gradient descent with constant learning rate. In *International Conference on Artificial Intelligence and Statistics*.
- [11] CHEN, H., WU, H.-C., CHAN, S.-C. and LAM, W.-H. (2019). A stochastic quasi-Newton method for large-scale nonconvex optimization with applications. *IEEE transactions on neural networks and learning systems* **31** 4776–4790.
- [12] CORNEBISE, J., MOULINES, E. and OLSSON, J. (2008). Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing* **18** 461–480.
- [13] DHAKA, A. K., CATALINA, A., ANDERSEN, M. R., MAGNUSSEN, M., HUGGINS, J. H. and VEHTARI, A. (2020). Robust, Accurate Stochastic Optimization for Variational Inference. In *Advances in Neural Information Processing Systems*.
- [14] DHAKA, A. K., CATALINA, A., WELANDAWA, M., ANDERSEN, M. R., HUGGINS, J. and VEHTARI, A. (2021). Challenges and Opportunities in High Dimensional Variational Inference. *Advances in Neural Information Processing Systems* **34**.
- [15] DIENG, A. B., TRAN, D., RANGANATH, R., PAISLEY, J. and BLEI, D. M. (2017). Variational Inference via χ Upper Bound Minimization. In *Advances in Neural Information Processing Systems*.
- [16] DIEULEVEUT, A., DURMUS, A. and BACH, F. (2020). Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. *The Annals of Statistics* **48** 1348–1382.
- [17] DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research* **12** 2121–2159.

- [18] DURMUS, A., MAJEWSKI, S. and MIASOJEDOW, B. (2019). Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research* **20** 1–46.
- [19] DURMUS, A. and MOULINES, E. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **25** 2854–2882.
- [20] EBERLE, A. and MAJKA, M. B. (2019). Quantitative contraction rates for Markov chains on general state spaces. *Electronic Journal of Probability* **24** 1–36.
- [21] GAL, Y. and GHAHRAMANI, Z. (2016). Dropout as a Bayesian Approximation - Representing Model Uncertainty in Deep Learning. In *International Conference on Machine Learning*.
- [22] GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7** 457–511.
- [23] GELMAN, A., CARLIN, J., STERN, H., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, Third ed. Chapman and Hall/CRC.
- [24] GEYER, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science* **7** 473–483.
- [25] GITMAN, I., LANG, H., ZHANG, P. and XIAO, L. (2019). Understanding the Role of Momentum in Stochastic Gradient Methods. In *Advances in Neural Information Processing Systems* 9633–9643.
- [26] HERNÁNDEZ-LOBATO, J. M., LI, Y., ROWLAND, M., BUI, T. D., HERNÁNDEZ-LOBATO, D. and TURNER, R. E. (2016). Black-Box Alpha Divergence Minimization. In *International Conference on Machine Learning*.
- [27] HINTON, G. E. and TIELEMAN, T. (2012). Lecture 6.5 – Rmsprop: Divide the gradient by a running average of its recent magnitude. In *Coursera: Neural networks for machine learning*.
- [28] HUGGINS, J. H., KASPRZAK, M., CAMPBELL, T. and BRODERICK, T. (2020). Validated Variational Inference via Practical Posterior Error Bounds. In *International Conference on Artificial Intelligence and Statistics*.
- [29] JOHNSON, M. J., DUVENAUD, D., WILTSCHKO, A. B., DATTA, S. R. and ADAMS, R. P. (2016). Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*.
- [30] JOULIN, A. and OLLIVIER, Y. (2010). Curvature, concentration and error estimates for Markov chain Monte Carlo. *The Annals of Probability* **38** 2418–2442.
- [31] KHAN, M. E. and LIN, W. (2017). Conjugate-Computation Variational Inference : Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models. In *International Conference on Artificial Intelligence and Statistics*.
- [32] KINGMA, D. P. and BA, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.
- [33] KINGMA, D. P. and WELLING, M. (2014). Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.
- [34] KUCUKELBIR, A., RANGANATH, R., GELMAN, A. and BLEI, D. M. (2015). Automatic Variational Inference in Stan. In *Advances in Neural Information Processing Systems*.
- [35] LANG, H., ZHANG, P. and XIAO, L. (2019). Using Statistics to Automate Stochastic Optimization. *arXiv.org arXiv:1909.09785 [stat.ML]*.
- [36] LI, Y. and TURNER, R. E. (2016). Rényi Divergence Variational Inference. In *Advances in Neural Information Processing Systems* 1073–1081.
- [37] LIU, S. and OWEN, A. B. (2021). Quasi-Monte Carlo Quasi-Newton in Variational Bayes. *Journal of Machine Learning Research* **22** 1–23.

- [38] MADDOX, W., GARIPPOV, T., IZMAILOV, P., VETROV, D. and WILSON, A. G. (2019). A Simple Baseline for Bayesian Uncertainty in Deep Learning. In *Advances in Neural Information Processing Systems*.
- [39] MADRAS, N. and SEZER, D. (2010). Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli* **16** 882–908.
- [40] MOHAMED, S., ROSCA, M., FIGURNOV, M. and MNIH, A. (2020). Monte Carlo Gradient Estimation in Machine Learning. *Journal of Machine Learning Research* **21** 1–62.
- [41] MUKKAMALA, M. C. and HEIN, M. (2017). Variants of RMSProp and Adagrad with Logarithmic Regret Bounds. In *International Conference on Machine Learning*.
- [42] MURPHY, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA.
- [43] NOCEDAL, J. and WRIGHT, S. (2006). *Numerical optimization*. Springer Science & Business Media.
- [44] PAPAMAKARIOS, G., NALISNICK, E., REZENDE, D. J., MOHAMED, S. and LAKSHMINARAYANAN, B. (2019). Normalizing Flows for Probabilistic Modeling and Inference. *arXiv.org arXiv:1912.02762 [stat.ML]*.
- [45] PESME, S., DIEULEVEUT, A. and FLAMMARION, N. (2020). On Convergence-Diagnostic based Step Sizes for Stochastic Gradient Descent. In *International Conference on Machine Learning*.
- [46] PFLUG, G. C. (1990). Non-asymptotic confidence bounds for stochastic approximation algorithms with constant step size. *Monatshefte für Mathematik* **110** 297–314.
- [47] POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal of Control and Optimization* **30** 838–855.
- [48] RANGANATH, R., GERRISH, S. and BLEI, D. M. (2014). Black Box Variational Inference. In *International Conference on Artificial Intelligence and Statistics* 814–822.
- [49] REZENDE, D. J., MOHAMED, S. and WIERSTRA, D. (2014). Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *International Conference on Machine Learning*.
- [50] ROBERT, C. P. (2007). *The Bayesian Choice*, 2nd ed. Springer, New York, NY.
- [51] RUDOLF, D. and SCHWEIZER, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* **4A** 2610–2639.
- [52] RUPPERT, D. (1988). Efficient estimations from a slowly convergent Robbins–Monro process Technical Report, Cornell University Operations Research and Industrial Engineering.
- [53] SAATCHI, Y. and WILSON, A. G. (2017). Bayesian GAN. In *Advances in Neural Information Processing Systems*.
- [54] SALIMANS, T., KINGMA, D. P. and WELLING, M. (2015). Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. In *International Conference on Machine Learning*.
- [55] STAN DEVELOPMENT TEAM (2020). Stan Modeling Language Users Guide and Reference Manual. **2.26**.
- [56] TITSIAS, M. K. and LÁZARO-GREDILLA, M. (2014). Doubly Stochastic Variational Bayes for non-Conjugate Inference. In *International Conference on Machine Learning*.
- [57] VATS, D. and KNUDSON, C. (2018). Revisiting the Gelman-Rubin Diagnostic. *arXiv.org arXiv: 1812.09384 [stats.CO]*.
- [58] VEHTARI, A., GELMAN, A., SIMPSON, D., CARPENTER, B. and BÜRKNER, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved \hat{R} for Assessing Convergence of MCMC. *Bayesian Analysis* **16** 667–718.

- [59] VILLANI, C. (2009). *Optimal transport: old and new*. Grundlehren der mathematischen Wissenschaften **338**. Springer.
- [60] VOLLMER, S. J., ZYGALAKIS, K. C. and TEH, Y. W. (2016). (Non-) asymptotic properties of Stochastic Gradient Langevin Dynamics. *Journal of Machine Learning Research* **17** 1–48.
- [61] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends[®] in Machine Learning* **1** 1–305.
- [62] WAN, N., LI, D. and HOVAKIMYAN, N. (2020). f-Divergence Variational Inference. *arXiv*.
- [63] WANG, D., LIU, H. and LIU, Q. (2018). Variational Inference with Tail-adaptive f-Divergence. In *Advances in Neural Information Processing Systems*.
- [64] YAIDA, S. (2019). Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*.
- [65] YAO, Y., VEHTARI, A., SIMPSON, D. and GELMAN, A. (2018). Yes, but Did It Work?: Evaluating Variational Inference. In *International Conference on Machine Learning*.
- [66] ZHANG, P., LANG, H., LIU, Q. and XIAO, L. (2020). Statistical Adaptive Stochastic Gradient Methods. *arXiv.org arXiv:2002.10597 [stat.ML]*.

Appendix A: Proofs

A.1. Proof of Proposition 4.4

Since the KL divergence of mean-field Gaussians factors across dimensions, without loss of generality we consider the case of $\dim = 1$. The symmetrized KL divergence between two Gaussians is

$$\begin{aligned} \text{SKL}(\lambda_1, \lambda_2) &= \frac{1}{2} \{ e^{2\psi_1 - 2\psi_2} + e^{2\psi_2 - 2\psi_1} + (\tau_1 - \tau_2)^2 (e^{-2\psi_1} + e^{-2\psi_2}) - 2 \} \\ &= \frac{1}{2} \{ e^{2\psi_1 - 2\psi_2} + e^{2\psi_2 - 2\psi_1} + (\tau_1 - \tau_2)^2 e^{-2\psi_1} (1 + e^{2\psi_1 - 2\psi_2}) - 2 \}. \end{aligned}$$

Recall our assumption that for some constants $A = (A_\tau, A_\psi)$ and $B = (B_\tau, B_\psi)$,

$$\bar{\lambda}_\gamma = \lambda_* + A\gamma + B\gamma^2 + o(\gamma^2).$$

Hence

$$\begin{aligned} e^{2\bar{\psi}_\gamma - 2\psi_*} &= 1 + 2A_\psi\gamma + 2B_\psi\gamma^2 + \frac{1}{2}(2A_\psi\gamma + 2B_\psi\gamma^2)^2 + o(\gamma^2) \\ &= 1 + 2A_\psi\gamma + 2B_\psi\gamma^2 + 2A_\psi^2\gamma^2 + o(\gamma^2) \end{aligned}$$

and similarly $e^{2\psi_* - 2\bar{\psi}_\gamma} = 1 - 2A_\psi\gamma - 2B_\psi\gamma^2 + 2A_\psi^2\gamma^2 + o(\gamma^2)$. Therefore $e^{2\bar{\psi}_\gamma - 2\psi_*} + e^{2\psi_* - 2\bar{\psi}_\gamma} - 2 = 4A_\psi^2\gamma^2 + o(\gamma^2)$ and $(\bar{\tau}_\gamma - \tau_*)^2 = A_\tau^2\gamma^2 + o(\gamma^2)$. Putting these results together, we have

$$\begin{aligned} \text{SKL}(\bar{\lambda}_\gamma, \lambda_*) &= \frac{1}{2} \left[4A_\psi^2\gamma^2 + o(\gamma^2) + A_\tau^2\gamma^2 e^{-\psi_*} \{ 2 - 2A_\psi\gamma - 2B_\psi\gamma^2 + 2\gamma^2 A_\psi^2 + o(\gamma^2) \} \right] \\ &= (2A_\psi^2 + A_\tau^2 e^{-2\psi_*})\gamma^2 + o(\gamma^2). \end{aligned}$$

Similarly, $\text{SKL}(\bar{\lambda}_\gamma, \bar{\lambda}_{\gamma'}) = (\gamma - \gamma')^2 (2A_\psi^2 + A_\tau^2 e^{-2\psi_*}) + o(\gamma^2)$ as long as $\gamma' = O(\gamma)$. In other words, letting $\bar{\lambda}_0 := \lambda_*$, we have the relation

$$\text{SKL}(\bar{\lambda}_\gamma, \bar{\lambda}_{\gamma'}) = C(\gamma - \gamma')^2 + o(\gamma^2)$$

for some unknown constant C .

A.2. Proof of Proposition 4.5

If $\kappa = 1$ corresponds to the statement of Proposition 4.4.

If $\kappa \in [1/2, 1)$, then for some constants $\Lambda = (\Lambda_\tau, \Lambda_\psi)$ and $A = (A_\tau, A_\psi)$,

$$\bar{\lambda}_\gamma = \lambda_* + \Lambda\gamma^\kappa + A\gamma + o(\gamma^{2\kappa}).$$

Assume that $\gamma_1 = \gamma$ and $\gamma_2 = O(\gamma)$. For $\kappa' > 0$, let $\epsilon_{\kappa'} := \gamma_1^{\kappa'} - \gamma_2^{\kappa'}$. Then we have

$$e^{\pm 2(\bar{\psi}_{\gamma_1} - \bar{\psi}_{\gamma_2})} = 1 \pm 2\Lambda_\psi\epsilon_\kappa \pm 2A_\psi\epsilon_1 + 2(\Lambda_\psi^2\epsilon_\kappa^2 + 2\Lambda_\psi A_\psi\epsilon_\kappa\epsilon_1 + A_\psi^2\epsilon_1^2) + o(\gamma^{2\kappa}),$$

so

$$e^{2\bar{\psi}_{\gamma_1} - 2\bar{\psi}_{\gamma_2}} + e^{2\bar{\psi}_{\gamma_2} - 2\bar{\psi}_{\gamma_1}} - 2 = 4\Lambda_\psi^2\epsilon_\kappa^2 + 8\Lambda_\psi A_\psi\epsilon_\kappa\epsilon_1 + 4A_\psi^2\epsilon_1^2 + o(\gamma^{2\kappa})$$

Moreover, $e^{\psi_* - \bar{\psi}_{\gamma_i}} = 1 + O(\gamma^\kappa)$ and

$$(\bar{\tau}_{\gamma_1} - \bar{\tau}_{\gamma_2})^2 = \Lambda_\tau^2 \epsilon_\kappa^2 + 2\Lambda_\tau A_\tau \epsilon_\kappa \epsilon_1 + A_\tau^2 \epsilon_1^2 + o(\gamma^{2\kappa}).$$

Putting these results together, we have

$$\begin{aligned} \text{SKL}(\bar{\lambda}_{\gamma_1}, \bar{\lambda}_{\gamma_2}) &= \frac{1}{2} \left[4\Lambda_\psi^2 \epsilon_\kappa^2 + 8\Lambda_\psi A_\psi \epsilon_\kappa \epsilon_1 + 4A_\psi^2 \epsilon_1^2 + o(\gamma^{2\kappa}) \right. \\ &\quad \left. + e^{-\psi_*} (\Lambda_\tau^2 \epsilon_\kappa^2 + 2\Lambda_\tau A_\tau \epsilon_\kappa \epsilon_1 + A_\tau^2 \epsilon_1^2) \{2 + O(\gamma^\kappa)\} \right] \\ &= 2\Lambda_\psi^2 \epsilon_\kappa^2 + \Lambda_\tau^2 e^{-\psi_*} \epsilon_\kappa^2 + O(\gamma^{1+\kappa}) \\ &= C(\gamma_1^\kappa - \gamma_2^\kappa)^2 + o(\gamma^{2\kappa}), \end{aligned} \tag{A.1}$$

where $C := 2\Lambda_\psi^2 + \Lambda_\tau^2 e^{-\psi_*}$.

A.3. Symmetric KL divergence termination rule

We can use Eq. (A.1) to derive a termination rule. If γ is the current learning rate, the previous learning rate was γ/ρ . Ignoring $o(\gamma^{2\kappa})$ terms, we have

$$\delta_\gamma := \text{SKL}(\bar{\lambda}_{\gamma/\rho}, \bar{\lambda}_\gamma) = C\gamma^{2\kappa}(1/\rho^\kappa - 1)^2.$$

Therefore, we can estimate p and C using a regression model of the form

$$\log \delta_\gamma = \log C + 2 \log(1/\rho^\kappa - 1) + 2p \log \gamma.$$

Given estimates $\hat{\kappa}$ and \hat{C} , we can estimate $\text{SKL}(\bar{\lambda}_\gamma, \lambda_*) \approx \hat{C}\gamma^{2\hat{\kappa}}$.

A.4. Proof of Proposition 5.1

Since for $|x| < 1/2$ it holds that $|\exp(x) - 1| \leq 1.5|x|$, we have

$$\frac{|\hat{\sigma}_i - \bar{\sigma}_i|}{\bar{\sigma}_i} = |\exp(\hat{\psi}_i - \bar{\psi}_i) - 1| \leq 1.5\varepsilon'$$

and

$$\frac{|\hat{\tau}_i - \bar{\tau}_i|}{\bar{\sigma}_i} \leq \varepsilon' \hat{\sigma}_i / \bar{\sigma}_i \leq \varepsilon'(1 + 1.5\varepsilon') \leq 1.75\varepsilon'.$$

A.5. Proof of Proposition 5.2

In the optimal case, the total cost for the iterations used for iterate averaging is

$$\begin{aligned} \text{OPT} &= C_O W_{\text{opt}} + C_E W_{\text{conv}} + C_E W_{\text{opt}} \\ &= C_E (r W_{\text{opt}} + W_{\text{conv}} + W_{\text{opt}}). \end{aligned}$$

On the other hand, if checking at window sizes $W_j := \chi^j W_{\text{conv}}$ ($j = 1, 2, \dots$), then window size at which convergence will be detected is $j_* := \lceil \log(W_{\text{opt}}/W_{\text{conv}})/\log(\chi) \rceil$, and in particular

$$W_{j_*} \leq \chi W_{\text{opt}}.$$

Therefore we can bound the actual computational cost as

$$\begin{aligned} & C_O W_{j_*} + C_E W_{\text{conv}} \log W_{\text{conv}} + C_E \sum_{j=1}^{j_*} W_j \\ & \leq C_O \chi W_{\text{opt}} + C_E W_{\text{conv}} \log W_{\text{conv}} + C_E \sum_{j=1}^{j_*} \chi^j W_{\text{conv}} \\ & \leq C_O \chi W_{\text{opt}} + C_E W_{\text{conv}} \log W_{\text{conv}} + C_E \frac{\chi(\chi^{j_*} - 1)W_{\text{conv}}}{\chi - 1} \\ & \leq C_O \chi W_{\text{opt}} + C_E W_{\text{conv}} \log W_{\text{conv}} + C_E \frac{\chi^2 W_{\text{opt}}}{\chi - 1} \\ & = C_E \left[\chi r W_{\text{opt}} + W_{\text{conv}} + \frac{\chi^2}{\chi - 1} W_{\text{opt}} \right]. \end{aligned}$$

If $\chi = 2$, then the actual computational cost is bounded by

$$\begin{aligned} C_E [2r W_{\text{opt}} + W_{\text{conv}} + 4W_{\text{opt}}] & \leq 4C_E [r W_{\text{opt}} + W_{\text{conv}} + W_{\text{opt}}] \\ & = 4 \text{OPT}. \end{aligned}$$

On the other hand, we can minimize the total cost by solving $\chi_* := \arg \min_{\chi > 1} \{r\chi + \chi^2/(\chi - 1)\} = \chi(r)$. Plugging this back in, we get the bound

$$\begin{aligned} & C_E \frac{2 + r + 2\sqrt{1+r}}{1+r} (r W_{\text{opt}} + W_{\text{conv}} \log W_{\text{conv}} + W_{\text{opt}}) \\ & = \frac{2 + r + 2\sqrt{1+r}}{1+r} \text{OPT} \\ & < 4 \text{OPT}. \end{aligned}$$

Appendix B: Further Details

B.1. Effective sample size (ESS) and Monte Carlo standard error (MCSE)

Let $v^{(1)}, v^{(2)}, \dots$ denote a stationary Markov chain, let $\bar{v} := \mathbb{E}[v^{(1)}]$ denote the mean at stationarity, and let $\hat{v} := K^{-1} \sum_{k=1}^K v^{(k)}$ denote the Monte Carlo estimate for \bar{v} . The (ideal) effective sample size is defined as

$$\text{ESS}(K) := K / \left(1 + \sum_{k=1}^{\infty} \rho_k\right),$$

where ρ_k is the autocorrelation of the Markov chain at lag k . The ESS can be efficiently estimated using a variety of methods [24, 58]. We write $\widehat{\text{ESS}}(v^{(1:K)})$ to denote an estimator

for $\text{ESS}(K)$ based on the sequence $v^{(1:K)} := (v^{(1)}, \dots, v^{(K)})$. The Monte Carlo standard error of \hat{v} is given by

$$\text{MCSE}(\hat{v}) := \sigma(v^{(1)})/\text{ESS}(K),$$

where $\sigma(v^{(1)})$ denote the standard deviation of the random variable $v^{(1)}$. Given the empirical standard deviation of $v^{(1)}, \dots, v^{(K)}$, which we denote $\hat{\sigma}(v^{(1:K)})$, the MCSE can be approximated by

$$\widehat{\text{MCSE}}(v^{(1:K)}) := \hat{\sigma}(v^{(1:K)})/\widehat{\text{ESS}}(v^{(1:K)}).$$

B.2. Total variation distance and KL divergence

For distributions η and ζ , $|I_{\eta,i,a,b} - I_{\zeta,i,a,b}| \leq \sqrt{\text{KL}(\eta | \zeta)/2}$ for all $a < b$ and i . This guarantee follows from Pinsker's inequality, which relates the KL divergence to the total variation distance

$$d_{\text{TV}}(\eta, \zeta) := \sup_{f: \|f\|_{\infty} \leq 1} \left| \int f(\theta) \eta(d\theta) - \int f(\theta) \zeta(d\theta) \right|,$$

where $\|f\|_{\infty} := \sup_{\theta} f(\theta) - \inf_{\theta} f(\theta)$. Specifically, Pinsker's inequality states that $d_{\text{TV}}(\eta, \zeta) \leq \sqrt{\text{KL}(\eta | \zeta)/2}$. Thus, small KL divergence implies small total variance distance, which implies the difference between expectations for any function f such that $\|f\|_{\infty}$ is small. In the case of interval probabilities, since $\|\mathbf{1}(\cdot \in [a, b])\|_{\infty} = 1$, it follows that

$$\begin{aligned} |I_{\eta,i,a,b} - I_{\zeta,i,a,b}| &= \left| \int \mathbf{1}(\theta \in [a, b]) \eta(d\theta) - \int \mathbf{1}(\theta \in [a, b]) \zeta(d\theta) \right| \\ &\leq \|\mathbf{1}(\cdot \in [a, b])\|_{\infty} d_{\text{TV}}(\eta, \zeta) \\ &\leq \sqrt{\text{KL}(\eta | \zeta)/2}. \end{aligned}$$

B.3. Adaptive SASA+

The SASA+ algorithm of Zhang et al. [66] generalizes the approach of Yaida [64]. The main idea is to find an appropriate *invariant function* $\Delta(d, \lambda)$ that satisfies $\int \Delta(d, \lambda) \mu_{\gamma}(dd, d\lambda) = 0$. Yaida [64] derived valid forms of Δ for specific choices of the descent direction, while Zhang et al. [66] showed that for any optimizer of the form Eq. (1) that is time-homogenous with $\gamma^{(k)} = \gamma$, the map $(d, \lambda) \mapsto 2\langle d, \lambda \rangle - \gamma \|d\|^2$ is a valid invariant function. The SASA+ algorithm proposed by Zhang et al. [66] uses a hypothesis test to determining when the iterates are sufficiently close to stationarity. Let $\Delta^{(k)} := 2\langle d^{(k)}, \lambda^{(k)} \rangle - \gamma \|d^{(k)}\|^2$ and let $W = \lceil \rho k \rceil$ denote the window size to use for checking stationarity. Once W is at least equal to a minimum window size W_{\min} , SASA+ uses $\Delta^{(k-W+1)}, \dots, \Delta^{(k)}$ to carry out a hypothesis test, where the null hypothesis is that $\mathbb{E}[\Delta^{(k)}] = 0$.

We make several adjustments to reduce the number of tuning parameters and to make the remaining ones more intuitive. Note that the SASA+ convergence criterion requires the choice of three parameters: ρ , W_{\min} , and the size of the hypothesis test α . Zhang et al. [66] showed empirically and our numerical experiments confirmed that the choice of α has little effect and therefore does not need to be adjusted by the user. The choices for ρ and W_{\min} ,

however, have a substantial effect on efficiency. If ϱ is too big, then early iterations that are not at stationarity will be included, preventing the detection of convergence. On the other hand, if ϱ is too small, then the total number of iterations must be large (specifically, greater than W_{\min}/ϱ) before the window size is large enough to trigger the first check for stationarity. Moreover, the correct choice of W_{\min} will vary depending on the problem. If the iterates have large autocorrelation then W_{\min} should be large, while if the autocorrelation is small or negative, then W_{\min} can be small.

Our approach to determining the optimal window size instead relies on the effective sample size (ESS). W_{opt} that maximizes $\widehat{\text{ESS}}(W) := \widehat{\text{ESS}}(\Delta^{(k-W+1)}, \dots, \Delta^{(k)})$, where k is the current iteration. To ensure reliability, we impose additional conditions on W_{opt} . First, we require that $\widehat{\text{ESS}}(W_{\text{opt}}) \geq N_{\min}$, a user-specified minimum effective sample size. Unlike W_{\min} , N_{\min} has an intuitive and direct interpretation. The second condition is, when finding W_{opt} , the search over values of W is constrained to the lower bound of N_{\min} (to ensure the estimator is sufficiently reliable) and the upper bound of $0.95k$ (to always allow for some “burn-in”). In practice we do not check all $W \in \{N_{\min}, \dots, 0.95k\}$, but rather perform a grid search over 5 equally spaced values ranging from N_{\min} to $0.95k$.

A slightly different version which may improve power is to instead define the multivariate invariant function $\bar{\Delta}(d, \lambda) = 2d \odot \lambda - d \odot d$, where we recall that \odot denotes component-wise multiplication. In this case a multivariate hypothesis test such as Hotelling’s T^2 test or the multivariate sign test, where a single effective sample size (e.g., the median component-wise ESS) could be used. Alternatively, a separate hypothesis test for each of the m components could be used, with stationarity declared once all tests confirm stationarity (using, e.g., a test size of α/m). While this approach might be more computationally efficient when m is large, it could come at the cost of test power.

B.4. Distance Based Convergence Detection

Pesme, Dieuleveut and Flammarion [45] proposed a distance based diagnostic algorithm to detect the stationarity of the SGD optimization algorithm. The main idea behind this algorithm is to find the distance between the current iterate λ_k and the optimal variational parameter λ_* , $\|\eta_k\| := \|\lambda_k - \lambda_*\|$. Since optimal variational parameter λ_* unknown, this distance cannot be directly observed. Therefore, they suggested using the distance between the current iterate λ_k and the initial iterate of the current learning rate, $\|\Omega_k\| := \|\lambda_k - \lambda_0\|$ and showed that $\|\eta_k\|$ and $\|\Omega_k\|$ have a similar behavior.

Under the setting of quadratic objective function with additive noise, they computed the behavior of the expectation of $\|\Omega_k\|^2$ in closed-form to detect the convergence of $\|\lambda_k - \lambda_0\|^2$. With the result, they have shown the asymptotic behavior of the $\mathbb{E}[\|\Omega_k\|^2]$ in the transient and stationary phases, where $\mathbb{E}[\|\Omega_k\|^2]$ has a slope greater than 1 and slope of 0 in a log-log plot respectively. Hence, the slope $S := \frac{\log\|\lambda_k - \lambda_0\|^2 - \log\|\lambda_k/q - \lambda_0\|^2}{\log k - \log k/q}$ computed between iterations q^n and q^{n+1} for $q > 1$ and $n \geq n_0$ where $n_0 \in \mathcal{N}^*$ and if $S < \text{thresh}$ ($\text{thresh} \in (0, 2]$) then the convergence will be detected.

B.5. Stochastic Quasi-Newton Optimization

The algorithm of Liu and Owen [37] provide a randomized approach of classical quasi-Newton (QN) method that is known as stochastic quasi-Newton (SQN) method. Even though, Liu and Owen [37] use randomized quasi-Monte Carlo (RMCQ) samples we use

Monte Carlo (MC) samples to compute the gradient. In classical quasi-Newton optimization the Newton update is

$$\lambda^{(k+1)} \leftarrow \lambda^{(k)} - (\nabla_{\lambda^{(k)}}^2 D_{\pi}(q_{\lambda^{(k)}}))^{-1} \nabla_{\lambda^{(k)}} D_{\pi}(q_{\lambda^{(k)}}),$$

where $\nabla_{\lambda^{(k)}}^2 D_{\pi}(q_{\lambda^{(k)}})$ is the Hessian matrix of $D_{\pi}(q_{\lambda^{(k)}})$. However, computation cost of Hessian matrix and its inverse is high and it also takes a large space. Therefore, we can use BFGS (discovered by Broyden, Fletcher, Goldfarb, and Shanno) method where it approximate the inverse of $\nabla_{\lambda^{(k)}}^2 D_{\pi}(q_{\lambda^{(k)}})$ using H_k at the k th iteration by initializing it with an identity matrix. Then the update is modified by

$$\lambda^{(k+1)} \leftarrow \lambda^{(k)} - \gamma^{(k)} H^{(k)} \nabla_{\lambda^{(k)}} D_{\pi}(q_{\lambda^{(k)}}),$$

where

$$H^{(k+1)} \leftarrow \left(I - \frac{s_k y_k^T}{s_k^T y_k} \right) H^{(k)} \left(I - \frac{y_k s_k^T}{s_k^T y_k} \right) + \frac{s_k s_k^T}{s_k^T y_k},$$

$s_k = \lambda^{(k+1)} - \lambda^{(k)}$, and $y_k = \nabla_{\lambda^{(k+1)}} D_{\pi}(q_{\lambda^{(k+1)}}) - \nabla_{\lambda^{(k)}} D_{\pi}(q_{\lambda^{(k)}})$. Even using the above Hessian approximation $H^{(k)}$ will require large space for storage. To overcome that problem we can use Limited-memory BFGS (L-BFGS) [43] that computes $H^{(k)} \nabla_{\lambda^{(k)}} D_{\pi}(q_{\lambda^{(k)}})$ using m most recent (s_k, y_k) correction pairs. Liu and Owen [37] use the stochastic quasi-Newton approach proposed by Chen et al. [11] and they compute the correction pairs after every B iterations by computing the iterate average of parameters using the most recent B iterations. To compute the y_k , the gradients of the objective function is estimated by using MC samples that is independent from the MC samples obtained to compute the gradient in the update step.

B.6. Natural Gradient Descent Optimization

Khan and Lin [31] proposed natural gradient descent (NGD) approach for variational inference that utilize the information geometry of the variational distribution. Given the variational distribution is a exponential family distribution

$$q_{\lambda} = h \exp(\phi^T \lambda - A(\lambda))$$

the NGD update is

$$\lambda^{(k+1)} \leftarrow \lambda^{(k)} - \gamma^{(k)} (F(\lambda^{(k)}))^{-1} \nabla_{\lambda^{(k)}} D_{\pi}(q_{\lambda^{(k)}}),$$

where $F(\lambda^{(k)})$ denote the Fisher Information Matrix (FIM) of the distribution and λ denote its natural-parameter. Without directly computing the inverse of FIM Khan and Lin [31] simplified the above update by using the relationship between natural parameter λ and expectation parameter $m = E_q[\phi]$ of the exponential family

$$F(\lambda)^{-1} \nabla_{\lambda} D_{\pi}(q_{\lambda}) = \nabla_m D_{\pi}(q_m).$$

Therefore, the natural-gradient update can be simplified as

$$\lambda^{(k+1)} \leftarrow \lambda^{(k)} - \gamma^{(k)} \nabla_{m^{(k)}} D_{\pi}(q_{m^{(k)}}).$$

This can be applied for the mean-field Gaussian variational family $q = \prod_{i=1}^d \mathcal{N}(\mu_i, \sigma_i^2)$ where we can define the natural parameters $\lambda_i = (\mu_i/\sigma_i^2, -0.5/\sigma_i^2)$ and expectation parameters $m = (\mu_i, \mu_i^2 + \sigma_i^2)$.

TABLE C.1
Datasets of PosteriorDB Package

Dataset	Short Name	Dimensions	Model Description
arK-arK	arK	7	AR(5) time series
bball_drive_event_0-hmm_drive_0	bball_0	8	Hidden Markov model
bball_drive_event_1-hmm_drive_1	bball_1	8	Hidden Markov model
diamonds-diamonds	diamonds	26	Log-Log
dogs-dogs	dogs	3	Logistic mixed-effects
dogs-dogs_log	dogs_log	2	Logarithmic mixed-effects
earnings-logearn_interaction	earnings	5	Log-linear
eight_schools-eight_schools_noncentered	8schools_nc	10	Non-centered hierarchical
eight_schools-eight_schools_centered	8schools_c	10	Centered hierarchical
garch-garch11	garch	4	GARCH(1,1) time series
gp_pois_regr-gp_pois_regr	gp_pois_regr	13	Gaussian process Poisson regression
gp_pois_regr-gp_regr	gp_regr	3	Gaussian process regression
hmm_example-hmm_example	hmm_example	6	Hidden Markov model
hudson_lynx-hare_lotka_volterra	hudson_lynx	8	Lotka-Volterra error
low_dim_gauss_mix-low_dim_gauss_mix	low_dim_gauss_mix	5	Two-dimensional Gaussian mixture
mcycle_gp_accel_gp	mcycle_gp	66	Gaussian process
nes2000-nes	nes2000	10	Multiple predictor linear
sblrc-blr	sblrc	6	Linear

Appendix C: Additional Experimental Details and Results

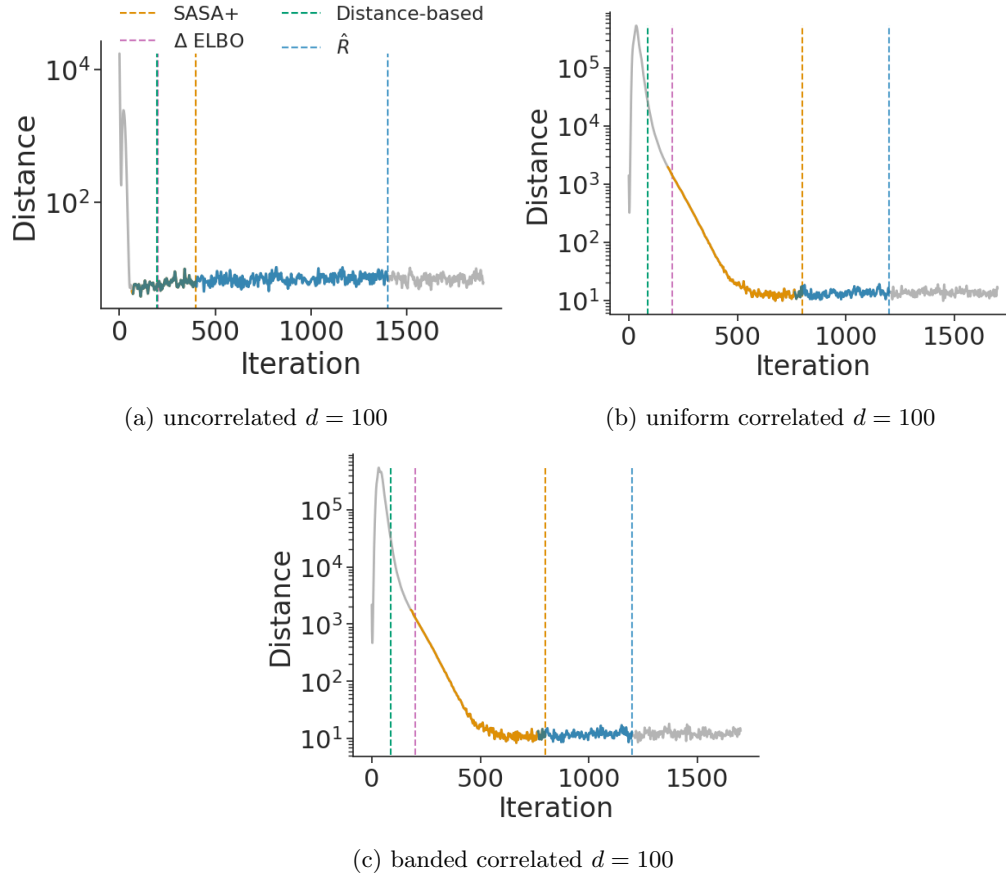


Fig C.1: Iteration number versus distance between iterate average and current iterate. The vertical lines indicate convergence detection trigger points and (for SASA+ and \hat{R}) the colored portion of the accuracy values indicate they are part of the window used for convergence detection.

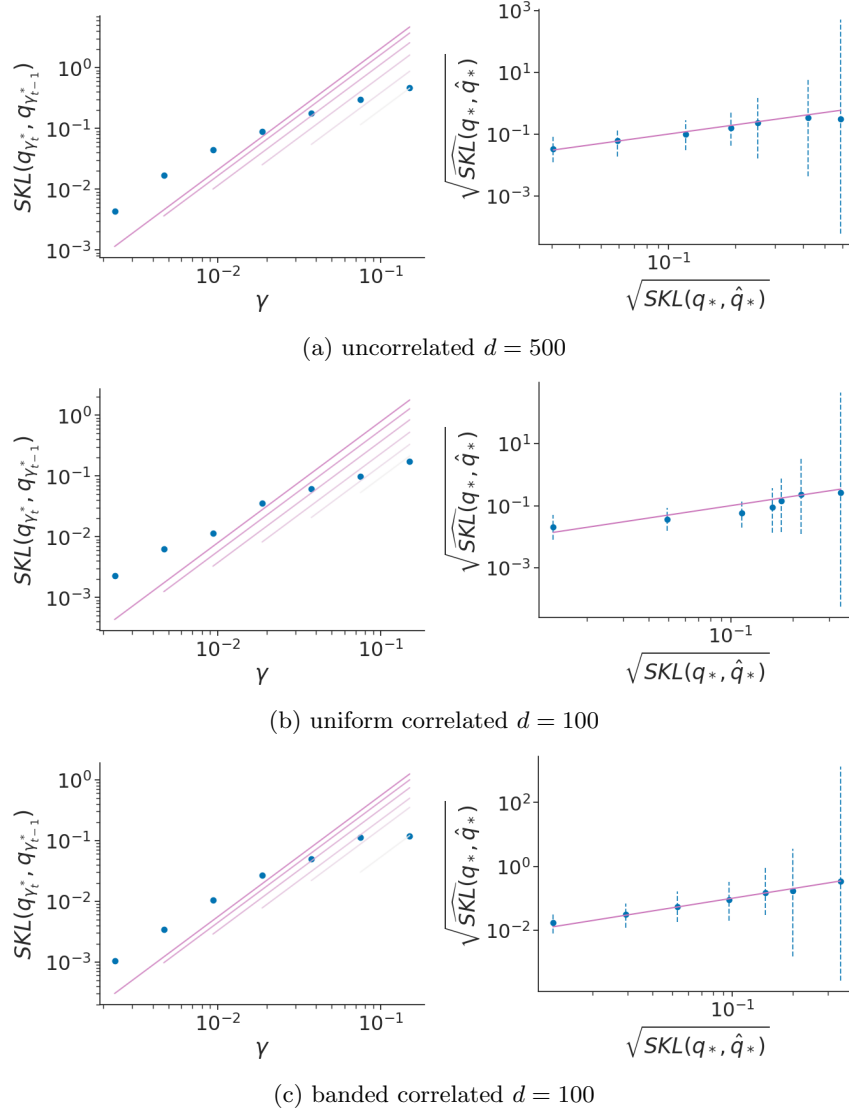


Fig C.2: Results for estimating the symmetrized KL divergence with avgAdam. **(left)** Learning rate versus symmetrized KL divergence of adjacent iterate averaged estimates of optimal variational distribution. The lines indicate the linear regression fits, with setting $\kappa = 1$. **(right)** Square root of true symmetrized KL divergence versus the estimated value with 95% credible interval. The uncertainty of the estimates decreases and remains well-calibrated as the learning rate decreases.

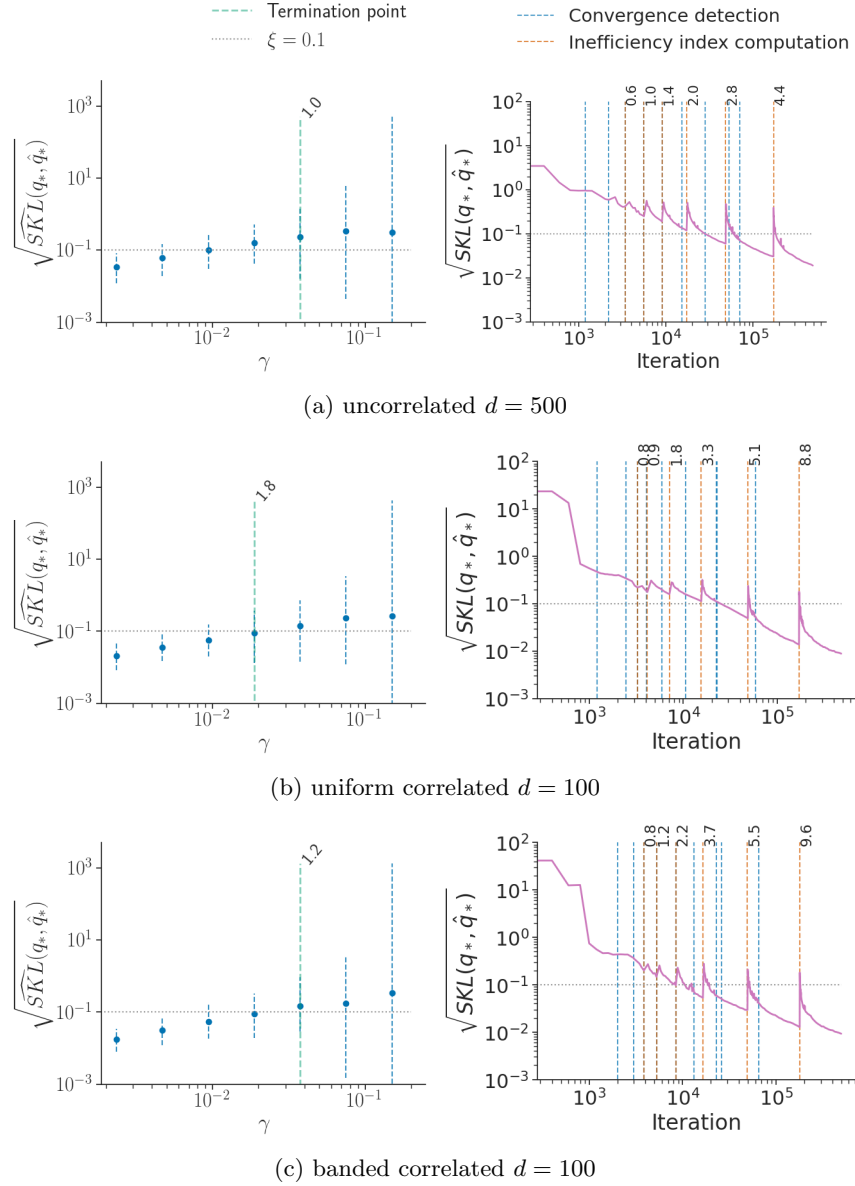


Fig C.3: Results of termination rule trigger point of Gaussian targets. **(left)** Learning rate versus square root of estimated symmetrized KL divergence with 95% credible interval (dashed blue line). The green vertical line indicates the termination rule trigger point with corresponding $\hat{\gamma}$ value. **(right)** Iterations versus square root of symmetrized KL divergence between iterate average and optimal variational approximation. The vertical lines indicate the convergence detection points using \hat{R} (blue) and inefficiency index computation (\hat{I}) points (orange) with corresponding values.

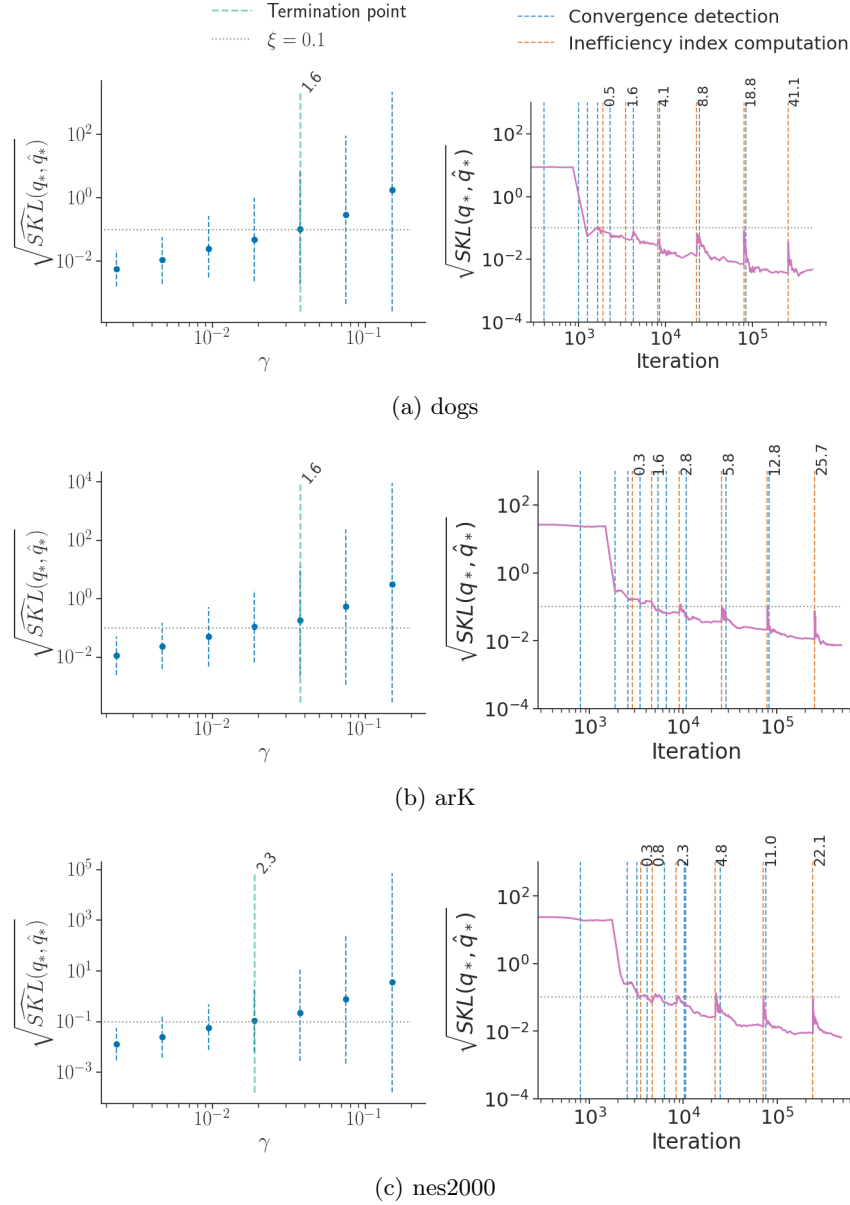
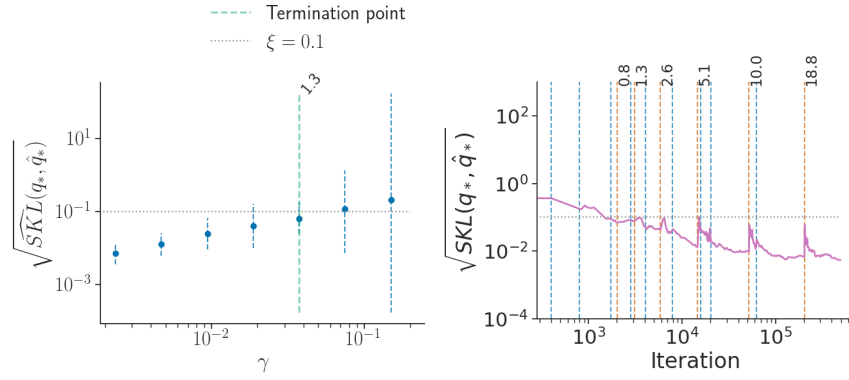
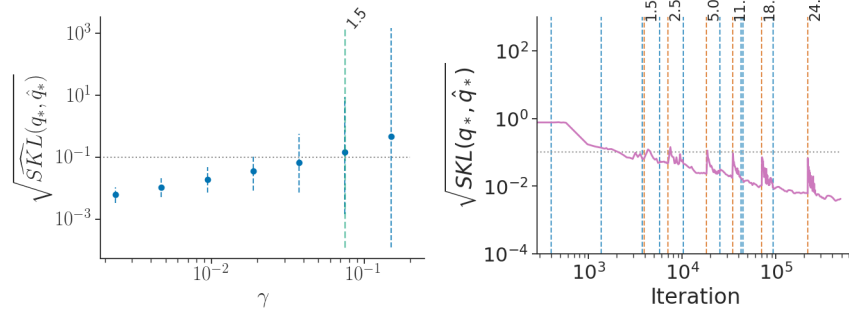


Fig C.4: Results of termination rule trigger point of `posteriorordb` package datasets. **(left)** Learning rate versus square root of estimated symmetrized KL divergence with 95% credible interval (dashed blue line). The green vertical line indicates the termination rule trigger point with corresponding \hat{T} value. **(right)** Iterations versus square root of symmetrized KL divergence between iterate average and optimal variational approximation. The vertical lines indicate the convergence detection points using \hat{R} (blue) and inefficiency index computation (\hat{T}) points (orange) with corresponding values.



(a) 8schools_nc



(b) 8schools_c

Fig C.5: Results of termination rule trigger point of `posteriordb` package datasets. **(left)** Learning rate versus square root of estimated symmetrized KL divergence with 95% credible interval (dashed blue line). The green vertical line indicates the termination rule trigger point with corresponding $\hat{\mathcal{I}}$ value. **(right)** Iterations versus square root of symmetrized KL divergence between iterate average and optimal variational approximation. The vertical lines indicate the convergence detection points using \hat{R} (blue) and inefficiency index computation ($\hat{\mathcal{I}}$) points (orange) with corresponding values.

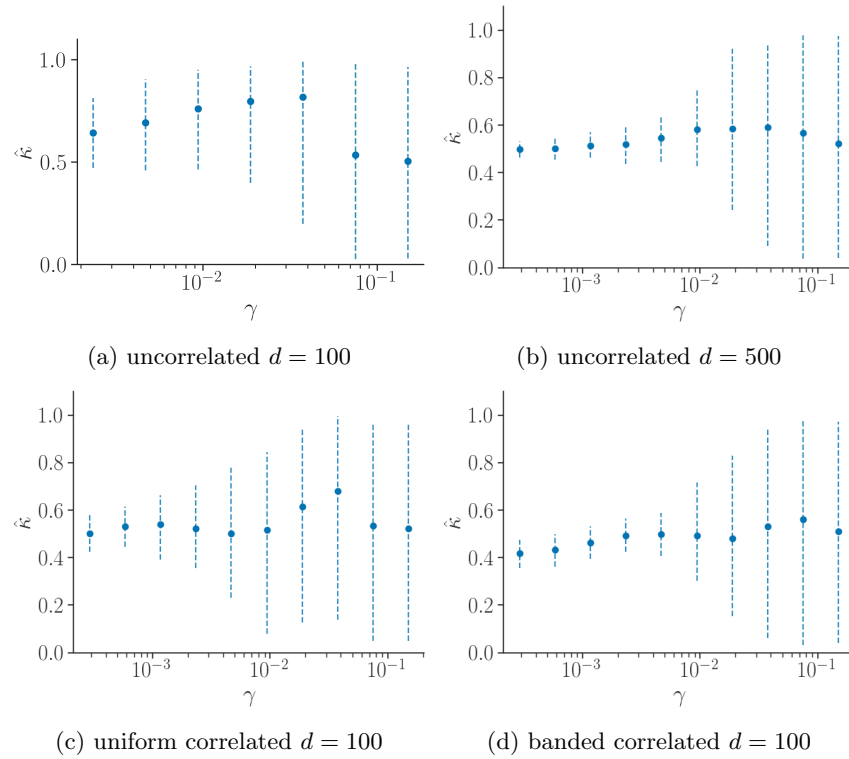


Fig C.6: Learning rate versus $\hat{\kappa}$ for Gaussian targets using adaptive SGD optimization (RMSProp) with 95% credible interval. The $\hat{\kappa}$ approximately reach to 0.5 at the final learning rate in all scenrios.

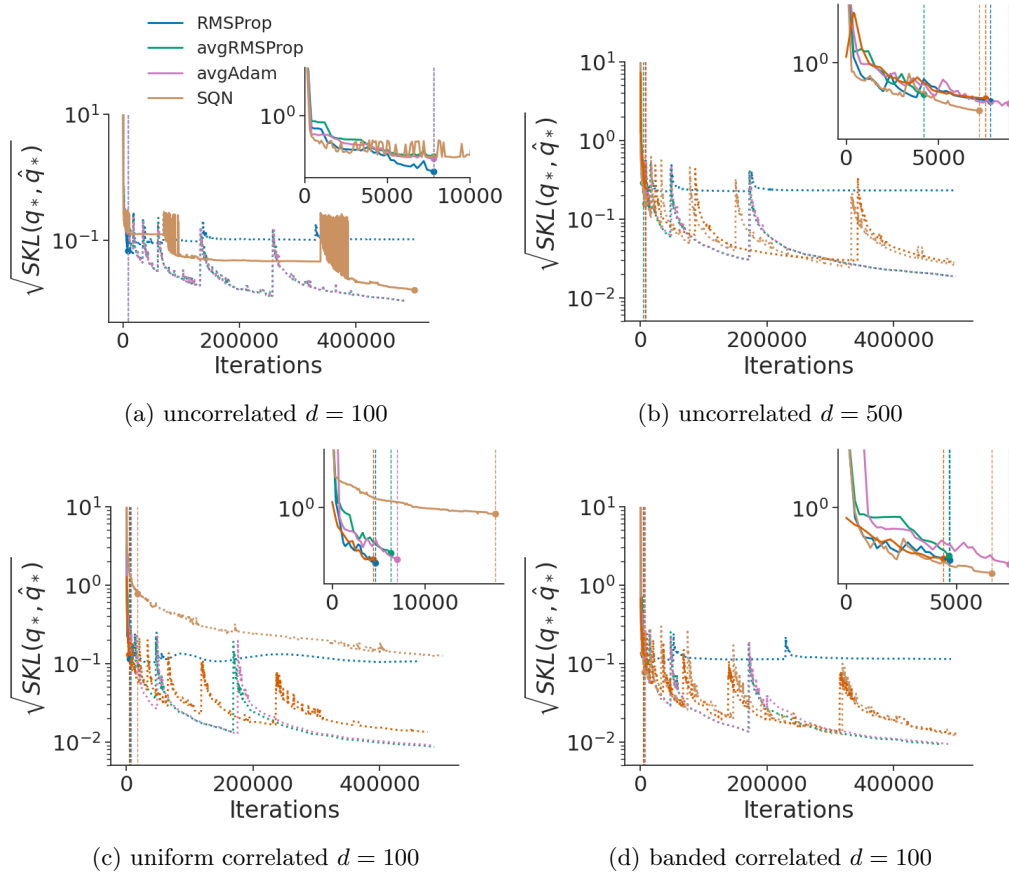


Fig C.7: Accuracy comparison of *RMSProp*, *avgRMSProp*, *avgAdam*, *SQN*, and *NGD* optimization methods in RAABBVI using Gaussian targets where accuracy is measured in terms of square root of symmetrized KL divergence between iterate average and optimal variational approximation. (*) The vertical lines indicate the termination rule trigger points and the behavior of optimization methods at the trigger points showed in zoom-in plots.

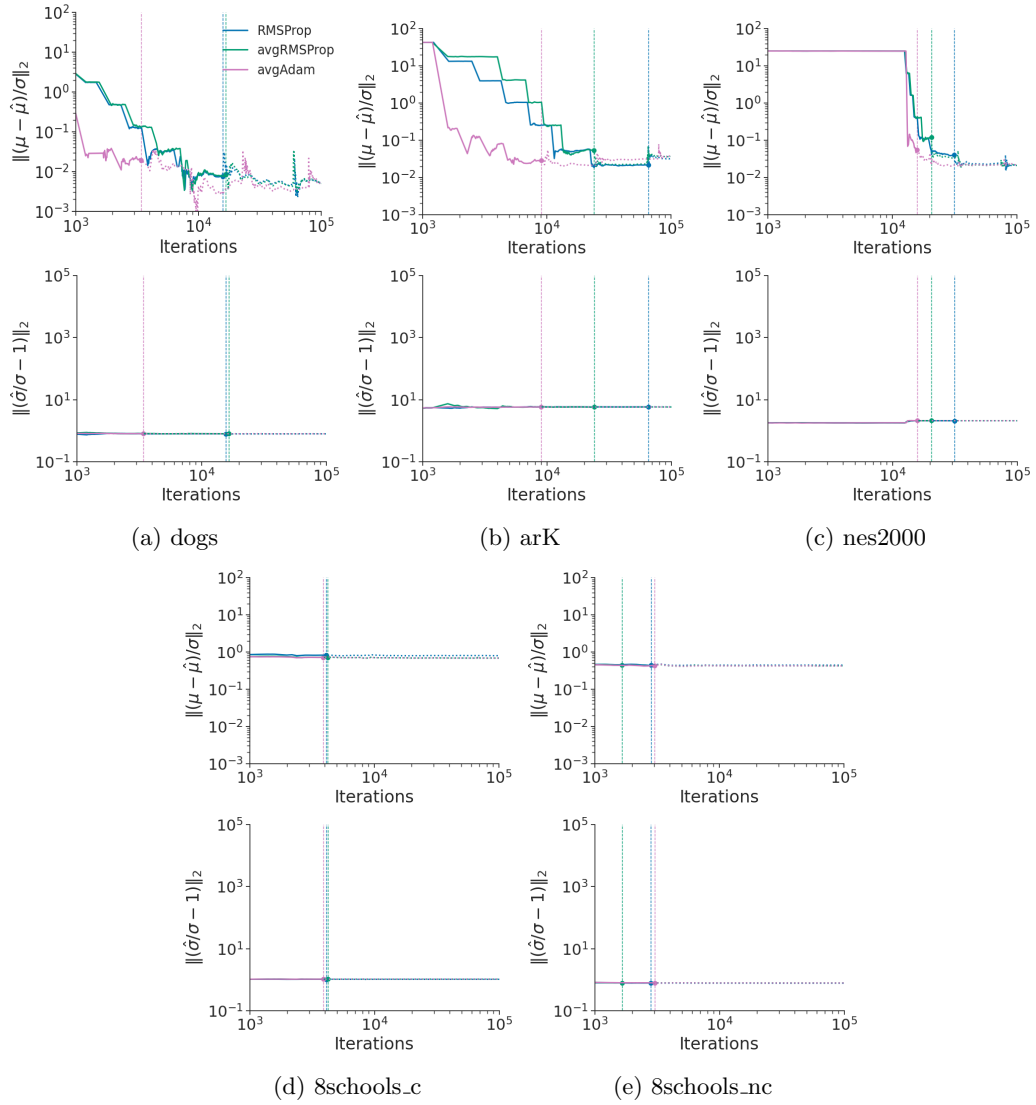


Fig C.8: Accuracy comparison of *RMSProp*, *avgRMSProp*, and *avgAdam* optimization methods in RAABVI using *posteriondb* datasets, where accuracy measured in terms of relative mean error (top) and relative standard deviation error (bottom). The vertical lines indicate the termination rule trigger points.

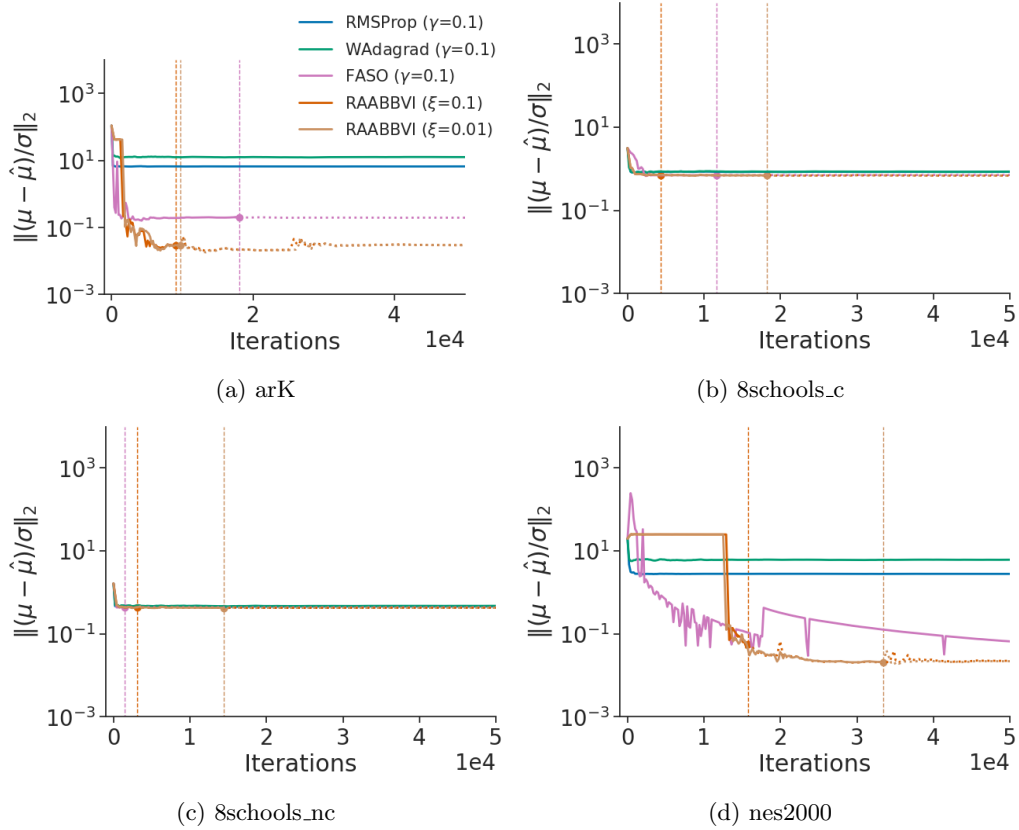


Fig C.9: Accuracy comparison of variational inference algorithms using `posteriordb` models and datasets, where accuracy is measure in terms of relative mean error. The vertical lines indicate the termination rule trigger points of FASO and RAABBVI. Iterate average for RMSProp and WAdagrad computed at every 200^{th} iteration using a window size of 20% of iterations.

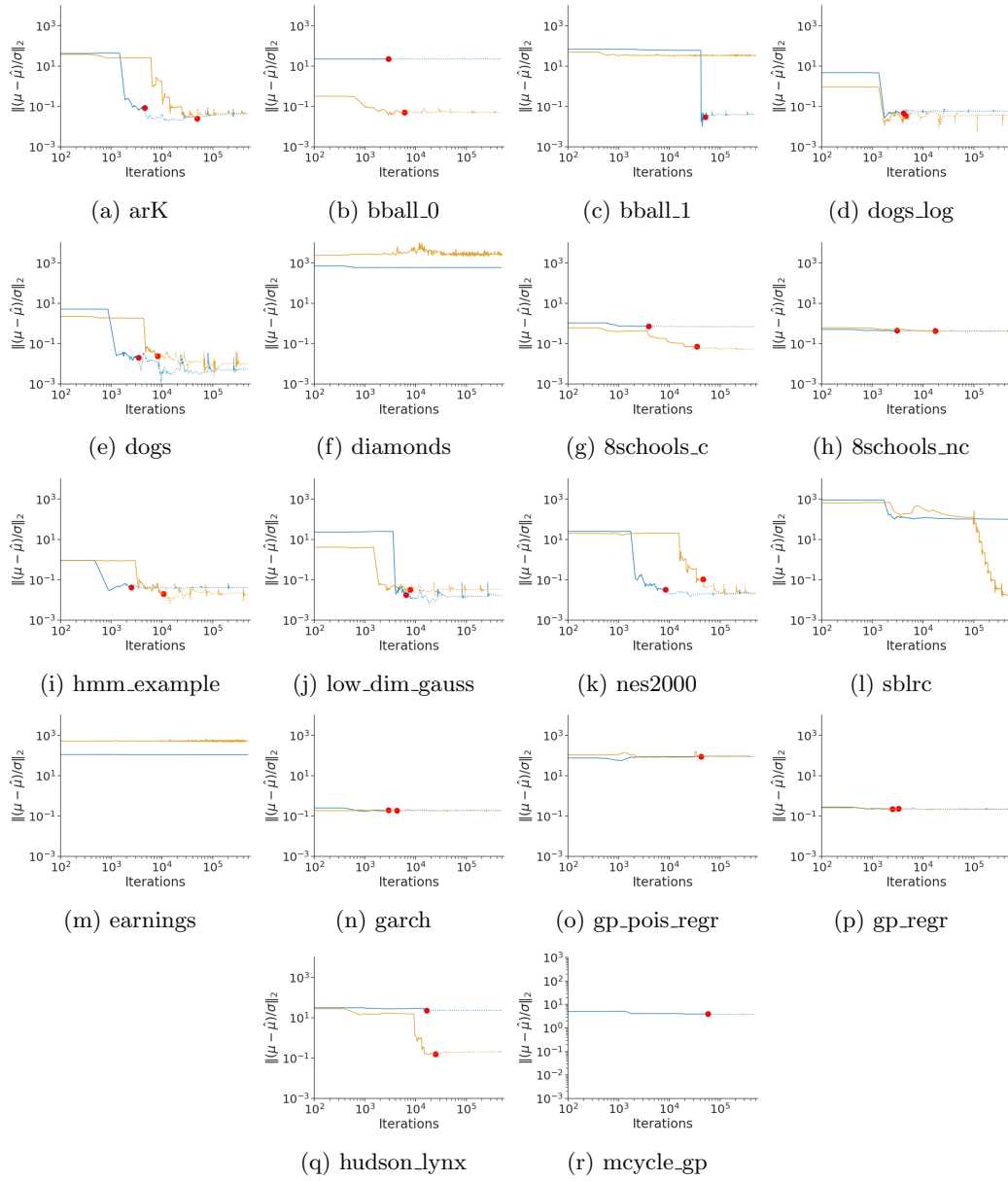


Fig C.10: Accuracy of mean-field (blue) and full-rank (orange) Gaussian family approximations for selected `posteriordb` data/models, where accuracy is measured in terms of relative mean error. The red dots indicate where the termination rule triggers

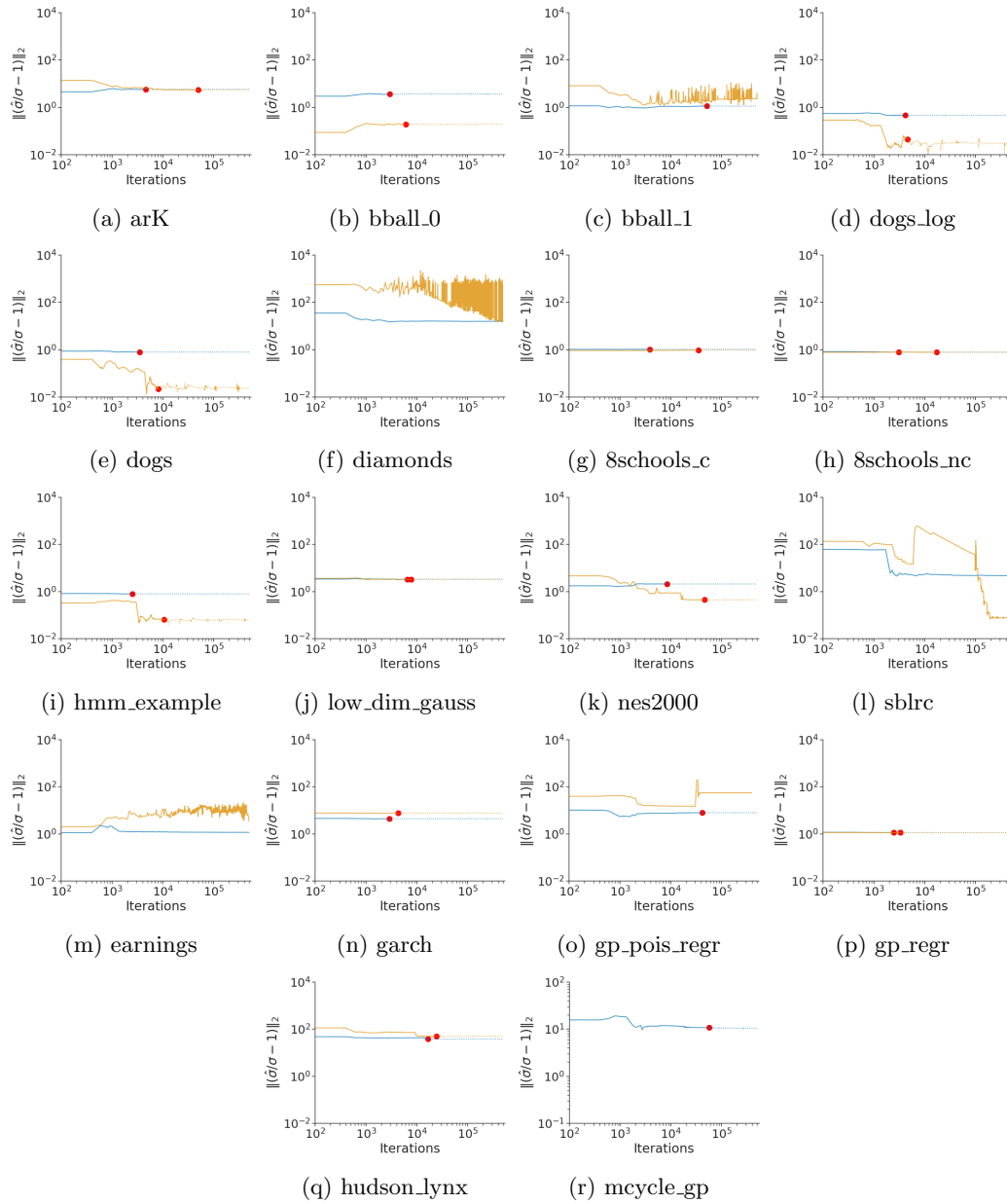


Fig C.11: Accuracy of mean-field (blue) and full-rank (orange) Gaussian family approximations for selected `posteriordb` data/models, where accuracy is measured in terms of relative standard deviation error. The red dots indicate where the termination rule triggers

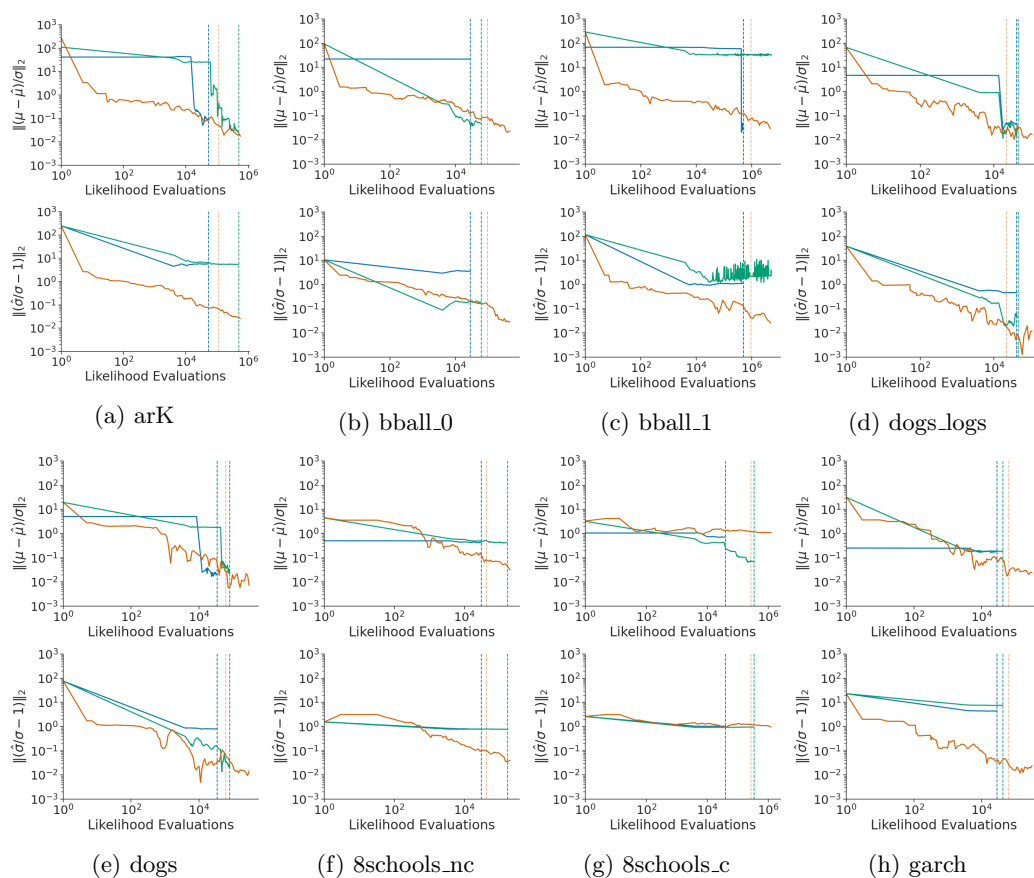


Fig C.12: Results of RAABBVI with mean-field Gaussian (blue) and full-rank Gaussian (green) family comparison to dynamic HMC (orange) in terms of relative mean error (top) and relative standard error (bottom). Blue and green vertical lines show the termination rule triggering points in RAABBVI and orange vertical line shows the end of warm-up period in dynamic HMC.

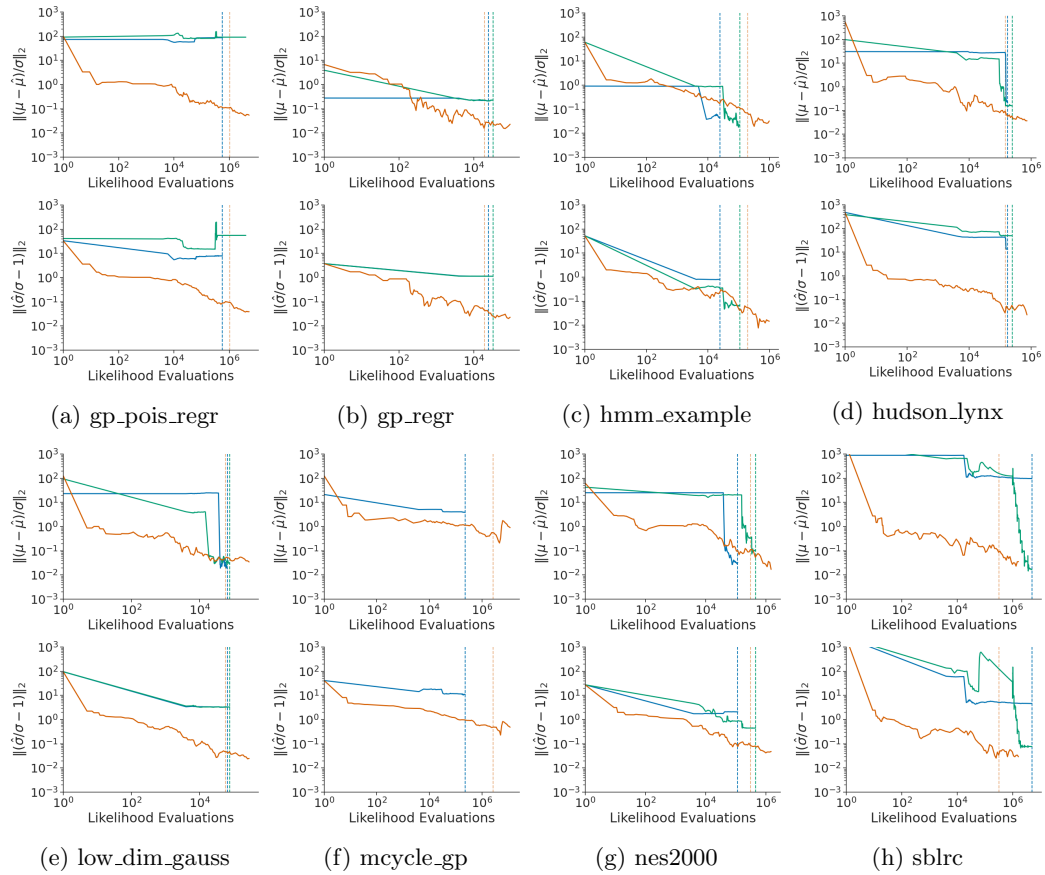


Fig C.13: Results of RAABBVI with mean-field Gaussian (blue) and full-rank Gaussian (green) family comparison to dynamic HMC (orange) in terms of relative mean error (top) and relative standard error (bottom). Blue and green vertical lines show the termination rule triggering points in RAABBVI and orange vertical line shows the end of warm-up period in dynamic HMC.

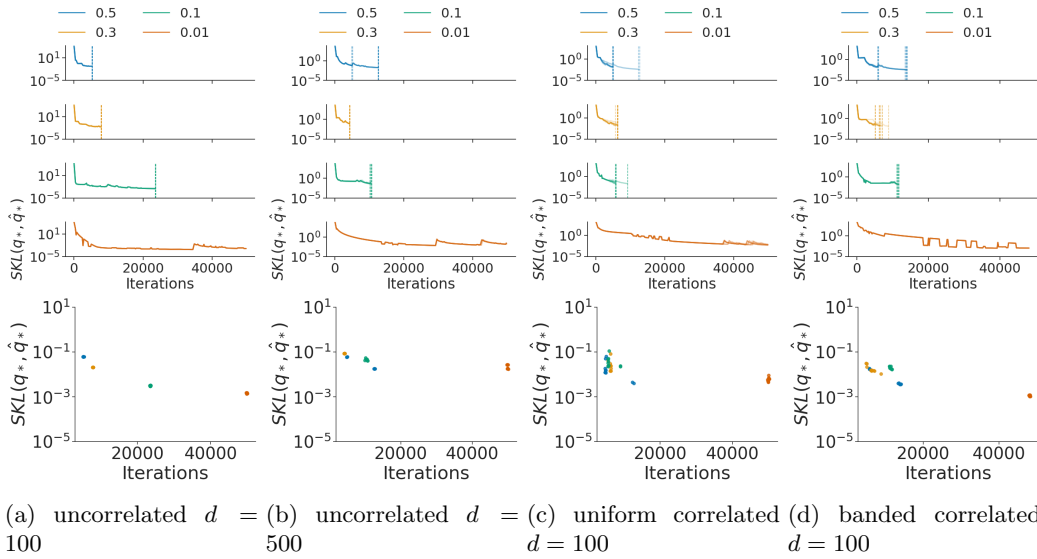


Fig C.14: Robustness to initial learning rate γ_0 using Gaussian targets. **(top)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation. The distinct lines represent repeated experiments and the vertical lines indicate the termination rule trigger points. **(bottom)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation at the termination rule trigger point.

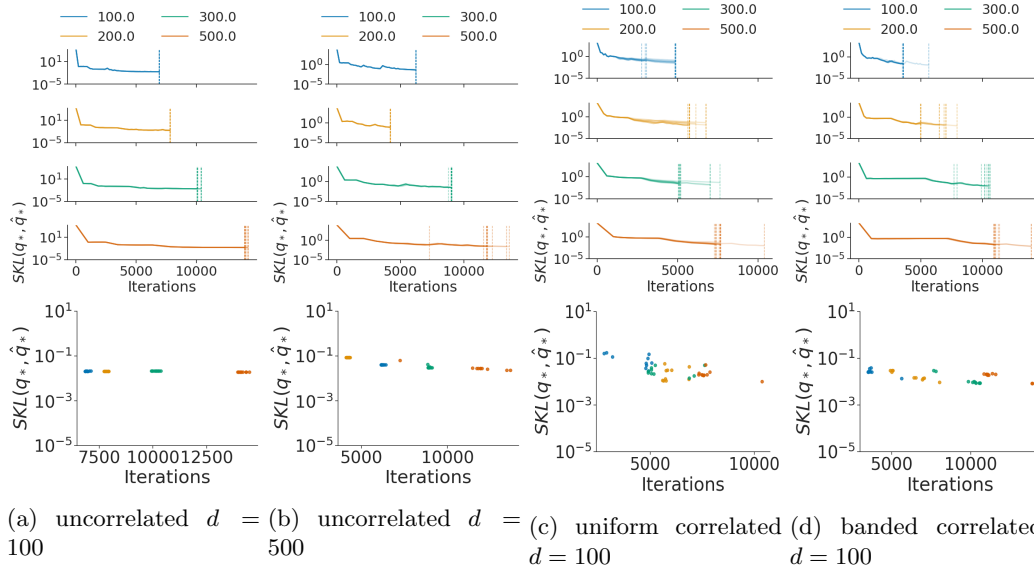


Fig C.15: Robustness to minimum window size $W_{min,,}$, using Gaussian targets. **(top)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation. The distinct lines represent repeated experiments and the vertical lines indicate the termination rule trigger points. **(bottom)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation at the termination rule trigger point.

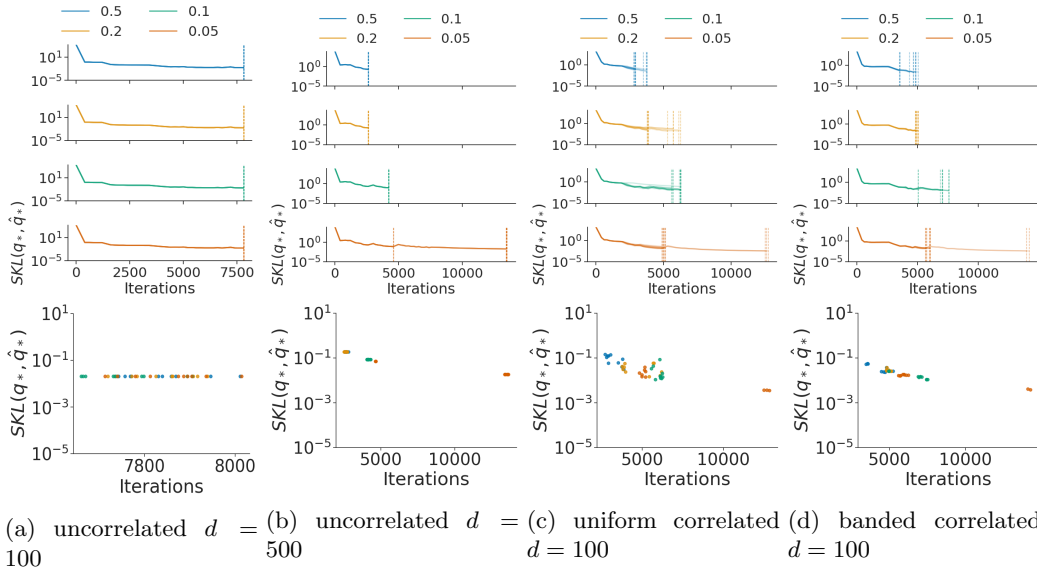


Fig C.16: Robustness to initial iterate average relative error threshold ε using Gaussian targets. **(top)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation. The distinct lines represent repeated experiments and the vertical lines indicate the termination rule trigger points. **(bottom)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation at the termination rule trigger point.

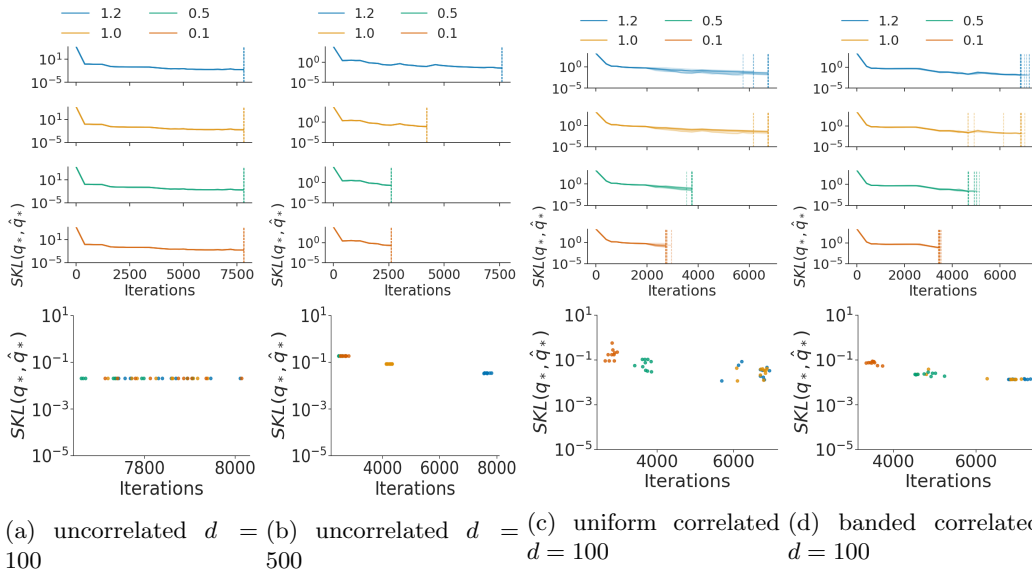


Fig C.17: Robustness to inefficiency threshold τ using Gaussian targets. **(top)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation. The distinct lines represent repeated experiments and the vertical lines indicate the termination rule trigger points. **(bottom)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation at the termination rule trigger point.

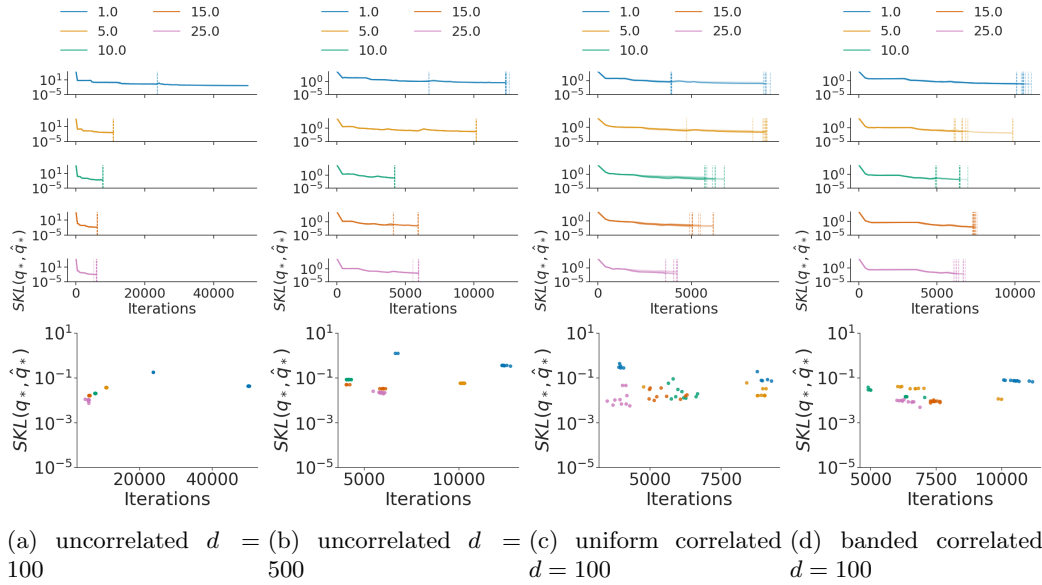


Fig C.18: Robustness to Monte Carlo samples M using Gaussian targets. **(top)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation. The distinct lines represent repeated experiments and the vertical lines indicate the termination rule trigger points. **(bottom)** Iterations versus symmetrized KL divergence between iterate average and optimal variational approximation at the termination rule trigger point.