Black Box Variational Inference with a Deterministic Objective: Faster, More Accurate, and Even More Black Box

Ryan Giordano,* Martin Ingram,* Tamara Broderick

January 16, 2024

Abstract

Automatic differentiation variational inference (ADVI) offers fast and easy-to-use posterior approximation in multiple modern probabilistic programming languages. However, its stochastic optimizer lacks clear convergence criteria and requires tuning parameters. Moreover, ADVI inherits the poor posterior uncertainty estimates of mean-field variational Bayes (MFVB). We introduce "deterministic ADVI" (DADVI) to address these issues. DADVI replaces the intractable MFVB objective with a fixed Monte Carlo approximation, a technique known in the stochastic optimization literature as the "sample average approximation" (SAA). By optimizing an approximate but deterministic objective, DADVI can use off-the-shelf second-order optimization, and, unlike standard mean-field ADVI, is amenable to more accurate posterior covariances via linear response (LR). In contrast to existing worst-case theory, we show that, on certain classes of common statistical problems, DADVI and the SAA can perform well with relatively few samples even in very high dimensions, though we also show that such favorable results cannot extend to variational approximations that are too expressive relative to mean-field ADVI. We show on a variety of real-world problems that DADVI reliably finds good solutions with default settings (unlike ADVI) and, together with LR covariances, is typically faster and more accurate than standard ADVI.

1 Introduction

The promise of "black-box" Bayesian inference methods is that the user need provide only a model and data. Then the black-box method should take care of approximating the posterior distribution and reporting any summaries of interest to the user. In settings where Markov chain Monte Carlo (MCMC) faces prohibitive computational costs, users of Bayesian inference have increasingly turned to variational methods. In turn, to improve ease of use in these cases, researchers have developed a variety of "black-box variational inference" (BBVI) methods [Ranganath et al., 2014, Blei et al., 2017]. "Automatic differentiation variational inference" (ADVI) represents a particularly widely used variant of BBVI [Kucukelbir et al., 2017], available in multiple modern probabilistic programming languages.

* These authors contributed equally.

- The code to reproduce this paper can be found at `https://github.com/rgiordan/DADVIPaper`.

- Our Python implementation of DADVI can be found at `https://github.com/martiningram/dadvi`.

However, researchers have observed that BBVI methods can face challenges with both automation [Dhaka et al., 2020, Welandawe et al., 2022] and accuracy (MacKay, 2003, Exercise 33.5; Bishop, 2006, Chapter 10.1.2; Turner and Sahani, 2011; Huggins et al., 2020, Propositions 3.1–3.3). In particular, BBVI takes an optimization-based approach to approximate Bayesian inference. The optimization objective in a typical BBVI method involves an intractable expectation over the approximating distribution. Most BBVI algorithms, including ADVI, avoid computing the intractable expectation by using stochastic gradient (SG) optimization, which requires only unbiased draws from the gradient of the intractable objective. However, the use of SG is not without a price: SG requires careful tuning of the step size schedule, can suffer from poor conditioning, and convergence can be difficult to assess. On the accuracy side, observe that ADVI minimizes the reverse Kullback-Leibler (KL) divergence over Gaussian approximating distributions. The especially common mean-field variant of this scheme, where the Gaussians are further constrained to fully factorize, notoriously produces poor posterior covariance estimates (MacKay, 2003, Exercise 33.5; Bishop, 2006, Chapter 10.1.2; Turner and Sahani, 2011), and research suggests variants beyond mean-field may suffer as well [Huggins et al., 2020, Proposition 3.2]. In many cases, these posterior covariance estimates can be efficiently corrected, without fitting a more complex approximation, through a form of sensitivity analysis known as "linear response" (LR) [Giordano et al., 2015, 2018]. However, LR cannot be used directly with SG, both because the optimum is only a rough approximation and because the objective function itself is intractable.

The stochastic optimization literature offers a well-studied alternative to SG: the "sample average approximation" (SAA), which uses a single set of draws — shared across all iterations — to approximate an intractable expected objective. See Kim et al. [2015] for a review of the SAA. In fact, a number of papers have applied SAA to BBVI [Giordano et al., 2018, Domke and Sheldon, 2018, 2019, Broderick et al., 2020, Wycoff et al., 2022, Giordano et al., 2023]. But before the present work and contemporaneous work by Burroni et al. [2023], there had not yet been a systematic study of the efficacy of SAA for BBVI. Burroni et al. [2023] chooses an increasing sequence of sample sizes in SAA, applied to variational inference with the full-rank Gaussian approximation family, in order to achieve an increasingly accurate approximation to the exact variational objective. In a complementary vein, we here instead explore the promise and challenges of using SAA in BBVI with a small, fixed number of samples — with a focus on both automation and accuracy. We call our method "DADVI" for "deterministic ADVI," and we use the unmodified "ADVI" to refer to the ADVI variational approximation optimized with SG.

When considering a general optimization problem, the case for SAA over SG may at first look weak. In full generality, SAA and SG require roughly the same number of draws, N, for a particular accuracy. And the total number of draws required for a given accuracy is expected to increase linearly in dimension [Nemirovski et al., 2009, Shapiro et al., 2021, Chapter 5]. Since SG uses each draw only once, and SAA uses each draw at each step of a multi-step optimization routine, SAA is, all else equal, expected to require more computation than SG in the worst-case scenario, particularly in high dimensions [Royset and Szechtman, 2013, Kim et al., 2015]. However, results in particular cases can be quite different than these general conclusions.

We demonstrate that the SAA can be competitive with SG in BBVI applications both theoretically and in experiments using real-world models and datasets. Theoretically, we

consider two cases common in Bayesian inference: (1) log posteriors that are approximately quadratic, and (2) posteriors that have a "global–local" structure: roughly, there are some (global) parameters of fixed dimension as the data set size grows, and some (local) parameters whose dimension grows with the data cardinality. We further assume, as is typically the case, that the user is interested in a relatively small number of quantities of interest that are specified in advance, as opposed to, say, the maximum value of a high-dimensional vector of posterior means. In these cases, our theory shows that DADVI does not suffer from the worst-case dimensional dependence that the classical SAA literature suggests. In our experiments, we show that DADVI produces competitive posterior approximations in very high-dimensional problems, even with only $N = 30$ draws, and even in models more complex than the cases that we analyze theoretically. Notably, in high dimensions, LR covariances are considerably more computationally efficient — and more accurate — than fitting a more complex variational approximation, such as a full-rank normal. To our knowledge, the advantages of SAA for performing sensitivity analysis, either within or beyond Bayesian inference, have not been widely recognized.

Conversely, we show that SAA is not applicable to all BBVI methods. For example, we show that, when using a full-rank ADVI approximation in high dimensions, the SAA approximation leads to a degenerate variational objective unless the number of draws used is very high — on the order of the number of parameters. The intuition behind how SAA fails in such a case applies to other highly expressive BBVI approximations such as normalizing flows [Rezende and Mohamed, 2015]. In high dimensions, it is thus a combination of the relative paucity of the mean-field ADVI approximation, together with special problem structure, that makes DADVI a useful tool. Nevertheless, such cases are common enough that the benefits of DADVI remain noteworthy.

In what follows, we start by reviewing ADVI (Section 2) and describing DADVI (Section 3), our SAA approximation. We highlight how DADVI, unlike ADVI, allows the use of LR covariances (Section 3.1). In Section 3.2, we demonstrate how to approximately quantify DADVI's Monte Carlo error, which arises from the single set of Monte Carlo draws, and we note that such a quantification is not readily available for ADVI due to its use of SG. We provide theory to support why DADVI can be expected to work in certain classes of high-dimensional problems (Sections 4.1 and 4.2), and we provide a counterexample to demonstrate how DADVI can fail with very expressive BBVI approximations (Section 4.3). In a range of real-world examples (Section 6.1), we show that DADVI inherits the generally recognized advantages of SAA, including the availability of off-the-shelf higher-order optimization and reliable convergence assessment. We find experimentally that DADVI, paired with LR covariances, can provide comparable posterior mean estimates and more accurate posterior uncertainties (Section 6.3) with less computation than corresponding ADVI methods (Section 6.2), including recent work that endeavors to improve and automate the tuning of SG for BBVI [Welandawe et al., 2022], and we show that our estimates of Monte Carlo sampling variability are accurate even for small values of N, around 30 (Section 6.5).

2 Setup

In what follows, we take data y and a finite-dimensional parameter $\theta \in \Omega_\theta$. We consider a user who is able to provide software implementations of the log density of the joint distribution

$\mathcal{P}(y, \theta)$ and is interested in reporting means and variances from an approximation of the exact Bayesian posterior $\mathcal{P}(\theta|y)$.

Black-box variational inference (BBVI) refers to a spectrum of approaches for approximating this posterior. We focus in the present paper on ADVI, a particularly popular instance of BBVI. Variational inference forms an approximation $\mathcal{Q}(\theta|\eta)$, with variational parameters $\eta \in \Omega_\eta$, to $\mathcal{P}(\theta|y)$. Let $\mathcal{N}(\cdot|\mu, \Sigma)$ denote a normal distribution with mean μ and covariance matrix Σ. The full-rank variant of ADVI approximately minimizes the reverse KL divergence $\mathrm{KL}\left(\mathcal{Q}(\cdot|\eta)||\mathcal{P}(\cdot|y)\right)$ between the exact posterior and an approximating family of multivariate normal distributions:

$$\Omega_{\mathcal{Q}} = \left\{\mathcal{Q}(\theta|\eta) : \mathcal{Q}(\theta|\eta) = \mathcal{N}\left(\theta|\mu(\eta), \Sigma(\eta)\right)\right\}, \tag{1}$$

where $\eta \mapsto (\mu(\eta), \Sigma(\eta))$ is a (locally) invertible map between the space of variational parameters and the mean and covariance of the normal distribution. When we optimize η over this family, we will refer to the resulting optimization problem as the "full-rank ADVI optimization problem."

In particular, we will typically focus on the following objective function, which is equivalent to the one above:

$$\mathscr{L}_{\mathrm{VI}}(\eta) := \mathop{\mathbb{E}}_{\mathcal{Q}(\theta|\eta)}\left[\log \mathcal{Q}(\theta|\eta)\right] - \mathop{\mathbb{E}}_{\mathcal{Q}(\theta|\eta)}\left[\log \mathcal{P}(\theta, y)\right]. \tag{2}$$

The objective $\mathscr{L}_{\mathrm{VI}}(\eta)$ in Equation (2) is equivalent to the KL divergence $\mathrm{KL}\left(\mathcal{Q}(\cdot|\eta)||\mathcal{P}(\cdot|y)\right)$ up to $\log \mathcal{P}(y)$, which does not depend on η, so that minimizing $\mathscr{L}_{\mathrm{VI}}(\eta)$ also minimizes the KL divergence. The negative of the objective, $-\mathscr{L}_{\mathrm{VI}}(\eta)$, is sometimes called the "evidence lower bound" (ELBO) [Blei et al., 2017].

To avoid degeneracy in the objective, typically one transforms any model parameters with restricted range before running the optimization — and performs the reverse transformation after. E.g., we might take the logarithm of any strictly positive parameters so that their transformed range is the full real line; see Kucukelbir et al. [2017] for further details. Therefore, we will henceforth assume that $\Omega_\theta = \mathbb{R}^{D_\theta}$ and $\mathcal{P}(\theta)$ is supported on all $\mathbb{R}^{D_\theta}$. Then in the full-rank case, η contains both the mean and some unconstrained representation of a D_θ-dimensional covariance matrix, so that $\eta \in \mathbb{R}^{D_\theta + D_\theta(D_\theta+1)/2}$.

The mean-field variant of ADVI restricts $\Sigma(\eta)$ to be diagonal.[1] That is, take the approximating family $\Omega_{\mathcal{Q}}$ to consist of independent normals with means $\mu \in \mathbb{R}^{D_\theta}$ and log standard deviations $\xi \in \mathbb{R}^{D_\theta}$.

$$\Omega_{\mathcal{Q}} = \left\{\mathcal{Q}(\theta|\eta) : \mathcal{Q}(\theta|\eta) = \prod_{d=1}^{D_\theta} \mathcal{N}\left(\theta_d|\mu_d, \exp(\xi_d)^2\right)\right\} \tag{3}$$
$$\mu = (\mu_1, \ldots, \mu_{D_\theta})^\intercal,\ \xi = (\xi_1, \ldots, \xi_{D_\theta})^\intercal,\ \eta = (\mu^\intercal, \xi^\intercal)^\intercal,\ \Omega_\eta = \mathbb{R}^{D_\eta}, \text{ and } D_\eta = 2D_\theta.$$

For mean-field ADVI, the variational parameter $\eta^\intercal = (\mu^\intercal, \xi^\intercal) \in \mathbb{R}^{2D_\theta}$. When the variational family satisfies the mean-field assumption, we will refer to the resulting optimization problem as the "mean-field ADVI optimization problem" and its objective as the "mean-field ADVI objective."

[1] Technically, in the mean-field variant of ADVI, $\Sigma(\eta)$ may sometimes be block diagonal; see Appendix A. We elide this special case in what follows for ease of exposition. Our experiments are fully diagonal.

Using the expression for univariate normal entropy and neglecting some constants, the mean-field ADVI objective in Equation (2) becomes

$$\mathscr{L}_{\mathrm{VI}}(\eta) := -\sum_{d=1}^{D_\theta} \xi_d - \mathop{\mathbb{E}}_{\mathcal{N}(\theta|\eta)} [\log \mathcal{P}(\theta, y)] \quad \text{and} \quad \overset{*}{\eta} := \mathop{\mathrm{argmin}}_{\eta \in \Omega_\eta} \mathscr{L}_{\mathrm{VI}}(\eta). \tag{4}$$

We would ideally like to compute $\overset{*}{\eta}$, but we cannot optimize $\mathscr{L}_{\mathrm{VI}}(\eta)$ directly, because the term $\mathop{\mathbb{E}}_{\mathcal{N}(\theta|\eta)} [\log \mathcal{P}(\theta, y)]$ is generally intractable. ADVI, like most current BBVI methods, employs stochastic gradient optimization (SG) to avoid computing $\mathscr{L}_{\mathrm{VI}}(\eta)$. Specifically, ADVI uses Monte Carlo and the "reparameterization trick" [Mohamed et al., 2020] as follows. Let $\mathcal{N}_{\mathrm{std}}(Z)$ denote the D_θ-dimensional standard normal distribution. If $Z \sim \mathcal{N}_{\mathrm{std}}(Z)$, then

$$\mathop{\mathbb{E}}_{\mathcal{N}(\theta|\eta)} [\log \mathcal{P}(\theta, y)] = \mathop{\mathbb{E}}_{\mathcal{N}_{\mathrm{std}}(Z)} [\log \mathcal{P}(\mu + Z \odot \exp(\xi), y)]. \tag{5}$$

For compactness, we write

$$\theta(\eta, Z) := \mu + Z \odot \exp(\xi), \tag{6}$$

where $\odot$ is the component-wise (Hadamard) product. For N independent draws[2] $\mathscr{Z} := \{Z_1, \ldots, Z_N\}$ from $\mathcal{N}_{\mathrm{std}}(Z)$, we can use Equation (5) to define an unbiased estimate for the mean-field $\mathscr{L}_{\mathrm{VI}}(\eta)$:

$$\widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z}) := -\sum_{d=1}^{D_\theta} \xi_d - \frac{1}{N} \sum_{n=1}^{N} \log \mathcal{P}(\theta(\eta, Z_n), y). \tag{7}$$

ADVI uses derivatives of $\widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z})$, with a new draw of $\mathscr{Z}$ at each iteration, to estimate $\overset{*}{\eta}$. The ADVI algorithm, which we will sometimes refer to as "ADVI" in shorthand, can be found in Algorithm 1. When we use the ADVI algorithm for the full-rank optimization problem, we will write "full-rank ADVI."

[2] A subscript Z_n will denote a particular member of the set $\mathscr{Z}$, though for the rest of the paper, subscripts will usually denote an entry of a vector.

Algorithm 1 ADVI (Existing method)

procedure ADVI
  $t \leftarrow 0$
  Fix N (typically $N = 1$)
  while Not converged **do**
    $t \leftarrow t + 1$
    Draw $\mathscr{Z}$
    $\Delta \leftarrow \nabla_\eta \widehat{\mathscr{L}_{\text{VI}}}(\eta_{t-1}|\mathscr{Z})$
    $\alpha_t \leftarrow$ SetStepSize(All past states)
    $\eta_t \leftarrow \eta_{t-1} - \alpha_t \Delta$
    AssessConvergence(All past states)
  end while
  $\tilde{\eta} \leftarrow \eta_t$ or $\tilde{\eta} \leftarrow \frac{1}{M}\sum_{t'=t-M+1}^{t} \eta_{t'}$
  return $\mathcal{Q}(\theta|\tilde{\eta})$
end procedure
Postprocessing (If possible)
  Assess MC error using $\eta_1, \ldots, \eta_t$
  if MC Error is too high **then**
    Re-run with smaller / more steps
  end if

Algorithm 2 DADVI (Our proposal)

procedure DADVI
  $t \leftarrow 0$
  Fix N (our default is $N = 30$)
  Draw $\mathscr{Z}$
  while Not converged **do**
    $t \leftarrow t + 1$
    $\Delta \leftarrow$ GetStep($\widehat{\mathscr{L}_{\text{VI}}}(\cdot|\mathscr{Z}), \eta_{t-1}$)
    $\eta_t \leftarrow \eta_{t-1} + \Delta$
    AssessConvergence($\widehat{\mathscr{L}_{\text{VI}}}(\cdot|\mathscr{Z}), \eta_t$)
  end while
  $\hat{\eta} \leftarrow \eta_t$
  return $\mathcal{Q}(\theta|\hat{\eta})$
end procedure
Postprocessing
  Compute LR covariances (Section 3.1)
  Assess MC error (Section 3.2)
  if MC Error is too high **then**
    Re-run with more samples in $\mathscr{Z}$
  end if

3 Our Method

Our method, DADVI, will start from the same optimization objective as ADVI, but it will use a different approximation to handle the intractable objective. As we have seen, in ADVI, each step of the optimization draws a new random variable. The key difference in our method, DADVI, is that the random approximation is instead made with a single set of draws and then fixed throughout optimization. The full DADVI algorithm appears in Algorithm 2. In the notation of Equation (7), for a particular $\mathscr{Z}$, the value $\hat{\eta}$ returned by DADVI in Algorithm 2 is given by

$$\hat{\eta} := \underset{\eta \in \Omega_\eta}{\text{argmin}} \widehat{\mathscr{L}_{\text{VI}}}(\eta|\mathscr{Z}). \tag{8}$$

The $\hat{\eta}$ of Equation (8) is an estimate of η^* insofar as its objective $\widehat{\mathscr{L}_{\text{VI}}}(\eta)$ is a random approximation to the true objective $\mathscr{L}_{\text{VI}}(\eta)$. In general, the idea of DADVI can be applied to either the mean-field or full-rank ADVI optimization problem (though see Section 4.3 below for some potential challenges when using DADVI with full-rank ADVI). In what follows, analogously to how we refer to the ADVI algorithm, we will assume that we are targeting the mean-field problem with DADVI unless explicitly stated that we are instead targeting the full-rank problem.

For DADVI, the reparameterization of Equation (5) is essential: it allows us to use the same set of draws $\mathscr{Z}$ for any value of η. With $\mathscr{Z}$ fixed, $\widehat{\mathscr{L}_{\text{VI}}}(\cdot|\mathscr{Z})$ in turn remains fixed

throughout optimization. This consistency would not be possible in general without a reparameterization like Equation (5) to separate the stochasticity from the shape of $\mathcal{Q}(\theta|\eta)$.

Note that, for a given $\mathscr{Z}$, all derivatives of $\widehat{\mathscr{L}_{\text{VI}}}(\eta|\mathscr{Z})$ required by either DADVI or ADVI can be computed using automatic differentiation and a software implementation of $\log \mathcal{P}(\theta, y)$. In this sense, both DADVI and ADVI are black-box methods. In practice, another key difference between ADVI and DADVI is that ADVI typically draws only a single random variable per iteration, whereas DADVI uses a larger number of draws; in particular, the default number of draws for DADVI in our experiments will be $N = 30$.

We will see in what follows that using DADVI instead of ADVI can reap large practical benefits.

3.1 Linear response covariances

We next review linear response (LR) covariances as an approximation for posterior covariances of interest. We then show how DADVI accommodates LR covariances in a way that ADVI does not. The key observation is that, since ADVI does not actually minimize a tractable objective, sensitivity measures such as LR covariances are not available, though they are for DADVI. To the authors' knowledge, the availability of such sensitivity measures for SAA but not SG is not yet a widely recognized advantage of SAA.

One well-documented failure of mean-field variational Bayes approximations (including mean-field ADVI) is the mis-estimation of posterior variance [Bishop, 2006, Turner and Sahani, 2011, Giordano et al., 2018, Margossian and Saul, 2023]. Even in cases for which mean-field approximations provide good approximations to posterior means (e.g. when a Bayesian central limit theorem can be approximately applied), the posterior variances are typically incorrect. Formally, we often find that, for some quantity of interest $\phi(\theta) \in \mathbb{R}$,

$$\mathop{\mathbb{E}}_{\mathcal{Q}(\theta|\eta^*)}[\phi(\theta)] \approx \mathop{\mathbb{E}}_{\mathcal{P}(\theta|y)}[\phi(\theta)] \quad \text{but} \quad \left| \mathop{\text{Var}}_{\mathcal{Q}(\theta|\eta^*)}(\phi(\theta)) - \mathop{\text{Var}}_{\mathcal{P}(\theta|y)}(\phi(\theta)) \right| \gg 0. \tag{9}$$

A classical motivating example is the case of multivariate normal posteriors, which we review in Section 4.1.

LR covariances comprise a technique for ameliorating the mis-estimation of posterior variances without fitting a more expressive approximating class and enduring the corresponding increase in computational complexity [Giordano et al., 2018]. Since posterior hyperparameter sensitivity takes the form of posterior covariances, posterior covariances can be estimated using the corresponding sensitivity of the VB approximation. Specifically, for some $\phi_2(\theta)$, consider the exponentially tilted posterior, $\mathcal{P}(\theta|y,t) \propto \mathcal{P}(\theta|y)\exp(t\phi_2(\theta))$. When we can exchange integration and differentiation, we find that

$$\mathcal{P}(\theta|y,t) \propto \mathcal{P}(\theta|y)\exp(t\phi_2(\theta)) \quad \Rightarrow \quad \left.\frac{d \mathop{\mathbb{E}}_{\mathcal{P}(\theta|y,t)}[\phi_1(\theta)]}{dt}\right|_{t=0} = \mathop{\text{Cov}}_{\mathcal{P}(\theta|y)}(\phi_1(\theta), \phi_2(\theta)). \tag{10}$$

A detailed proof of Equation (10) is given in Theorem 1 of Giordano et al. [2018]; see also the classical score estimator of the derivative of an expectation [Mohamed et al., 2020]. Together,

Equations (9) and (10) motivate the LR approximation

$$\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}\left(\phi_1(\theta), \phi_2(\theta)\right) := \left.\frac{d \underset{\mathcal{Q}(\theta|\overset{*}{\eta}(t))}{\mathbb{E}}\left[\phi_1(\theta)\right]}{dt}\right|_{t=0} = \left.\frac{\partial \underset{\mathcal{Q}(\theta|\eta)}{\mathbb{E}}\left[\phi_1(\theta)\right]}{\partial \eta^\intercal}\right|_{\eta=\overset{*}{\eta}} \left.\frac{d\overset{*}{\eta}(t)}{dt}\right|_{t=0} \tag{11}$$

where $\overset{*}{\eta}(t)$ minimizes the KL divergence to the tilted posterior $\mathcal{P}(\theta|y,t)$. By applying the implicit function theorem to the first-order condition $\nabla_\eta \mathscr{L}_{\mathrm{VI}}\left(\overset{*}{\eta}\right) = 0$, together with the chain rule, Giordano et al. [2018] show that

$$\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}\left(\phi_1(\theta), \phi_2(\theta)\right) = \left.\frac{\partial \underset{\mathcal{Q}(\theta|\eta)}{\mathbb{E}}\left[\phi_1(\theta)\right]}{\partial \eta^\intercal}\right|_{\eta=\overset{*}{\eta}} \left(\nabla^2_\eta \mathscr{L}_{\mathrm{VI}}\left(\overset{*}{\eta}\right)\right)^{-1} \left.\frac{\partial \underset{\mathcal{Q}(\theta|\eta)}{\mathbb{E}}\left[\phi_2(\theta)\right]}{\partial \eta}\right|_{\eta=\overset{*}{\eta}} . \tag{12}$$

As discussed in Giordano et al. [2018] — and demonstrated in our experiments to follow — it can often be the case that $\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}\left(\phi_1(\theta), \phi_2(\theta)\right) \approx \underset{\mathcal{P}(\theta|y)}{\mathrm{Cov}}\left(\phi_1(\theta), \phi_2(\theta)\right)$, even when $\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{Cov}}\left(\phi_1(\theta), \phi_2(\theta)\right)$ is quite a poor approximation to $\underset{\mathcal{P}(\theta|y)}{\mathrm{Cov}}\left(\phi_1(\theta), \phi_2(\theta)\right)$. For example, in the case of multivariate normal posteriors, the LR covariances are exact, as we discuss in Section 4.1 below. See Giordano et al. [2018] for more extended discussion of the intuition behind Equation (11).

Unfortunately, the derivative $d\overset{*}{\eta}(t)/dt$ required by Equation (11) cannot be directly computed for ADVI. First, observe that the Hessian matrix $\nabla^2_\eta \mathscr{L}_{\mathrm{VI}}\left(\overset{*}{\eta}\right)$ in Equation (12) cannot be computed for ADVI since neither $\overset{*}{\eta}$ nor $\mathscr{L}_{\mathrm{VI}}(\cdot)$ is computable. One might instead approximate $\nabla^2_\eta \mathscr{L}_{\mathrm{VI}}(\eta)$ with $\nabla^2_\eta \widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z})$ by using additional Monte Carlo samples, and then evaluate at the ADVI optimum. However, due to noise in the SG algorithm, the ADVI optimum typically does not actually minimize $\mathscr{L}_{\mathrm{VI}}(\eta)$ nor $\widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z})$, so one is not justified in applying the implicit function theorem at the ADVI optimum.

In contrast, DADVI does not suffer from these difficulties because its objective function is available, and DADVI typically finds a parameter that minimizes that objective to a high degree of numerical accuracy; one can ensure directly that $\hat{\eta}$ is, to high precision, a local minimum of $\widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z})$. Therefore, we are justified in applying the implicit function theorem to the first-order condition $\nabla_\eta \widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z}) = 0$. If we follow the derivation of Equation (12) but with $\widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z})$ in place of $\mathscr{L}_{\mathrm{VI}}(\eta)$, we find the following tractable LR covariance estimate:

$$\widehat{\underset{\mathcal{Q}(\theta|\hat{\eta})}{\mathrm{LRCov}}}\left(\phi_1(\theta), \phi_2(\theta)\right) = \left.\frac{\partial \underset{\mathcal{Q}(\theta|\eta)}{\mathbb{E}}\left[\phi_1(\theta)\right]}{\partial \eta^\intercal}\right|_{\eta=\hat{\eta}} \hat{\mathcal{H}}^{-1} \left.\frac{\partial \underset{\mathcal{Q}(\theta|\eta)}{\mathbb{E}}\left[\phi_2(\theta)\right]}{\partial \eta}\right|_{\eta=\hat{\eta}} \tag{13}$$

$$\text{where} \quad \hat{\mathcal{H}} := \nabla^2_\eta \widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z}) .$$

We note that the same reasoning that leads to a tractable version of LR covariances applies to other sensitivity measures, such as prior sensitivity measures [Giordano et al., 2023] or the infinitesimal jackknife [Giordano et al., 2019]. Though we do not explore these uses of sensitivity analysis in the present work, one expects DADVI but not ADVI to support such analyses.

3.2 Monte Carlo error estimation

In this section we show how to estimate the Monte Carlo error of the output of DADVI. Since this estimate is based on use of the implicit function theorem, as we saw for LR covariances, it is again not as readily available to ADVI.

Let $f(\eta)$ denote some quantity of interest, such as a posterior expectation of the form $f(\eta) = \mathop{\mathbb{E}}_{\mathcal{Q}(\theta|\eta)}[\phi_1(\theta)]$ as in the previous Section 3.1. We are now interested in the sampling variance of $f(\hat{\eta}) - f(\hat{\eta}^*)$ due to the Monte Carlo randomness in $\mathscr{Z}$. We can apply standard asymptotic theory for the variance of M-estimators to find that this sampling variance is, in the notation of Section 3.1, consistently estimated by

$$\mathop{\mathrm{Var}}_{\mathcal{N}_{\mathrm{std}}(Z)}\left(f(\hat{\eta}) - f(\hat{\eta}^*)\right) \approx \frac{1}{\sqrt{N}} \nabla_\eta f(\hat{\eta})^\intercal \hat{\mathcal{H}}^{-1} \hat{\Sigma}_s \hat{\mathcal{H}}^{-1} \nabla_\eta f(\hat{\eta}) \,, \tag{14}$$

$$\text{where} \quad \hat{\Sigma}_s := \frac{1}{N} \sum_{n=1}^{N} \nabla_\eta \widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|Z_n) \, \nabla_\eta \widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|Z_n)^\intercal .$$

Equation (14) is analogous to the "sandwich covariance" estimate for misspecified maximum likelihood models [Stefanski and Boos, 2002]. Indeed, the question of how variable the DADVI estimate $\hat{\eta}$ is under sampling of $\mathscr{Z}$ is exactly the same as asking how variable a misspecified maximum likelihood estimator (or any M-estimator) is under sampling of the data, and the same conceptual tools can be applied. To complete the analogy, our $\hat{\Sigma}_s$ plays the role of the empirical score covariance, and $\hat{\mathcal{H}}$ plays the role of the empirical Fisher information.

Analogously to our discussion of LR covariances in Section 3.1, we briefly note that the classical derivation of Equation (14) is based on a Taylor series expansion of the first-order condition $\nabla_\eta \widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z}) = 0$, and so is not applicable to estimators like ADVI that do not satisfy any computable first-order conditions.

3.3 Computational considerations

We next describe best practices in computing LR covariances and the Monte Carlo sampling variability of the DADVI estimate. First, we delineate how to use these quantities to check that the number of samples N is adequate. Second, we discuss how to handle the primary computational difficulty of computing both quantities, namely the inverse of the Hessian matrix of the DADVI objective at the optimum.

In the postprocessing step of Algorithm 2, we recommend computing both Equation (13) and Equation (14) for each quantity of interest. One might consider a Bayesian analysis non-robust to sampling uncertainty if decisions based on the Bayesian analysis might change due to the sampling uncertainty. For instance, in a typical Bayesian analysis, one might make decisions based on how far a posterior mean is from a decision boundary in units of posterior standard deviation. Therefore, we might expect that sampling variability could be decision-changing if the estimated sampling variability dominated the estimated posterior uncertainty. In turn, then, we recommend using a comparison of the estimated quantities from Sections 3.1 and 3.2 to check the adequacy of the sample size N. If the estimated sampling variability dominates or might generally be sufficiently large as to be decision-changing, we recommend increasing N. In the present work we will not attempt to formalize nor to analyze such a procedure, although Burroni et al. [2023] and the general SAA literature [Royset and

Szechtman, 2013, Kim et al., 2015] attempt to estimate properties of the optimization and optimally allocate computing resources in a schedule of increasing sample sizes.

In Equations (13) and (14), the quantities $\hat{\Sigma}_s$ and $\nabla_\eta f(\hat{\eta})$ are typically straightforward to efficiently compute with automatic differentiation, but direct computation of $\hat{\mathcal{H}}^{-1}$ would incur a computational cost on the order of roughly D_η^3, which can be prohibitive in high-dimensional problems. However, for a given quantity of interest $f(\eta)$, it suffices for both Equations (13) and (14) to compute the D_η-vector $\hat{\mathcal{H}}^{-1}\nabla_\eta f(\hat{\eta})$. For models with very large D_η, we recommend evaluating $\hat{\mathcal{H}}^{-1}\nabla_\eta f(\hat{\eta})$ using the conjugate gradient method, which requires only Hessian-vector products of the form $\hat{\mathcal{H}}v$ [Nocedal and Wright, 1999, Chapter 5]. These products can be evaluated quickly using standard automatic differentiation software. As long as the number of quantities of interest is not large, both LR and sampling uncertainties can thus be computed at considerably less computational cost than a full matrix inversion.

4 Considerations in high dimensions

As discussed in Section 1, classical analysis in the optimization literature argues that, in the worst case, SAA is expected to require more computation than SG for a given approximation accuracy. The reason is that the total number of samples required for a given accuracy scales linearly with dimension, for both SG and SAA [Nemirovski et al., 2009, Shapiro et al., 2021, Chapter 5]. Since SAA requires more computation per sample than SG, one would correspondingly expect SAA to require more computation than SG for the same accuracy in high dimensional problems.

In this section we discuss why the aforementioned classical analysis of dimension dependence does not necessarily apply to the particular structure of the mean-field ADVI problem and some of its typical applications. We argue that, for problems that are approximately normal, or problems that are high dimensional due only to a having a large number of low-dimensional "local" parameters, DADVI can be effective with a relatively small number of samples which, in particular, need not grow linearly as the dimension of the problem grows. In contrast, we show that SAA may be inappropriate for more expressive BBVI approximations, such as full-rank ADVI. A key assumption of our analysis is that the user is interested in a relatively small number of scalar-valued quantities of interest, even though these quantities of interest may depend in some sense on the whole variational distribution.

4.1 High dimensional normals

We will show that in the normal model, the number of samples required to estimate any particular posterior mean do not depend on the dimension. Further, the LR covariances from DADVI are exact, irrespective of the problem dimension, and are in fact independent of the particular $\mathscr{Z}$ used. Conversely, the worst-case error in the DADVI posterior mean estimates across all dimensions will grow as dimension grows.

Take the quadratic model

$$\log \mathcal{P}(\theta, y) = -\frac{1}{2}\theta^\intercal A\theta + B^\intercal\theta = -\frac{1}{2}\mathrm{Tr}\left(A\theta\theta^\intercal\right) + B^\intercal\theta \tag{15}$$

for a known matrix $A \in \mathbb{R}^{D_\theta \times D_\theta}$ and vector $B \in \mathbb{R}^{D_\theta}$, possibly depending on the data y. Such a model arises, for example, when approximating the posterior of a conjugate normal location model, in which case the posterior mean $A^{-1}B$ and covariance matrix A^{-1} would depend on the sufficient statistics of the data y. Additionally, as we show below, the exact variational objective is available in closed form for the quadratic model. Of course, there is no need for a variational approximation to a posterior which is available in closed form, nor any need for a stochastic approximation to a variational objective which is available in closed form. However, studying quadratic models can provide intuition for the dimension dependence of DADVI approximations when the problem is *approximately* quadratic.

We first derive the exact variational objective function and its optimum. Recall our notation of Section 2, in which the variational posterior mean is denoted μ and the log standard deviation is denoted ξ. For compactness, we additionally write the vector of variational standard deviation parameters as $\sigma = \exp(\xi)$, where $\exp(\cdot)$ is applied component-wise. Let σ^2 be the corresponding vector of variance parameters. Note that

$$\mathop{\mathbb{E}}_{\mathcal{Q}(\theta|\eta)}[\theta] = \mu \quad \text{and} \quad \mathop{\mathbb{E}}_{\mathcal{Q}(\theta|\eta)}[\theta\theta^\intercal] = \mu\mu^\intercal + \mathrm{Diag}\left(\sigma^2\right),$$

$$\text{so} \quad \mathscr{L}_{\mathrm{VI}}(\eta) = \frac{1}{2}\mu^\intercal A \mu - \frac{1}{2}\sigma^\intercal (A \odot I_{D_\theta})\sigma - B^\intercal \mu - \sum_{d=1}^{D_\theta} \log \sigma_d.$$

The exact optimal parameters are thus

$$\overset{*}{\mu} = A^{-1}B \quad \text{and} \quad \overset{*}{\sigma}_d = (A_{dd})^{-1/2}.$$

If the objective had arisen from a multivariate normal posterior, observe the variational approximation to the mean is exactly correct, but the covariances are, in general, mis-estimated, since $1/A_{dd} \neq (A^{-1})_{dd}$ unless the true posterior covariance is diagonal.

We next make an asymptotic argument that we can expect any particular DADVI output to be a good estimate of the optimum of the exact objective, even for a small N.

Proposition 1. *Consider any parameter dimension index $d \in \{1, \ldots, D_\theta\}$, selected independently of $\mathscr{Z}$. In the quadratic model, we have $\hat{\sigma}_d^{-2} - \overset{*}{\sigma}_d^{-2} = O_p(N^{-1/2})$ and $\hat{\mu}_d - \overset{*}{\mu}_d = O_p(N^{-1/2})$. The constants do not depend on D_θ.*

Proof. We can compare the optimal parameters with the DADVI estimates. Let $\bar{z} := \frac{1}{N}\sum_{n=1}^N Z_n$. Let $\overset{d}{=}$ denote equality in distribution and let $Q \sim \chi^2_{N-1}$ denote a chi-squared random variable with $N-1$ degrees of freedom. We show in Appendix B.1 that, irrespective of the dimension of the problem,

$$\hat{\mu} = \overset{*}{\mu} - \hat{\sigma} \odot \bar{z} \quad \text{and} \quad \hat{\sigma}_d \overset{d}{=} \left(\frac{Q}{N} A_{dd}\right)^{-1/2}.$$

So $\mathop{\mathbb{E}}_{\mathcal{N}_{\mathrm{std}}(Z)}\left[\hat{\sigma}_d^{-2}\right] = \frac{N-1}{N} A_{dd} = \frac{N-1}{N}\overset{*}{\sigma}_d^{-2}$, and $\hat{\sigma}_d^{-2} - \overset{*}{\sigma}_d^{-2} = O_p(N^{-1/2})$. It follows that $\hat{\mu}_d - \overset{*}{\mu}_d = O_p(N^{-1/2})$ as well. □

The next remark suggests that, in cases with $N \ll D_\theta$, the worst-estimated linear combination of means is poorly estimated; note that in choosing the worst case we can overfit the draws $\mathscr{Z}$. It follows that the behaviors for any particular element of $\hat{\mu}$ and $\hat{\sigma}$ above do not imply that DADVI performs uniformly well across all parameters.

Remark 1. *In the quadratic model, we have*

$$\mathop{\mathbb{E}}_{\mathcal{N}_{\text{std}}(Z)}\left[\sup_{\nu:\|\nu\|_2=1} \nu^\intercal \frac{\hat{\mu} - \overset{*}{\mu}}{\hat{\sigma}}\right] = \mathop{\mathbb{E}}_{\mathcal{N}_{\text{std}}(Z)}\left[\sup_{\nu:\|\nu\|_2=1} \nu^\intercal \bar{z}\right] = \mathop{\mathbb{E}}_{\mathcal{N}_{\text{std}}(Z)}\left[\sqrt{\bar{z}^\intercal \bar{z}}\right] \approx \sqrt{\frac{D_\theta}{N}}.$$

In the first term of the preceding display, the division in the term $(\hat{\mu} - \overset{*}{\mu})/\hat{\sigma}$ is elementwise. The final relation follows since $\sqrt{N}\bar{z}$ is a D_θ-dimensional standard normal, so $N\bar{z}^\intercal\bar{z}$ is a $\chi^2_{D_\theta}$ random variable.

Finally, we show that the LR covariances reported from DADVI are exact, regardless of how small N is or, indeed, the particular values of $\mathscr{Z}$. Recall that, by contrast, the exact mean-field variance estimates are notoriously unreliable as estimates of the posterior variance.

Proposition 2. *In the quadratic model, we have*

$$\underset{\mathcal{Q}(\theta|\hat{\eta})}{\widehat{\text{LRCov}}}\,(\theta) = \left.\frac{d\hat{\mu}}{dt^\intercal}\right|_{\hat{\eta}} = A^{-1},$$

with no $\mathscr{Z}$ dependence.

See Appendix B.2 for a proof. Since A^{-1} is in fact the exact posterior variance, the linear response covariance is exact in this case irrespective of how small N is, in contrast to $\overset{*}{\sigma}$, which can be a poor estimate of the marginal variances unless A is diagonal.

4.2 High dimensional local variables

We next show that the number of samples required for DADVI estimation grows only logarithmically in dimension when the target joint distribution can be written as a large number of nearly independent problems that share a single, low-dimensional global parameter.

Formally, we say a problem has a "global–local" structure if we have the following decomposition:[3]

$$\theta = \begin{pmatrix} \gamma \\ \lambda^1 \\ \vdots \\ \lambda^P \end{pmatrix} \quad \text{and} \quad \log \mathcal{P}(\theta, y) = \sum_{p=1}^{P} \ell^p(\gamma, \lambda^p) + \ell^\gamma(\gamma), \tag{16}$$

where $\lambda^p \in \mathbb{R}^{D_\lambda}$ and $\gamma \in \mathbb{R}^{D_\gamma}$, and any data dependence is implicit in the functions ℓ^p and ℓ^γ. Here, the "global" parameters γ are shared among all "observations," and the "local" λ^p parameters do not occur with one another. We assume that the dimensions D_γ and D_λ are small, but that the total dimension $D_\theta = D_\gamma + PD_\lambda$ is large because P is large, i.e., because there are many local parameters.

Each vector involved in the variational approximation — the variational parameter η, the variational mean μ and standard deviation σ, the normal random variables Z, and the sets $\mathscr{Z}$ of normal random variables — can be partitioned into sub-vectors related to the global and

[3] Each local parameter is, itself, a vector, so we use superscripts to distinguish local parameters, retaining subscripts for particular elements of vectors.

local parameters. We will denote these subvectors with γ and p superscripts, respectively, so that, e.g., $\eta^\intercal = (\eta^\gamma, \eta^1, \ldots, \eta^p, \ldots, \eta^P)$, and so on. We will write Ω_η^γ for the domain of η^γ and Ω_η^p for the domain of η^p.

If there were no global parameters γ, then the high dimensionality would be no problem for DADVI. Without shared global parameters, the variational objective would consist of P completely independent D_λ-dimensional sub-problems. According to the classical optimization results referred to at the beginning of this section (e.g. Shapiro et al. [2021, Chapter 5]), under typical regularity conditions, each of these sub-problems' solutions could be accurately approximated with DADVI using no more than $N = O(D_\lambda)$ standard normal draws, each of length D_λ. The corresponding $\mathscr{Z}$ for the combined problem would stack the N vectors for each sub-problem, resulting in a $\mathscr{Z}$ consisting again of only N standard normal draws, each of length PD_λ. For this combined problem, any particular posterior mean of the combined problem (chosen independently of $\mathscr{Z}$) would then be well-estimated using only N draws, although we would expect more adversarial quantities such as $\max_p \sup_{v:\|v\|_2=1} v^\intercal (\hat{\eta}^p - \hat{\eta}^{*p})$ to be poorly estimated, as we saw in the quadratic problem (see Remark 1 in Section 4.1 above).

The goal of the present section is to state conditions under which the extra dependence induced by the shared finite-dimensional global parameter does not depart too strongly from the fully independent case described in the preceding paragraph. Our two key assumptions, stated respectively in Assumptions 1 and 2 below, are that each local problem obeys a sufficiently strong uniform law of large numbers, and that the local problems do not, in a certain sense, provide contradictory information about the global parameters.

To state our assumptions, let us first introduce some notation. Similar to around Equation (5), we write $\gamma(\eta^\gamma, Z^\gamma) = \mu^\gamma + \exp(\xi^\gamma) \odot Z^\gamma$, with analogous notation for $\lambda^p(\eta^p, Z^p)$.

Our first step is to write the variational objectives as the sum of "local objectives."

Definition 1. *Define the "local objective"*

$$f^p(\eta^\gamma, Z^\gamma, \eta^p, Z^p) := \ell^p\left(\gamma(\eta^\gamma, Z^\gamma), \lambda^p(\eta^p, Z^p)\right) + \sum_{d=1}^{D_\lambda} \xi_d^p + \frac{1}{P}\left(\ell^\gamma(\gamma(\eta^\gamma, Z^\gamma)) + \sum_{d=1}^{D_\gamma} \xi_d^\gamma\right).$$

We then define its expected value $\bar{f}^p$ and corresponding sample approximation $\hat{f}^p$:

$$\bar{f}^p(\eta^\gamma, \eta^p) := \mathop{\mathbb{E}}_{\mathcal{N}_{\text{std}}(Z)}\left[f^p(\eta^\gamma, Z^\gamma, \eta^p, Z^p)\right], \quad \hat{f}^p(\eta^\gamma, \mathscr{Z}^\gamma, \eta^p, \mathscr{Z}^p) := \frac{1}{N}\sum_{n=1}^{N} f^p(\eta^\gamma, Z_n^\gamma, \eta^p, Z_n^p).$$

With these definitions in hand, we observe that the mean-field objective for this model and its sample approximation can be written as functions of the local quantities.

$$\mathscr{L}_{\text{VI}}(\eta) = -\sum_{p=1}^{P} \bar{f}^p(\eta^\gamma, \eta^p), \qquad \widehat{\mathscr{L}}_{\text{VI}}(\eta|\mathscr{Z}) = -\sum_{p=1}^{P} \hat{f}^p(\eta^\gamma, \mathscr{Z}^\gamma, \eta^p, \mathscr{Z}^p).$$

Our key assumption is that a sub-Gaussian uniform law of numbers (ULLN) applies to each local objective.

Assumption 1 (A uniform law of large numbers applies to the local problems)**.** *Assume that, for any $\delta > 0$, there exist positive constants C_1, C_2, and N_0 depending on D_λ and D_γ but not on P such that for $N \geq N_0$,*

$$\mathcal{P}\left(\sup_{(\eta^\gamma, \eta^p) \in \Omega_\eta^\gamma \times \Omega_\eta^p} \left| \hat{f}^p(\eta^\gamma, \mathscr{Z}^\gamma, \eta^p, \mathscr{Z}^p) - \bar{f}^p(\eta^\gamma, \eta^p) \right| > \delta \right) \leq \varepsilon := C_1 \exp\left(-C_2 N_0\right).$$

Example 1. *Recall the definition of the "local objective" given in Definition 1. Assume that $\Omega_\eta^\gamma \times \Omega_\eta^p$ is compact and $\bar{f}^p$ is Lipschitz. Assume that, for all parameters in $\Omega_\eta^\gamma \times \Omega_\eta^p$, the moment generating function of $\hat{f}^p(\eta^\gamma, Z^\gamma, \eta^p, Z^p)$ is finite in a neighborhood of 0, and that $\underset{Z}{\mathrm{Var}}\left(\hat{f}^p(\eta^\gamma, Z^\gamma, \eta^p, Z^p)\right)$ is finite. Then Shapiro [2003, Theorem 12 and Equation 3.17] implies that Assumption 1 holds.*[4]

Though restrictive, the conditions of Example 1 are those that give rise to the commonly cited linear dimensional dependence for the SAA [e.g. Nemirovski et al., 2009, Kim et al., 2015, Homem-de Mello and Bayraksan, 2014]. Similar conditions to Example 1 can be also found in the statistics literature. For example, Wainwright [2019, Theorem 4.10] provides a bound of the form in Assumption 1 for bounded f^p with Rademacher complexity that decreases in N. Note that ADVI objectives, like many maximum likelihood problems, are typically over unbounded domains, with non-Lipschitz objective functions. In such cases, one can still use Assumption 1 by showing first that an estimator converges suitably quickly to a compact set with high probability, and then use Assumption 1 on that compact set; see, e.g., the discussion in Section 3.2.1 of Van der Vaart and Wellner [2013]. Our present purpose is not to survey the extensive literature on circumstances under which Assumption 1 holds, only to demonstrate simple, practically relevant conditions under which the SAA does not suffer from the worst-case dimensional dependence suggested by the SAA literature.

Next, we assume that the optima are well-defined for the local problems.

Assumption 2 (A strict minimum exists)**.** *Assume that there exists a strict optimum at $\overset{*}{\eta}$ in the sense that there exists a positive constant C_3, not depending on P, that satisfies*

$$\mathscr{L}_{\mathrm{VI}}(\eta) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta}) \geq P C_3 \left\|\eta^\gamma - \overset{*}{\eta}^\gamma\right\|_2^2 \quad \text{and} \quad \mathscr{L}_{\mathrm{VI}}(\eta) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta}) \geq C_3 \sum_{p=1}^{P} \left\|\eta^p - \overset{*}{\eta}^p\right\|_2^2.$$

As illustrated by Example 2 below, a key aspect of Assumption 2 is that each local objective function is informative about the global parameter, so that as the dimension P grows, the global objective function grows "steeper" as a function of η^γ.

Example 2. *Recall Definition 1. Suppose that, for each p, the expected local objective $\bar{f}^p(\eta^\gamma, \eta^p)$ is twice-differentiable and uniformly convex, in the sense that there exists a lower*

[4] The connection between our notation and Shapiro's is as follows. Shapiro's α is our $1 - \varepsilon$. Shapiro's ε is our δ. Shapiro's $\delta = 0$ in our case because we assume that $\hat{\eta}$ is an exact optimum. Shapiro's diameter D is bounded because $\Omega_\eta^\gamma \times \Omega_\eta^p$ is compact. Shapiro's L is our Lipschitz constant. Shapiro's n is our $D_\gamma + D_\lambda$. And Shapiro's $\sigma_{\max}^2$ is bounded by our assumption on the variance of $\hat{f}^p$. A similar but more detailed result can also be found in Shapiro et al. [2021, Section 5.3.2].

bound $C_3 > 0$ on the eigenvalues of the second derivative matrices of $\bar{f}^p(\eta^\gamma, \eta^p)$, uniformly in both p and η. Then, by a Taylor series expansion,

$$\mathscr{L}_{\mathrm{VI}}(\eta) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta}) \geq C_3 \left(P \left\| \eta^\gamma - \overset{*}{\eta}^\gamma \right\|_2^2 + \sum_{p=1}^{P} \left\| \eta^p - \overset{*}{\eta}^p \right\|_2^2 \right),$$

from which Assumption 2 follows. (See Appendix C for more details.)

Theorem 3. *Under Assumptions 1 and 2, for any $\varepsilon > 0$ and $\delta > 0$, there exists an N_0, depending only logarithmically on P, such that $N \geq N_0$ implies that*

$$\mathcal{P}\left(\left\| \hat{\eta}^\gamma - \overset{*}{\eta}^\gamma \right\|_2^2 \leq \delta \text{ and, for all } p, \ \left\| \hat{\eta}^p - \overset{*}{\eta}^p \right\|_2^2 \leq \delta \right) \geq 1 - \varepsilon.$$

sketch. By Assumption 2, closeness of $\bar{f}^p(\eta^\gamma, \eta^p)$ and $\hat{f}^p(\eta^\gamma, \mathscr{Z}^\gamma, \eta^p, \mathscr{Z}^p)$ implies closeness of $\hat{\eta}^p$ and $\overset{*}{\eta}^p$, and closeness of $\frac{1}{P}\widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z})$ and $\frac{1}{P}\mathscr{L}_{\mathrm{VI}}(\eta)$ implies closeness of $\hat{\eta}^\gamma$ and $\overset{*}{\eta}^\gamma$. Thus, for $\hat{\eta}$ to be close to $\overset{*}{\eta}$, it suffices for $\left| \bar{f}^p(\eta^\gamma, \eta^p) - \hat{f}^p(\eta^\gamma, \mathscr{Z}^\gamma, \eta^p, \mathscr{Z}^p) \right| < \delta'$ simultaneously for all p, and for some δ' that is a function of δ and the constants in Assumption 2. To apply a union bound to Assumption 1 requires decreasing ε by a factor of P, which requires increasing N by a factor of no more than $\log P$.

See Appendix C for a detailed proof. □

The key difference between classical results such as Shapiro et al. [2021, Chapter 5] and our Theorem 3 is that, in the classical results, $N = O(P)$, whereas for Theorem 3, $N = O(\log P)$. Intuitively, N need grow only logarithmically in P because the global parameters are sharply identified, which approximately decouples the remaining local problems.

4.3 DADVI fails for full-rank ADVI

The preceding sections demonstrated that, in certain cases, DADVI can work well to estimate the optimum of the mean-field ADVI problem even in high dimensions. By contrast, we now show that DADVI will behave pathologically for the full-rank ADVI problem in high dimensions unless a prohibitively large number of draws are used. The intuition we develop for full-rank ADVI also extends to other highly expressive variational approximations such as normalizing flows.

In forming the full-rank optimization problem, ADVI parameterizes $\mathcal{Q}(\theta|\eta)$ using a mean μ and a $D_\theta \times D_\theta$ matrix R in place of σ. Formally, the full-rank approximation taking $\theta(\eta, Z) = \mu + RZ$ in place of the mean-field reparameterization is given in Equation (6). Letting $|\cdot|$ denote the matrix determinant, the KL divergence becomes

$$\widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z}) := -\frac{1}{2}\log|RR^\intercal| - \frac{1}{N}\sum_{n=1}^{N} \log \mathcal{P}(\mu + RZ_n, y). \tag{17}$$

The preceding display can be compared with the corresponding mean-field objective in Equation (7). For the present section, we will write $\widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z}) = \widehat{\mathscr{L}_{\mathrm{VI}}}((\mu, R)|\mathscr{Z})$. Under this parameterization, $\underset{\mathcal{Q}(\theta|\eta)}{\mathrm{Cov}}(\theta) = \underset{\mathcal{N}_{\mathrm{std}}(Z)}{\mathrm{Cov}}(\theta(\eta, Z)) = RR^\intercal$, so the matrix R can be taken to be

any square root of the covariance matrix of $\mathcal{Q}(\theta|\eta)$. In practice, R is typically taken to be lower-triangular (i.e., a Cholesky decomposition), though the particular form of the square root used will not matter for the present discussion.

Suppose we are attempting to optimize the full-rank ADVI problem with DADVI when $D_\theta > N$, so that θ has more dimensions than there are draws Z_n. Our next result shows that, in such a case, DADVI will behave pathologically.

Theorem 4. *Consider a full-rank ADVI optimization problem with $D_\theta > N$. Then, for any μ, we have $\inf_R \widehat{\mathscr{L}_{\text{VI}}}((\mu, R)|\mathscr{Z}) = -\infty$, so the DADVI estimate is undefined.*

Proof. In the full-rank case, the objective function $\widehat{\mathscr{L}_{\text{VI}}}(\eta|\mathscr{Z})$ in Equation (17) depends on R only through the products RZ_n and the entropy term, which is $\frac{1}{2}\log|RR^\intercal| = \log|R|$. Since $N < D_\theta$, we can write $R = R^{\mathscr{Z}} + R^\perp$, where $R^{\mathscr{Z}}$ is a rank-N matrix operating on the subspace spanned by $\mathscr{Z}$ and $R^\perp$ is a rank-$(D_\theta - N)$ matrix satisfying $R^\perp z_n = 0$ for all $n = 1, \ldots, N$. Then we can rewrite the DADVI objective as

$$\widehat{\mathscr{L}_{\text{VI}}}(\eta|\mathscr{Z}) = -\log|R^{\mathscr{Z}} + R^\perp| - \frac{1}{N}\sum_{n=1}^{N} \log \mathcal{P}(\mu + R^{\mathscr{Z}} Z_n, y). \tag{18}$$

Since $\sup_{R^\perp} \log|R^{\mathscr{Z}} + R^\perp| = \infty$, the result follows.[5] □

What will happen, in practice, if one tries to use DADVI in the full-rank case? Denote the maximum a posteriori (MAP) estimate as $\hat{\theta} := \text{argmax}_\theta \log \mathcal{P}(\theta, y)$, and note that the first term on the right hand side of Equation (18) is most negative when $\mu = \hat{\theta}$ and $R^{\mathscr{Z}}$ is the zero matrix. A zero $R^{\mathscr{Z}}$ is impermissible because, when $R^{\mathscr{Z}}$ is actually the zero matrix, then $R^{\mathscr{Z}} + R^\perp$ is singular, and $\log|R^{\mathscr{Z}} + R^\perp| = -\infty$.[6] However, for any $\varepsilon > 0$ and $M > 0$, we can take $R^{\mathscr{Z}} Z_n = \varepsilon Z_n$ and $R^\perp v = Mv$ for any $v \perp \mathscr{Z}$, so that R is full-rank. When $\mu = \hat{\theta}$, one can always decrease both terms on the right hand side of Equation (18) via the following two-step procedure. First, decrease ε by any amount and thereby decrease the first term. Second, given that ε, increase M by a sufficient amount to decrease the second term as well.

The degeneracy described in Theorem 4 can be avoided if one uses at least as many draws as there are model parameters, i.e., if $N \geq D_\theta$. However, based on the classical optimization results discussed above (e.g. Shapiro et al. [2021, Chapter 5]), one might expect full-rank ADVI to require N to be on the order of D_θ^2, since the full-rank variational parameters have dimension of order D_θ^2 due to the inclusion of a full-rank covariance matrix. We proved above that the classical dimension dependence of N on the dimension of the variational parameters is unnecessarily pessimistic for certain mean-field ADVI objectives. It is an interesting question for future work to ask whether the classical dimension dependence is also pessimistic for the full-rank approximation: that is, whether DADVI for full-rank ADVI actually requires N to be on the order D_θ rather than D_θ^2, or somewhere in between.

Finally, we note that the failure of DADVI in the full-rank case appears to be indicative of a general phenomenon. Any smooth function mapping the columns of $\mathscr{Z}$ into Ω_θ must

[5] Recall that the log determinant is the sum of the logs of the eigenvalues of R, which can be made arbitrarily large as $R^\perp$ varies freely.

[6] Indeed, if $R^{\mathscr{Z}} = 0$, then $\mathcal{Q}(\theta|\eta)$ would have zero variance in any direction spanned by $\mathscr{Z}$, $\mathcal{P}(\theta, y)$ would not be absolutely continuous with respect to $\mathcal{Q}(\theta|\eta)$, and the reverse KL divergence would be undefined.

span an N-dimensional sub-manifold of Ω_θ. If a variational approximation is rich enough to increase the entropy to an arbitrary degree on the complement of this submanifold, then DADVI will lead to a degenerate solution. In this sense, it is in fact the inexpressivity of the mean-field variational assumption that allows DADVI to work in high dimensions.

5 Related work

As discussed above in Section 1, the idea of approximating an intractable optimization objective $F(\eta) := \mathop{\mathbb{E}}_{\mathcal{N}_{\text{std}}(Z)}[f(\eta, Z)]$ by $\hat{F}(\eta|\mathscr{Z}) := \frac{1}{N}\sum_{n=1}^{N} f(\eta, Z_n)$ is well-studied in the optimization literature as the "sample average approximation" (SAA) [Nemirovski et al., 2009, Royset and Szechtman, 2013, Kim et al., 2015, Shapiro et al., 2021, Chapter 5]. A key theoretical conclusion of the optimization literature is that, in general, SAA should perform worse than SG in high dimensions in terms of computational cost of providing an accurate optimum. Our theoretical results of Section 4 and experimental results of Section 6 suggest that these general-case analyses may be unduly pessimistic for many BBVI problems, though we believe more work remains to be done establishing guarantees for SAA applied to BBVI in high dimensions.

The present work and the concurrent work by Burroni et al. [2023] together form the first systematic studies of the accuracy of SAA for BBVI, though the idea of applying SAA to BBVI has occurred several times in the literature in the context of other methodological results [Giordano et al., 2018, Domke and Sheldon, 2018, 2019, Broderick et al., 2020, Wycoff et al., 2022, Giordano et al., 2023]. The methods and experiments of Burroni et al. [2023] provide a complement to our present work. Burroni et al. [2023] propose and study a method for iteratively increasing the number of draws used for the SAA approximation until a desired accuracy is reached (see also Royset and Szechtman [2013] for a similar approach in the optimization literature); in contrast, we keep the number of draws fixed in our theoretical analysis and our experiments. Additionally, the models considered by Burroni et al. [2023] are relatively low-dimensional, which allow the authors to use a very large number of draws (up to $N = 2^{18}$) without incurring a prohibitive computational cost. In contrast, almost all of our experiments in Section 6 use $N = 30$; only in our investigation of Monte Carlo error in Section 6.5 do we examine changing N, and there we consider only N up to 64. Studying relatively lower-dimensional models with a large number of draws allows Burroni et al. [2023] to apply SAA with the full-rank approximation (see our discussion of the SAA with the full-rank approximation in Section 4.3). In contrast, we emphasize the computation of LR covariances with the SAA approximation (as in Giordano et al. [2018]) and on the use of DADVI in higher dimensions more generally. One could imagine combining our approaches: for example, by computing the size of the sampling error relative to the LR covariance, and increasing the number of draws as necessary as recommended in Burroni et al. [2023], though we leave such a synthesis for future work.

6 Experiments

We consider a range of models and datasets. We find that, despite using out-of-the box optimization and convergence criteria, DADVI optimization (using the SAA approximation)

typically converges much faster than classical (stochastic) ADVI. DADVI performs comparably to ADVI in posterior mean estimation while allowing much better posterior covariance estimation via linear response. Upon examination of optimization trajectories, we find that ADVI tends to eventually find better ADVI objective values than DADVI but typically takes longer to do so. And we confirm that the sampling variability estimates available from DADVI are of high quality, even for just tens of draws.

Below, DADVI exhibits good performance on a number of high-dimensional models. These models do not obviously satisfy any of the theoretical conditions for good performance of the SAA established above (Section 4). So our experimental results point to a gap between theory and experiment that is an interesting subject for future work.

In our experiments, as in the rest of the paper, we follow the convention that "ADVI" refers to methods that use stochastic optimization, and "DADVI" refers to our proposal of using SAA with the ADVI objective function.

6.1 Models and data

We evaluate DADVI and ADVI on the following models and datasets.

- ARM: 53 models and datasets taken from a hierarchical-modeling textbook [Gelman and Hill, 2006]. The datasets are relatively small and the models consist of textbook linear and generalized linear models, with and without random effects.

- Microcredit: A hierarchical model from development economics [Meager, 2019] that performs shrinkage on seven randomized controlled trials. The model accounts for heavy tails, asymmetric effects, and zero-inflated observations.

- Occupancy: A multi-species occupancy model from ecology [Ingram et al., 2022, Kery and Royle, 2009]. In occupancy models, the question of interest is whether a particular species is present at (i.e. occupying) a particular site. The data consist of whether the species was observed at repeated visits to the site. At any given visit, the species may be present but not observed. Occupancy models estimate both (1) the suitability of a site as a function of environmental covariates such as temperature or rainfall and (2) the probability of observing the species given that it is present (the observation process). The resulting likelihood makes it a non-standard regression model and thus a good candidate for a black-box inference method. Here we use a multi-species occupancy model that places a hierarchical prior on the coefficients of the observation process. Our dataset comprises 1387 sites, 43 environmental covariates at each site, 32 different species, and 2000 visits; this dataset represents a subset of the eBird dataset used by Ingram et al. [2022].[7]

- Tennis: A Bradley-Terry model with random effects for ranking tennis players. In this model, each tennis player has a rating, assumed fixed throughout their career. The probability of a given player beating another is determined by the inverse logit of their rating difference. The ratings are modeled as random effects, and the data comprises all men's professional tennis matches on the ATP tour since 1969. Overall, this is a

[7] We used a subset so that our ground-truth MCMC method would complete in a reasonable amount of time.

large dataset of 164,936 matches played between 5,013 different players, each of whom has their own random effect, making this a high-dimensional mixed model.

- POTUS: A time series polling model for the US presidential election [Heidemanns et al., 2020]. This model is both complex and high-dimensional. It models logit polling probabilities with a reverse autoregressive time series and random effects for various polling conditions.

Throughout this section, by a "model" we will mean a model with its corresponding dataset.

Model Name	D_θ	NUTS runtime
ARM (53 models)	2 to 176 (median 5)	15 seconds to 16 minutes (median 39 seconds)
Microcredit	124	597 minutes
Occupancy	$1,884$	251 minutes
Tennis	$5,014$	57 minutes
POTUS	$15,098$	643 minutes

Table 1: Model summaries.

These models differ greatly in their complexity, as can be seen in Table 1. The 53 ARM models from Gelman and Hill [2006] are generally simple,[8] ranging from fixed effects models with a handful of parameters to generalized linear mixed models with a few hundred parameters. The other four models are more complex, with total parameter dimension D_θ ranging from 124 for the Microcredit model to $15,098$ for the POTUS model. We restricted attention to posteriors that could be tractably sampled from with the NUTS MCMC algorithm [Hoffman and Gelman, 2014] as implemented in PyMC [Salvatier et al., 2016] in order to have access to "ground truth" posterior means and variances. However, outside the relatively simple ARM models, NUTS samplers were time-consuming, which motivates the use of faster variational approximations.

We fit each model using the following methods, including three different versions of ADVI.

- NUTS: The "no-U-turn" MCMC sampler in PyMC [Salvatier et al., 2016].

- DADVI: Except where otherwise indicated, we report results with $N = 30$ draws for DADVI for each model. We optimized using an off-the-shelf second-order Newton trust region method (`trust-ncg` in `scipy.optimize.minimize`). Our implementation of DADVI is available as a Python package at `https://github.com/martiningram/dadvi`.

- LRVB: Using the optimum found by DADVI, we computed linear response covariance estimates. In the high-dimensional models Occupancy, Tennis, and POTUS, we selected a small number of quantities of interest and used the conjugate gradient (CG) algorithm to compute the LR covariances and frequentist standard errors. For Occupancy, the quantities of interest were predictions of organism presence at 20 sites; for Tennis the quantities of interest were win predictions of 20 randomly chosen matchups; and, for

[8] Indeed, many of the ARM models can be fit quickly enough with MCMC that BBVI is arguably not necessary. We include all the ARM models in our results to show that DADVI works well in both simple and complex cases.

POTUS, the quantity of interest was the national vote share received by the democratic candidate on election day. When using the CG algorithm, we preconditioned using the estimated variational covariance as described in Appendix E. When reporting metrics for the computational cost of computing LRVB, we always report the total cost of the posterior approximation — i.e., the cost of DADVI optimization plus the additional cost of computing the LR covariances.

- Mean field ADVI (ADVI): We used the PyMC implementation of ADVI, together with its default termination criterion. Every 100 iterations, this termination criterion compares the current parameter vector with the one 100 iterations ago. It then computes the relative difference for each parameter and flags convergence if it falls below 10^{-3}. We ran ADVI for up to 100,000 iterations if convergence was not flagged before then.

- RAABBVI (ADVI): RAABVI represents a state-of-the art stochastic mean field ADVI method employing principled step size selection and convergence assessment [Welandawe et al., 2022]. To run RAABBVI, we used the public package `viabel`,[9] provided by Welandawe et al. [2022]. By default, `viabel` supports the packages `autograd` and `Stan`. To be able to run RAABBVI with PyMC, we provide it with gradients of the objective function computed with PyMC's `JAX` backend, which we use also for DADVI.

- Full-rank ADVI (ADVI): When possible, we used the PyMC implementation of full-rank ADVI, together with the default termination criterion for ADVI described above. Full-rank was computationally prohibitive for all but the ARM and Microcredit models.

6.2 Computational cost

We first show that, despite using out-of-the box optimization and convergence criteria, DADVI optimization typically converges faster than the ADVI methods. DADVI also converges much more reliably; in many cases, the ADVI methods either converged early according to their own criteria or failed to converge and had to be terminated after a large, pre-determined number of draws.

We measured the computational cost of a method in two different ways: the wall time ("runtime"), and the number of model gradient or Hessian-vector product evaluations ("model evaluations"). Neither is a complete measure of a method's computational cost, and we hope to provide a more thorough picture by reporting both. For example, we were able to naively parallelize DADVI by evaluating the model on each draw of $\mathscr{Z}$ in parallel, whereas ADVI uses a single draw per gradient step and cannot be parallelized in this way. As a consequence, DADVI will have a favorable runtime relative to ADVI for the same number of model evaluations.

We included NUTS runtime results as a baseline. We do not include model evaluations for NUTS, since standard NUTS packages do not typically report the number of model evaluations used for leapfrog steps that are not saved as part of the MCMC output.

The results for ARM and non-ARM models are shown respectively in Figures 1 and 2. Both DADVI and LRVB are faster than all competing methods in terms of both runtime and model evaluations on most models, with the exception of a small number of ARM models and

[9] `https://github.com/jhuggins/viabel`

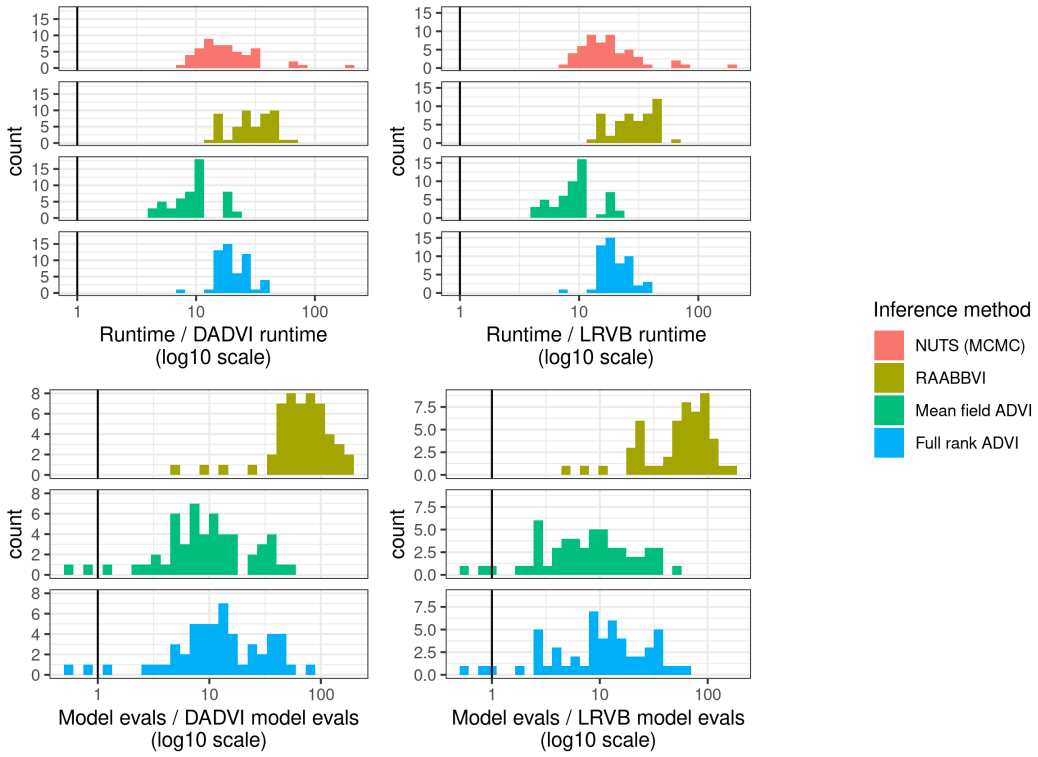

Figure 1: Runtimes and model evaluation counts for the ARM models. Results are reported divided by the corresponding value for DADVI or LRVB. Numbers greater than one (shown by the black line) indicate favorable performance by DADVI or LRVB. Recall that the reported LRVB numbers include the cost of the DADVI optimization as well as the LR covariances. Most of the ARM models are relatively low-dimensional, so the LR covariances added little to the computation.

the Occupancy model. These computational benefits are favorable for DADVI and LRVB given the results of Section 6.3 below showing that the posterior approximations provided by DADVI and LRVB are similar to or better than the posterior approximations from the other methods.

6.3 Posterior Accuracy

We next see that the quality of posterior mean estimates for DADVI and the ADVI methods are comparable. The LRVB posterior standard deviations are much more accurate than the ADVI methods, including full-rank ADVI.

Each method produced a posterior mean estimate for each model parameter, μ_{METHOD}, and a posterior standard deviation estimate, σ_{METHOD}. Above, we used μ to denote the posterior expectation of the full θ vector, but here we are using it more generically to denote a posterior expectation of some sub-vector of θ, or even the posterior mean of a transformed parameter as estimated using Monte Carlo draws from the variational approximation in the unconstrained space. We use the NUTS estimates, μ_{NUTS} and σ_{NUTS}, as the ground truth to

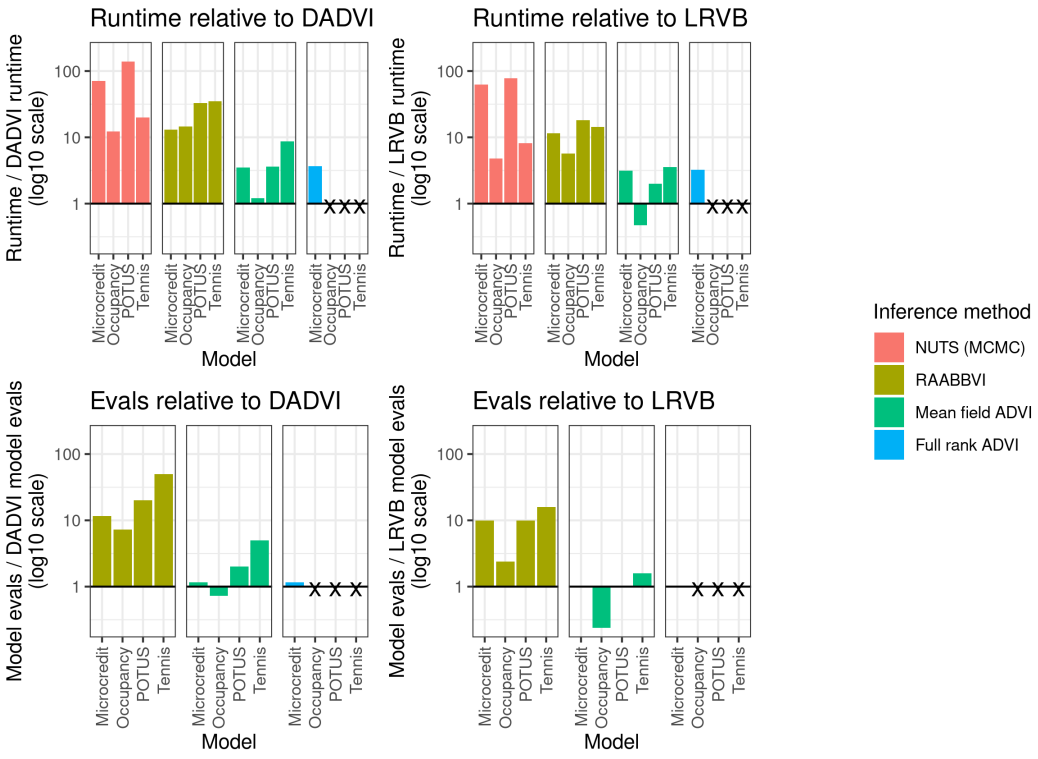

Figure 2: Runtimes and model evaluation counts for the non-ARM models. Results are reported divided by the corresponding value for DADVI or LRVB. Numbers greater than one (shown by the black line) indicate favorable performance by DADVI or LRVB. Recall that the reported LRVB numbers include the cost of the DADVI optimization as well as the LR covariances. Missing model and method combinations are marked with an X.

which we compare the various variational methods. In order to form a common scale for the accuracy of the posterior means and variances, we define the relative error in the posterior mean and standard deviation as follows:

$$\varepsilon^{\mu}_{\text{METHOD}} := \frac{\mu_{\text{METHOD}} - \mu_{\text{NUTS}}}{\sigma_{\text{NUTS}}} \quad \text{and} \quad \varepsilon^{\sigma}_{\text{METHOD}} := \frac{\sigma_{\text{METHOD}} - \sigma_{\text{NUTS}}}{\sigma_{\text{NUTS}}}.$$

For example, if, on a particular parameter of a particular model, we find that $\|\varepsilon^{\mu}_{\text{DADVI}}\| < \|\varepsilon^{\mu}_{\text{MF-ADVI}}\|$, we would say that DADVI has provided better mean estimates of that model parameter than mean-field ADVI. For posterior covariances we will always report $\varepsilon^{\sigma}_{\text{LRVB}}$ rather than $\varepsilon^{\sigma}_{\text{DADVI}}$, since we expect σ_{DADVI} to suffer from the same deficiencies as the ADVI methods due to their shared use of the mean-field approximation.

As discussed in Section 2, any parameters with restricted ranges will typically be transformed before running ADVI. In our plots, then, we include one point each for the original and transformed versions, respectively, of each distinctly named parameter in the PyMC model. For Occupancy, Tennis, and POTUS, we reported posterior mean accuracy measures for all parameters, but posterior uncertainty measures only for a small number of quantities of interest. When a named parameter is multi-dimensional, we report the norm of the error

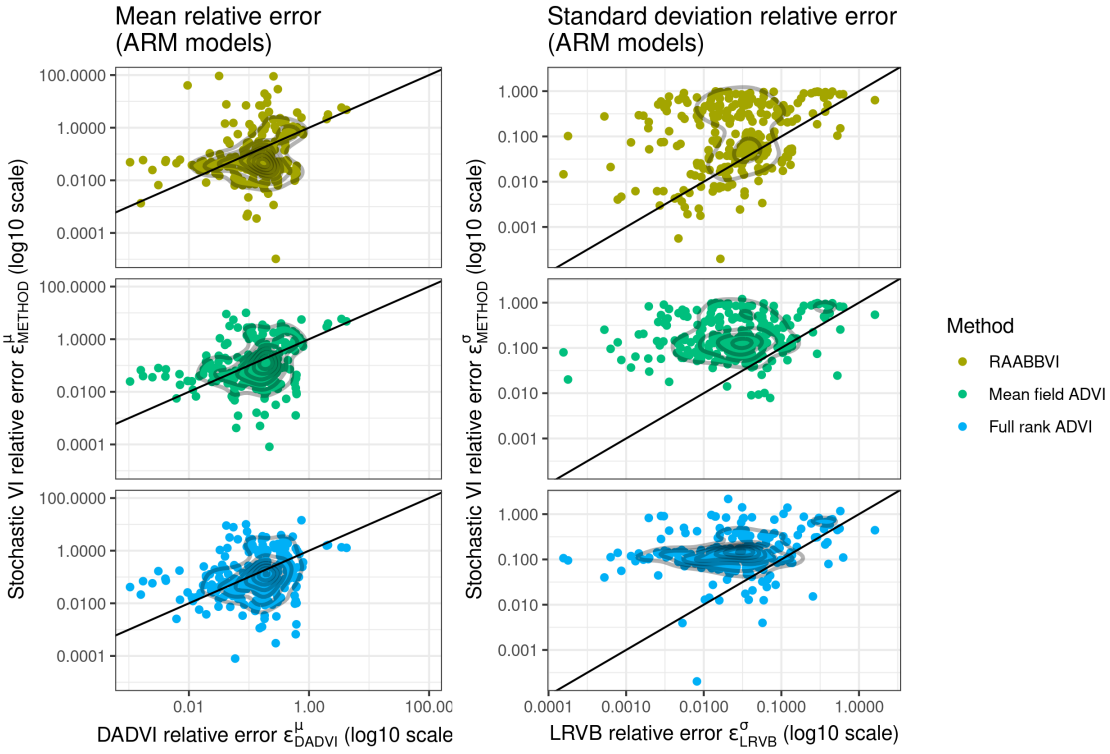

Figure 3: Posterior accuracy measures for the ARM models. Each point is a single named parameter in a single model. Points above the diagonal line indicate better DADVI or LRVB performance. Level curves of a 2D density estimator are shown to help visualize overplotting.

vector over all dimensions in order to avoid giving too much visual weight to a small number of high-dimensional parameters.

The posterior accuracy results for ARM and the larger models are shown respectively in Figures 3 and 4. Recall that, of the non-ARM models, only the Microcredit model was small enough for full-rank ADVI.

The estimates for the posterior means are comparable across methods, with RAABBVI performing the best on average. However, there are parameters for which RAABBVI's mean estimates are off by up to a hundred standard deviations while the DADVI estimates are fairly accurate. In contrast, when the DADVI mean estimates are severely incorrect, the RAABBVI ones are also severely incorrect. This pattern suggests that severe errors in the DADVI posterior means are primarily due to the mean-field approximation, whereas severe errors in ADVI methods can additionally occur due to problems in optimization.

The LRVB posterior standard deviation estimates are almost uniformly better than the ADVI and RAABBVI estimates based on the mean-field approximation. This performance is not surprising since the mean field approximation is known to produce poor posterior standard deviation estimates.[10] Interestingly, for the ARM models, even the full-rank ADVI posterior

[10] Note that the relative standard deviation errors for ADVI tend to cluster around 1 because MFVB posterior standard deviations tend to be under-estimated, and so a small posterior standard deviation estimate leads to a relative error of one.

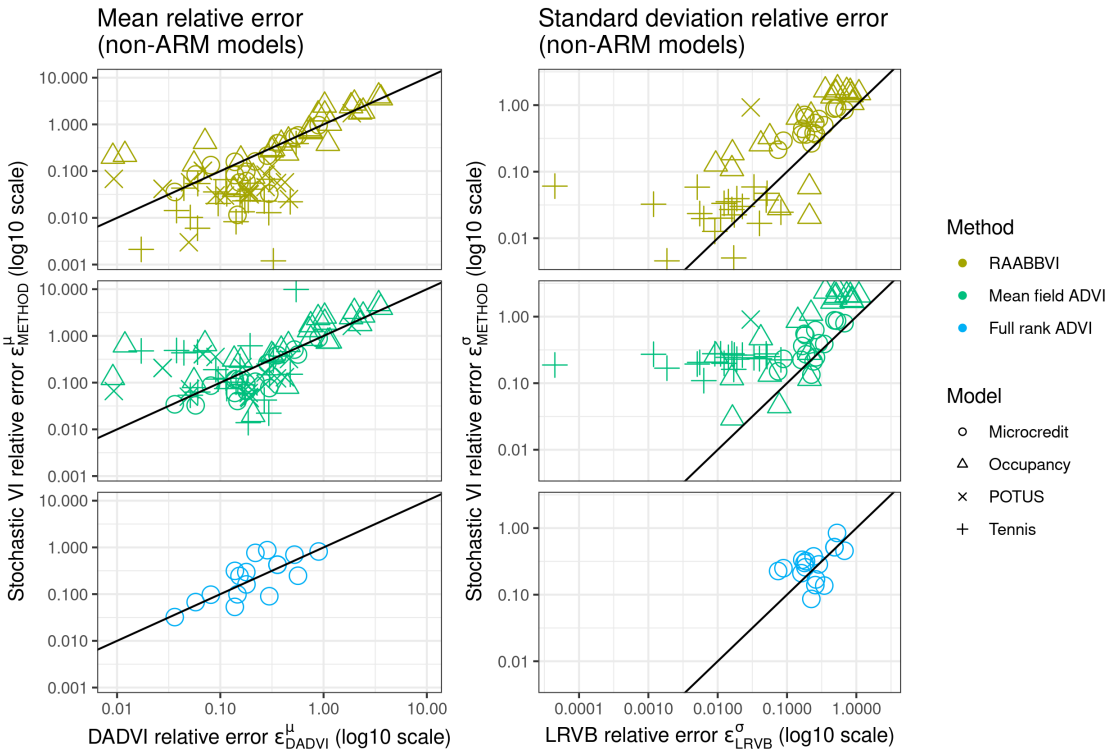

Figure 4: Posterior accuracy measures for the non-ARM models. Each point is a single named parameter in a single model. Points above the diagonal line indicate better DADVI or LRVB performance.

covariance estimates are worse than the LRVB covariance estimates, which is probably due to the difficulty of optimizing the full-rank ADVI objective.

6.4 Assessing convergence

By examining the optimization traces, we next see that the ADVI methods eventually find better optima (in terms of the variational objective) than DADVI, but they typically take longer than DADVI to terminate, in agreement with Section 6.2.

In order to understand the progress of ADVI and RAABBVI towards their optimum, we evaluated the variational objective on a set of 1000 independent draws[11] for each method along its optimization. This evaluation is computationally expensive, but gives a good estimate of the true objective $\mathscr{L}_{\text{VI}}(\cdot)$ along the optimization paths. Specifically, letting η^{i}_{METHOD} denote the variational parameters for method METHOD after i model evaluations, and letting $\tilde{\mathscr{Z}}$ denote the set of 1000 independent draws, we evaluated $\widehat{\mathscr{L}_{\text{VI}}}\left(\eta^{i}_{\text{METHOD}} | \tilde{\mathscr{Z}}\right)$ for each method and for steps i up to convergence.

In order to place the optimization traces on a common scale, for each method we center and scale the objective values by the DADVI optimum and sampling standard deviation. In

[11] We used the same set of independent draws for each method to ensure a like-to-like comparison.

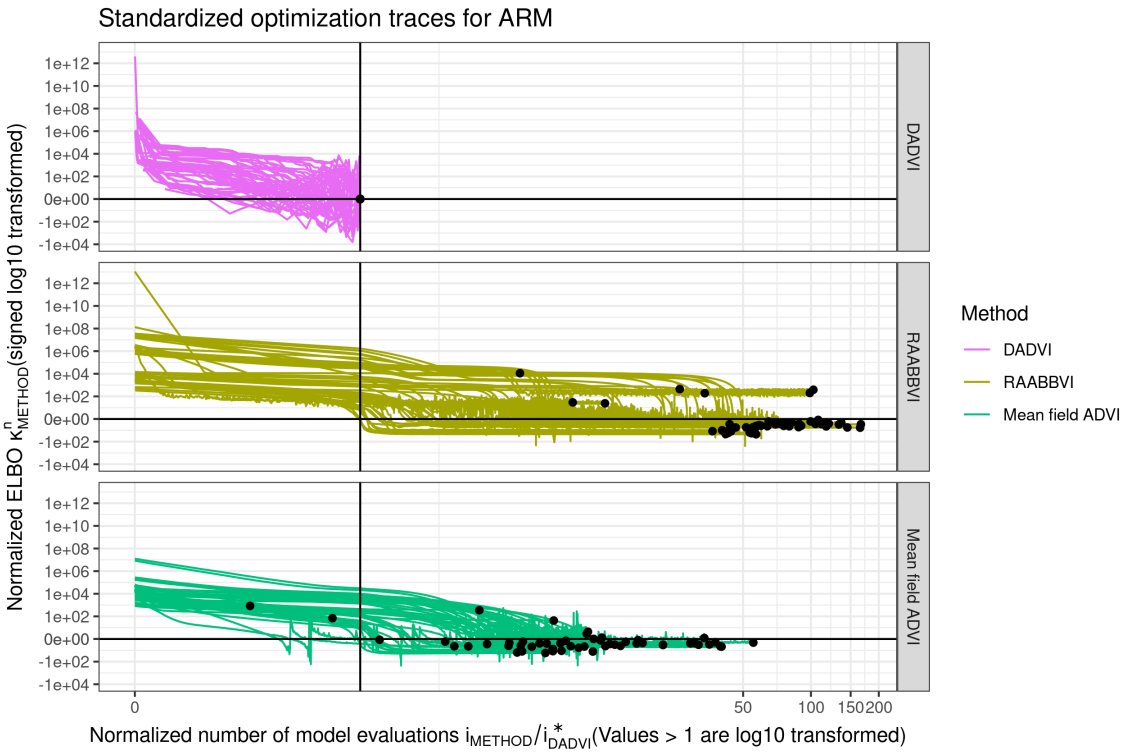

Figure 5: Optimization traces for the ARM models. Black dots show the termination point of each method. Dots above the horizontal black line mean that DADVI found a better ELBO. Dots to the right of the vertical black line mean that DADVI terminated sooner in terms of model evaluations.

particular, we report κ^i_{METHOD}, which is equal to

$$\kappa^i_{\text{METHOD}} := \frac{\widehat{\mathscr{L}}_{\text{VI}}\left(\eta^i_{\text{METHOD}}|\tilde{\mathscr{Z}}\right) - \widehat{\mathscr{L}}_{\text{VI}}\left(\hat{\eta}_{\text{DADVI}}|\tilde{\mathscr{Z}}\right)}{\sqrt{\widehat{\underset{\tilde{\mathscr{Z}}}{\text{Var}}}\left(\widehat{\mathscr{L}}_{\text{VI}}\left(\hat{\eta}_{\text{DADVI}}|Z\right)\right)}} \tag{19}$$

where $\widehat{\underset{\tilde{\mathscr{Z}}}{\text{Var}}}\left(\widehat{\mathscr{L}}_{\text{VI}}\left(\hat{\eta}_{\text{DADVI}}|Z\right)\right)$ denotes an approximation to $\underset{\mathcal{N}_{\text{std}}(Z)}{\text{Var}}\left(\widehat{\mathscr{L}}_{\text{VI}}\left(\hat{\eta}_{\text{DADVI}}|Z\right)\right)$ using the sample variance over $\tilde{\mathscr{Z}}$. Let i^*_{METHOD} denote the number of model evaluations taken by a method at convergence. Then, under Equation (19), $\kappa^{i^*_{\text{DADVI}}}_{\text{DADVI}} = 1$ by definition, $\kappa^{i^*_{\text{METHOD}}}_{\text{METHOD}} < 1$ indicates a better optimum at convergence for METHOD relative to DADVI, and $i^*_{\text{METHOD}} < i^*_{\text{DADVI}}$ indicates faster convergence for METHOD in terms of model evaluations relative to DADVI. The paths traced by κ^i_{METHOD} may be non-monotonic because the algorithms do not have access to $\tilde{\mathscr{Z}}$.

The optimization traces for ARM and non-ARM models are shown respectively in Figures 5 and 6, with suitably transformed axes for easier visualization. In many cases, the ADVI methods eventually find better optima (in terms of the variational objective) than DADVI, but ADVI typically takes longer to do so (the slower convergence is also shown in Figures 1

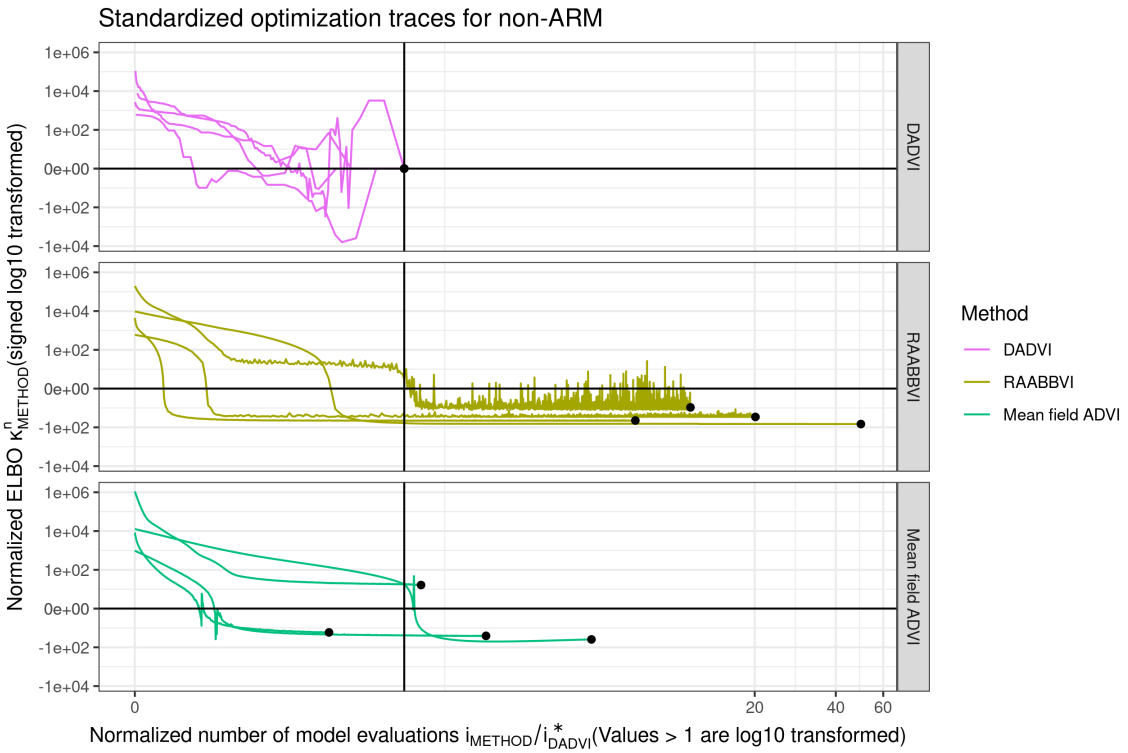

Figure 6: Traces for non-ARM models. Black dots show the termination point of each method. Dots above the horizontal black line mean that DADVI found a better ELBO. Dots to the right of the vertical black line mean that DADVI terminated sooner in terms of model evaluations.

and 2). As can be seen on the non-ARM models in Figure 6, the ADVI methods sometimes reach lower objective function values sooner than DADVI, but continue to optimize because they do not have access to the computationally expensive $\widehat{\mathscr{L}_{\text{VI}}}\left(\eta^i_{\text{METHOD}}|\widehat{\mathscr{Z}}\right)$ and have not detected convergence according to their own criteria. Similarly, DADVI sometimes finds lower values of $\mathscr{L}_{\text{VI}}(\cdot)$ along its path to optimization, but does not terminate because these points correspond to sub-optimal values of $\widehat{\mathscr{L}_{\text{VI}}}(\cdot|\widehat{\mathscr{Z}})$.

The results in Figures 5 and 6 suggest the possibility of initializing ADVI with DADVI and then optimizing further with stochastic methods in cases when low values of the objective function are of interest. However, as seen in Section 6.3 above, lower values of the variational objective do not necessarily translate into better posterior moment estimates.

6.5 Sampling variability

We next show that frequentist standard error estimates from DADVI provided good estimates of the sampling variability of the DADVI mean estimates, particularly for $N \geq 32$.

As discussed in Section 3.2, the sampling variability of DADVI estimates are straightforward to compute using standard formulas for the sampling variability of M-estimators. For the DADVI mean estimates, we computed the sampling standard deviation as described in

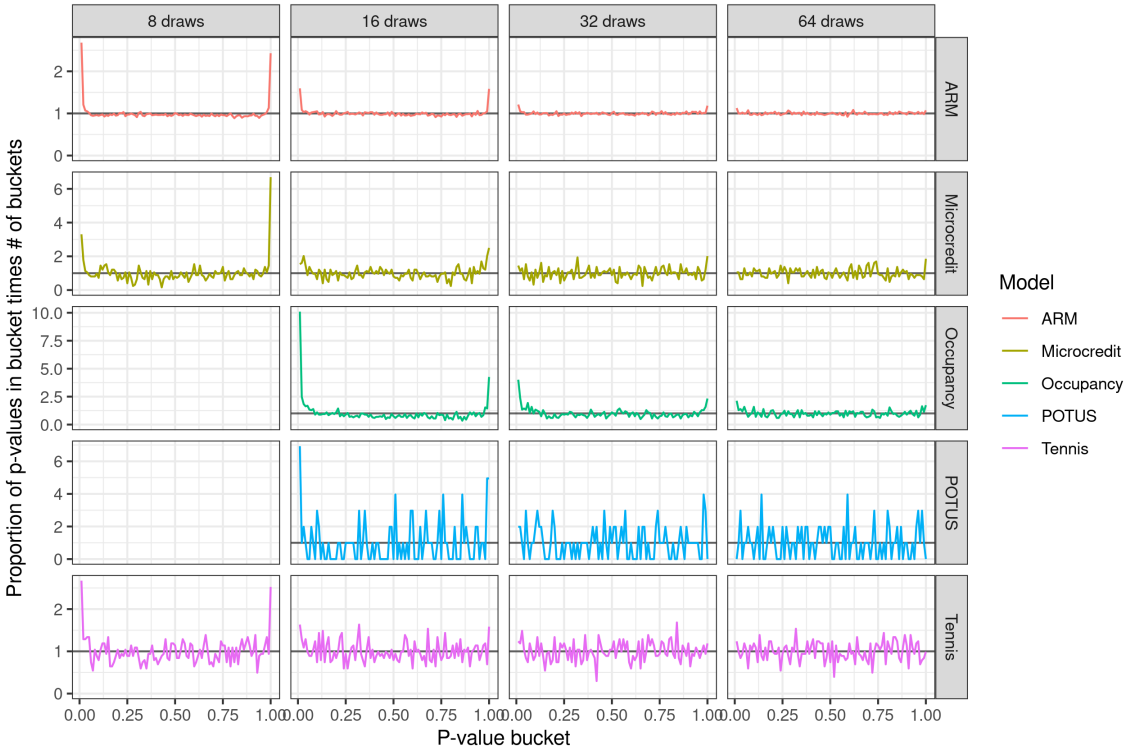

Figure 7: Density estimates of $\Phi(\varepsilon^\xi)$ for difference models. All the ARM models are grouped together for ease of visualization. Each panel shows a binned estimate of the density of $\Phi(\varepsilon^\xi)$ for a particular model and number of draws N. Values close to one (a uniform density) indicate good frequentist performance. CG failed for the Occupancy and POTUS models with only 8 draws, possibly indicating poor optimization performance with so few samples.

Sections 3.2 and 3.3.[12] We denote by ξ our estimate of $\sqrt{\underset{\mathcal{N}_{\text{std}}(Z)}{\text{Var}}(\mu_{\text{DADVI}})}$ as computed using Equation (14), that is, of the sampling standard deviation of the DADVI mean estimate under sampling of $\mathscr{Z}$. We can evaluate the accuracy of ξ by computing μ_{DADVI} with a large number of draws, which we denote as μ_∞, and checking whether

$$\varepsilon^\xi := \frac{\mu_{\text{DADVI}} - \mu_\infty}{\xi}$$

has an approximately standard normal distribution under many draws of μ_{DADVI}. We evaluated μ_∞ by taking the average of 100 runs with $N = 64$ each.[13]

To evaluate whether ε^ξ has a normal distribution, we can take Φ to be the cumulative distribution function of the standard normal distribution, and check whether $\Phi(\varepsilon^\xi)$ has a

[12] For the large POTUS, Occupancy, and Tennis models, we used CG to compute frequentist coverage for the same select quantities of interest for which we computed LR covariances.

[13] The values shown in the $N = 64$ panel of Figure 7 are the same as those whose average was taken to estimate μ_∞. In theory, this induces some correlation between the ε^ξ values for $N = 64$. However, the sampling variability of μ_∞ was so small that the induced correlation is practically negligible.

uniform distribution. Since the parameters returned from a particular model are not independent under sampling from $\mathscr{Z}$, the $\Phi(\varepsilon^{\xi})$ are not independent, and standard tests of uniformity like the Kolmogorov-Smirnov test are not valid. However, we can visually inspect the quality of the standard errors by checking whether $\Phi(\varepsilon^{\xi})$ has an approximately uniform distribution, without attempting to quantify how close it should be to uniform by chance alone. As can be seen in Figure 7, for $N = 8$ and $N = 16$ the $\Phi(\varepsilon^{\xi})$ values are over-dispersed to varying degrees for different models; this behavior indicates that the sampling variance ξ is under-estimated. In contrast, the intervals provide good marginal coverage when $N \geq 32$, though some over-dispersion remains in the Occupancy model.

7 Conclusion

In this paper, we proposed performing deterministic optimization on an approximate objective instead of using traditional stochastic optimization on the intractable objective from the mean-field ADVI problem. We found that using our DADVI approach can be faster, more accurate, and more automatic. The benefits of a deterministic objective can be attributed to the ability to use off-the-shelf second-order optimization algorithms with simple convergence criteria and linear response covariances. Additionally, the use of a deterministic objective allows computation of Monte Carlo sampling errors for the resulting approximation. And these errors can facilitate an explicit tradeoff between computation and accuracy. In contrast to the worst-case analyses in the optimization literature, we show theoretically that the number of samples needed for the deterministic objective need not scale linearly in the dimension in types of statistical models commonly encountered in practice. Although a deterministic objective cannot be used with highly expressive approximating families (such as full-rank ADVI), there is reason to believe that deterministic objectives can provide practical benefits for many black-box variational inference problems.

8 Acknowledgements

Ryan Giordano and Tamara Broderick were supported in part by an NSF CAREER Award and an ONR Early Career Grant. We are indebted to Ben Recht and Jonathan Huggins for helpful discussions and suggestions. We are also grateful for the feedback from our anonymous reviewers. All mistakes are our own.

References

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. Chapter 10.

D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

T. Broderick, R. Giordano, and R. Meager. An automatic finite-sample robustness metric: When can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999*, 2020.

J. Burroni, J. Domke, and D. Sheldon. Sample average approximation for Black-Box VI. *arXiv preprint arXiv:2304.06803*, 2023.

A. Dhaka, A. Catalina, M. Andersen, M. Magnusson, J. Huggins, and A. Vehtari. Robust, accurate stochastic optimization for variational inference. *Advances in Neural Information Processing Systems*, 33:10961–10973, 2020.

J. Domke and D. Sheldon. Importance weighting and variational inference. *Advances in Neural Information Processing Systems*, 31, 2018.

J. Domke and D. Sheldon. Divide and couple: Using Monte Carlo variational objectives for posterior approximation. *Advances in Neural Information Processing Systems*, 32, 2019.

R. Dudley. *Real analysis and probability*. CRC Press, 2018.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2006. doi: 10.1017/CBO9780511790942.

R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In *International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019.

R. Giordano, R. Liu, M. I. Jordan, and T. Broderick. Evaluating sensitivity to the stick-breaking prior in bayesian nonparametrics (with discussion). *Bayesian Analysis*, 18(1): 287–366, 2023.

R. J. Giordano, T. Broderick, and M. I. Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449, 2015.

T. Giordano, T. Broderick, and M. I. Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 19(51):1–49, 2018.

M. Heidemanns, A. Gelman, and G. Morris. An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election. *Harvard Data Science Review*, 2(4), 10 2020. doi: 10.1162/99608f92.fc62f1e1. URL `https://hdsr.mitpress.mit.edu/pub/nw1dzd02`.

M. Hoffman and A. Gelman. The No-U-Turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

M. Hoffman, D. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

T. Homem-de Mello and G. Bayraksan. Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.

J. Huggins, M. Kasprzak, T. Campbell, and T. Broderick. Validated variational inference via practical posterior error bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 1792–1802. PMLR, 2020.

M. Ingram, D. Vukcevic, and N. Golding. Scaling multi-species occupancy detection models to large citizen science datasets. In preparation, 2022.

M. Kery and A. Royle. *Inference about species richness and community structure using species-specific occupancy models in the National Swiss Breeding Bird Survey MUB*, pages 639–656. Modeling demographic processes in marked populations. Springer, New York and London, 2009. URL `http://pubs.er.usgs.gov/publication/5211455`.

S. Kim, R. Pasupathy, and S. Henderson. A guide to sample average approximation. *Handbook of simulation optimization*, pages 207–243, 2015.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. Blei. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017.

D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

C. Margossian and L. Saul. The shrinkage-delinkage trade-off: An analysis of factorized gaussian approximations for variational inference. *arXiv preprint arXiv:2302.09163*, 2023.

R. Meager. Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, January 2019. doi: 10.1257/app.20170299.

S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

J. Nocedal and S. Wright. *Numerical optimization*. Springer, 1999.

R. Ranganath, S. Gerrish, and D. Blei. Black Box Variational Inference. In *International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pages 814–822. PMLR, 2014.

D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

J. Royset and R. Szechtman. Optimal budget allocation for sample average approximation. *Operations Research*, 61(3):762–776, 2013.

J. Salvatier, T. Wiecki, and C. Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016. doi: 10.7717/peerj-cs.55. URL `https://doi.org/10.7717/peerj-cs.55`.

A. Shapiro. Monte Carlo sampling methods. *Handbooks in operations research and management science*, 10:353–425, 2003.

A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on stochastic programming: Modeling and theory*. SIAM, 2021.

L. Stefanski and D. Boos. The calculus of M-estimation. *The American Statistician*, 56(1): 29–38, 2002.

R. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In D. Barber, A. T. Cemgil, and S. Chiappa, editors, *Bayesian Time Series Models*. Cambridge University Press, 2011.

A. Van der Vaart and J. Wellner. *Weak convergence and empirical processes: With applications to statistics*. Springer Science & Business Media, 2013.

M. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

M. Welandawe, M. Andersen, A. Vehtari, and J. Huggins. Robust, automated, and accurate black-box variational inference. *arXiv preprint arXiv:2203.15945*, 2022.

S. Wright and J. Nocedal. *Numerical Optimization*, volume 35. Springer Series in Operation Research and Financial Engineering, 1999.

N. Wycoff, A. Arab, K. Donato, and L. Singh. Sparse bayesian lasso via a variable-coefficient ℓ_1 penalty. *arXiv preprint arXiv:2211.05089*, 2022.

L. Zhang, B. Carpenter, A. Gelman, and A. Vehtari. Pathfinder: Parallel quasi-newton variational inference. *The Journal of Machine Learning Research*, 23(1):13802–13850, 2022.

A Elaboration on the mean-field assumption

In practice, the mean-field assumption in variational inference need not always correspond to factorization over every single one-dimensional component of each parameter. Rather, it often represents a factorization into individual parameters as described in a model. For instance, consider a parameter within a model that represents a distribution over K outcomes, so that its elements are positive and sum to one. A natural prior for such a parameter might be a Dirichlet distribution. If this parameter exists as one parameter among multiple parameters in our model, a mean-field assumption will typically provide a separate factor for this parameter, but it will not further factorize across components within the parameter. So $\Sigma(\eta)$ may, in fact, be block-diagonal rather than purely diagonal, where each block size will correspond to the size of a parameter.

Researchers have explored other options between the extremes of the mean-field and full-rank assumptions for Gaussian approximations within variational inference; see, for instance, [Zhang et al., 2022].

B Behavior of high-dimensional normals

B.1 Proof of Theorem 1

We begin by deriving the DADVI optimal estimates. Let $\bar{z} := \frac{1}{N}\sum_{n=1}^{N} Z_n$ and $\overline{zz^\intercal} := \frac{1}{N}\sum_{n=1}^{N} Z_n Z_n^\intercal$. Also, let $S := \text{Diag}\,(\sigma)$, noting that $Sv = \sigma \odot v$ for any vector v. We can write $\theta_n = \mu + SZ_n$, so

$$\widehat{\mathop{\mathbb{E}}_{\mathscr{Z}}}\left[\theta(Z,\eta)\right] = \mu + S\bar{z} \quad \text{and} \quad \widehat{\mathop{\mathbb{E}}_{\mathscr{Z}}}\left[\theta(Z,\eta)\theta(Z,\eta)^\intercal\right] = \mu\mu^\intercal + \mu\bar{z}^\intercal S + S\bar{z}\mu^\intercal + S\overline{zz^\intercal}S,$$

so

$$\widehat{\mathscr{L}_{\text{VI}}}\left(\eta\right) = \frac{1}{2}\mu^\intercal A\left(\mu + 2S\bar{z}\right) + \frac{1}{2}\text{Tr}\left(AS\overline{zz^\intercal}S\right) - B^\intercal(\mu + S\bar{z}) - \sum_{d=1}^{D_\theta}\log\sigma_d. \tag{20}$$

For a fixed σ (and so a fixed S), the DADVI optimal mean parameter then satisfies

$$A\left(\hat{\mu} + S\bar{z}\right) - B = 0 \quad \Rightarrow \quad \hat{\mu} = A^{-1}B - S\bar{z} = \mu^* - S\bar{z}. \tag{21}$$

Thus, for any particular entry d, $\hat{\mu}_d - \mu_d^* = O_p(N^{-1/2})$ as long as $\sigma_d = O_p(1)$, both as the number of samples, N, goes to infinity.

We can now turn to the behavior of $\hat{\sigma}$. By plugging $\hat{\mu}$ as a function of S, which is given by Equation (21), into each term of Equation (20) that depends on μ, we get

$$\begin{aligned}
\frac{1}{2}\hat{\mu}^\intercal A\left(\hat{\mu} + 2S\bar{z}\right) &= \frac{1}{2}(\hat{\mu} + S\bar{z} - S\bar{z})^\intercal A\left(\hat{\mu} + S\bar{z} + S\bar{z}\right) \\
&= \frac{1}{2}(A^{-1}B - S\bar{z})^\intercal A\left(A^{-1}B + S\bar{z}\right) \\
&= \frac{1}{2}B^\intercal A^{-1}B - \frac{1}{2}\bar{z}^\intercal SAS\bar{z} \quad \text{and} \\
B^\intercal(\hat{\mu} + S\bar{z}) &= B^\intercal A^{-1}B.
\end{aligned}$$

Plugging the preceding two equations into the corresponding terms of Equation (20) gives, up to a constant C that does not depend on σ,

$$\widehat{\mathscr{L}_{\mathrm{VI}}}(\sigma) = \frac{1}{2}\mathrm{Tr}\left(AS\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)S\right) - \sum_{d=1}^{D_\theta} \log \sigma_d + C. \tag{22}$$

Let R denote the symmetric square root of the symmetric, positive definite A matrix (so $A = RR$ and $R = R^\intercal$). Then we have

$$\mathrm{Tr}\left(AS(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal)S\right) = \mathrm{Tr}\left(RS\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)(RS)^\intercal\right).$$

Let $\stackrel{d}{=}$ denote equality in distribution, i.e., $X \stackrel{d}{=} Y$ means that X and Y have the same law. Then

$$RSZ_n \stackrel{d}{=} (RSSR)^{1/2} z_n,$$

since both the left and the right hand sides of the preceding display have a $\mathcal{N}(\cdot|0_{D_\theta}, RSSR)$ distribution. (We have used the fact that S and R are both symmetric.) Thus, for any σ,

$$\widehat{\mathscr{L}_{\mathrm{VI}}}(\sigma) \stackrel{d}{=} \frac{1}{2}\mathrm{Tr}\left(RSSR\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)\right) - \frac{1}{2}\sum_{d=1}^{D_\theta} \log \sigma_d^2 + C. \tag{23}$$

Though the dependence on $\mathscr{X}$ of the left and right hand sides of the preceding equation is different, for a given σ, the two have the same distribution, and their optima have the same distribution as well. The product SS is simply $\mathrm{Diag}\left(\sigma^2\right)$, so expanding the trace gives

$$\begin{aligned}
\mathrm{Tr}\left(RSSR\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)\right) &= \sum_{i,j,k=1}^{D_\theta} R_{ij}\sigma_j^2 R_{jk}\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)_{ki} \Rightarrow \\
\frac{\partial}{\partial \sigma_d^2}\mathrm{Tr}\left(RSSR\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)\right) &= \sum_{i,k=1}^{D_\theta} R_{dk}\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)_{ki} R_{id} \\
&= \left(R\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)R^\intercal\right)_{dd}.
\end{aligned}$$

So the optimal value of σ_d^2 for the right hand side of Equation (23) is

$$\hat{\sigma}_d^2 = \frac{1}{\left(R\left(\overline{zz^\intercal} - \bar{z}\bar{z}^\intercal\right)R^\intercal\right)_{dd}}.$$

Note that $RZ_n \sim \mathcal{N}(\cdot|0_{D_\theta}, A)$. Therefore, if $w_n \sim \mathcal{N}(\cdot|0_{D_\theta}, A)$, then

$$\hat{\sigma}_d^{-2} \stackrel{d}{=} \frac{1}{N}\sum_{n=1}^{N} w_{nd}^2 - \left(\frac{1}{N}\sum_{n=1}^{N} w_{nd}\right)^2.$$

So $\mathop{\mathbb{E}}_{\mathcal{N}_{\mathrm{std}}(Z)}\left[\hat{\sigma}_d^{-2}\right] = \frac{N-1}{N}A_{dd} = \frac{N-1}{N}(\sigma_d^*)^{-2}$, and $\hat{\sigma}_d^{-2} - (\sigma_d^*)^{-2} = O_p(N^{-1/2})$. From this it follows that $\hat{\mu}_d - \mu_d^* = O_p(N^{-1/2})$ as well.

B.2 Proof of Theorem 2

Recall that the linear response covariance estimate for θ in this model considers the perturbed model

$$\log \mathcal{P}(\theta, y|t) := \log \mathcal{P}(\theta, y) + t^\intercal \theta$$

and computes

$$\underset{\mathcal{Q}(\theta|\hat{\eta})}{\widehat{\mathrm{LRCov}}}\,(\theta) = \left.\frac{d\hat{\mu}}{dt^\intercal}\right|_{\hat{\eta}} = A^{-1},$$

where the final equality follows from Equations (21) and (22) by identifying B with $B + t$ and observing that $\hat{\sigma}$ does not depend on t. Since A^{-1} is in fact the true posterior variance, the linear response covariance is exact in this case irrespective of how small N is, in contrast even to σ^*, which can be a poor estimate of the marginal variances unless A is diagonal.

C High-dimensional global–local problems

Proof. of Theorem 3.

We can write

$$\begin{aligned}
&\mathscr{L}_{\mathrm{VI}}(\hat{\eta}) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta}) = \\
&\quad \mathscr{L}_{\mathrm{VI}}(\hat{\eta}) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta}) + \widehat{\mathscr{L}_{\mathrm{VI}}}(\overset{*}{\eta}|\mathscr{Z}) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\overset{*}{\eta}|\mathscr{Z}) + \widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|\mathscr{Z}) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|\mathscr{Z}) = \\
&\quad \left(\mathscr{L}_{\mathrm{VI}}(\hat{\eta}) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|\mathscr{Z})\right) + \left(\widehat{\mathscr{L}_{\mathrm{VI}}}(\overset{*}{\eta}|\mathscr{Z}) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta})\right) + \left(\widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|\mathscr{Z}) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\overset{*}{\eta}|\mathscr{Z})\right) \leq \\
&\quad \left(\mathscr{L}_{\mathrm{VI}}(\hat{\eta}) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|\mathscr{Z})\right) + \left(\widehat{\mathscr{L}_{\mathrm{VI}}}(\overset{*}{\eta}|\mathscr{Z}) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta})\right) \leq \\
&\quad \left|\mathscr{L}_{\mathrm{VI}}(\hat{\eta}) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|\mathscr{Z})\right| + \left|\widehat{\mathscr{L}_{\mathrm{VI}}}(\overset{*}{\eta}|\mathscr{Z}) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta})\right| \leq \\
&\quad 2 \sup_{\eta \in \Omega_\eta} \left|\mathscr{L}_{\mathrm{VI}}(\eta) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z})\right| \qquad (24)
\end{aligned}$$

where the penultimate inequality uses the fact that $\widehat{\mathscr{L}_{\mathrm{VI}}}(\hat{\eta}|\mathscr{Z}) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\overset{*}{\eta}|\mathscr{Z}) \leq 0$. By Assumption 2, we then have

$$\|\hat{\eta}^\gamma - \overset{*}{\eta}{}^\gamma\|_2^2 \leq \frac{2}{PC_3} \sup_{\eta \in \Omega_\eta} \left|\mathscr{L}_{\mathrm{VI}}(\eta) - \widehat{\mathscr{L}_{\mathrm{VI}}}(\eta|\mathscr{Z})\right|. \qquad (25)$$

Similarly, for any given p, apply Assumption 2 with the components of η matching $\hat{\eta}^p$ in the components corresponding to the variational distribution for λ^p, and matching $\overset{*}{\eta}$ otherwise, giving

$$\bar{f}^p(\overset{*}{\eta}{}^\gamma, \hat{\eta}^p) - \bar{f}^p(\overset{*}{\eta}{}^\gamma, \overset{*}{\eta}{}^p) \geq C_3 \|\hat{\eta}^p - \overset{*}{\eta}{}^p\|_2^2. \qquad (26)$$

Since $\hat{\eta}^p$ minimizes $\eta^p \mapsto \hat{f}^p(\overset{*}{\eta}{}^\gamma, \eta^p)$, the same reasoning as Equation (24) implies that

$$\bar{f}^p(\overset{*}{\eta}{}^\gamma, \hat{\eta}^p) - \bar{f}^p(\overset{*}{\eta}{}^\gamma, \overset{*}{\eta}{}^p) \leq 2 \sup_{\eta^p \in \Omega_\eta} \left|\bar{f}^p(\overset{*}{\eta}{}^\gamma, \eta^p) - \hat{f}^p(\overset{*}{\eta}{}^\gamma, \mathscr{Z}^\gamma, \eta^p, \mathscr{Z}^p).\right|$$

Combining the previous two displays gives

$$\|\hat{\eta}^p - \overset{*}{\eta}{}^p\|_2^2 \leq \frac{2}{C_3} \sup_{\eta \in \Omega_\eta} \left| \bar{f}^p(\eta^\gamma, \eta^p) - \hat{f}^p(\eta^\gamma, \mathscr{Z}^\gamma, \eta^p, \mathscr{Z}^p) \right|.$$

Next, we use Assumption 1 to control the difference between the samples and limiting objectives. Take $\delta' = C_3\delta/2$. Let

$$\mathcal{E}^p := \sup_{\eta^\gamma, \eta^p} \left| \hat{f}^p(\eta^\gamma, \mathscr{Z}^\gamma, \eta^p, \mathscr{Z}^p) - f^p(\eta^\gamma, \eta^p) \right| \quad \text{and} \quad \mathcal{E} := \frac{1}{P} \sup_{\eta \in \Omega_\eta} \left| \widehat{\mathscr{L}_{\mathrm{VI}}}(\eta | \mathscr{Z}) - \mathscr{L}_{\mathrm{VI}}(\eta) \right|.$$

Since we can only increase the error by allowing the global parameter to vary separately for each local ULLN, we have $\mathcal{E} \leq \frac{1}{P}\sum_{p=1}^P \mathcal{E}^p$. Therefore, $\{\forall p : \mathcal{E}^p \leq \delta'\} \Rightarrow \{\mathcal{E} \leq \delta'\}$ and $\{\mathcal{E} > \delta'\} \Rightarrow \{\exists p : \mathcal{E}^p > \delta'\}$. A union bound then gives

$$\mathcal{P}(\mathcal{E} > \delta) \leq \mathcal{P}\left(\bigcup_p \left\{ \mathcal{E}^p > \delta' \right\} \right) \leq \sum_{p=1}^P \mathcal{P}\left(\mathcal{E}^p > \delta' \right) \leq C_1 \exp\left(-C_2 N + \log P \right) \leq \varepsilon, \tag{27}$$

where the final inequality follows from taking $N \geq N_0$ large enough to satisfy Assumption 1 and $N_0 \geq C_2^{-1}\left(\log P - \log\left(C_1^{-1}\varepsilon\right)\right)$.

By Equations (25) and (26),

$$\bigcap_{p=1}^P \left\{ \mathcal{E}^p < \delta' \right\} \Rightarrow \bigcap_{p=1}^P \left\{ \|\hat{\eta}^p - \overset{*}{\eta}{}^p\|_2^2 \leq \delta \right\} \text{ and } \bigcap_{p=1}^P \left\{ \mathcal{E}^p < \delta' \right\} \Rightarrow \mathcal{E} < \delta' \Rightarrow \|\hat{\eta}^\gamma - \overset{*}{\eta}{}^\gamma\|_2^2 \leq \delta.$$

The conclusion then follows from Equation (27). □

Proof. of Example 2.

Suppose that, for each p, $\bar{f}^p(\eta^\gamma, \eta^p)$ is twice-differentiable and convex, and the domain is compact. Let the first and second-order derivatives be denoted by $\nabla \bar{f}^p$ and $\nabla^2 \bar{f}^p$ respectively, and let C_3 lower bound the minimum eigenvalue of all $\nabla^2 \bar{f}^p$.

Then a Taylor series expansion with integral remainder gives

$$\bar{f}^p(\eta^\gamma, \eta^p) - \bar{f}^p(\overset{*}{\eta}{}^\gamma, \overset{*}{\eta}{}^p) = \nabla \bar{f}^p(\overset{*}{\eta}{}^\gamma, \overset{*}{\eta}{}^p) \begin{pmatrix} \eta^\gamma - \overset{*}{\eta}{}^\gamma \\ \eta^p - \overset{*}{\eta}{}^p \end{pmatrix} + R^p(\overset{*}{\eta}, \eta)$$

where

$$R^p(\overset{*}{\eta}, \eta) = \int_0^1 \begin{pmatrix} \eta^\gamma - \overset{*}{\eta}{}^\gamma \\ \eta^p - \overset{*}{\eta}{}^p \end{pmatrix}^\intercal \nabla^2 \bar{f}(\overset{*}{\eta}{}^\gamma + t(\eta^\gamma - \overset{*}{\eta}{}^\gamma), \overset{*}{\eta}{}^p + t(\eta^p - \overset{*}{\eta}{}^p)) \begin{pmatrix} \eta^\gamma - \overset{*}{\eta}{}^\gamma \\ \eta^p - \overset{*}{\eta}{}^p \end{pmatrix} (1-t) dt.$$

(Apply Dudley [2018, Theorem B.2] with $t \mapsto \bar{f}(\overset{*}{\eta}{}^\gamma + t(\eta^\gamma - \overset{*}{\eta}{}^\gamma), \overset{*}{\eta}{}^p + t(\eta^p - \overset{*}{\eta}{}^p))$.) Since $\overset{*}{\eta}$ is an optimum, $\sum_{p=1}^P \nabla \bar{f}^p(\overset{*}{\eta}{}^\gamma, \overset{*}{\eta}{}^p) = 0$. Since $\nabla^2 \bar{f}^p$ is positive definite for every p, there exists a $C_3 \geq 0$ such that

$$R^p(\overset{*}{\eta}, \eta) \geq C_3 \left(\|\eta^\gamma - \overset{*}{\eta}{}^\gamma\|_2^2 + \|\eta^p - \overset{*}{\eta}{}^p\|_2^2 \right).$$

It follows that

$$\mathscr{L}_{\mathrm{VI}}(\eta) - \mathscr{L}_{\mathrm{VI}}(\overset{*}{\eta}) \geq C_3 \left(P \|\eta^\gamma - \overset{*}{\eta}{}^\gamma\|_2^2 + \sum_{p=1}^P \|\eta^p - \overset{*}{\eta}{}^p\|_2^2 \right),$$

from which Assumption 2 follows.

□

D Model details

D.1 ARM models

We selected 53 from the `Stan` example models repository[14]. The models we used are as follows, with their parameter dimension in parentheses:

`separation (2), wells_dist100 (2), nes2000_vote (2), wells_d100ars (3), earn_height (3), sesame_one_pred_b (3), radon_complete_pool (3), earnings1 (3), kidscore_momiq (3), kidscore_momhs (3), electric_one_pred (3), sesame_one_pred_a (3), sesame_one_pred_2b (3), logearn_height (3), electric_multi_preds (4), congress (4), wells_interaction_c (4), earnings2 (4), logearn_logheight (4), kidiq_multi_preds (4), wells_dae (4), wells_interaction (4), logearn_height_male (4), ideo_reparam (5), logearn_interaction (5), kidscore_momwork (5), kidiq_interaction (5), wells_dae_c (5), mesquite_volume (5), earnings_interactions (5), wells_dae_inter (5), wells_daae_c (6), wells_dae_inter_c (7), mesquite_vash (7), wells_predicted_log (7), mesquite (8), mesquite_vas (8), mesquite_log (8), sesame_multi_preds_3b (9), sesame_multi_preds_3a (9), pilots (17), election88 (53), radon_intercept (88), radon_no_pool (89), radon_group (90), electric (100), electric_1b (101), electric_1a (109), electric_1c (114), hiv (170), hiv_inter (171), radon_vary_si (174), radon_inter_vary (176).`

Some models were eliminated from consideration for being duplicates of other models, and a small number were eliminated for poor NUTS performance (low effective sample size or poor R hat).

D.2 Tennis

In the tennis model, each player, $i = 1, ..., M$ has a rating θ_i. These ratings are drawn from a prior distribution with a shared variance:

$$\theta_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \tag{28}$$

The standard deviation σ is given a half-Normal prior with a scale parameter of 1. The likelihood for a match $n = 1, ..., N$ between player i and j is given by:

$$y_n \sim \text{Bernoulli}(\text{logit}^{-1}(\theta_i - \theta_j)) \tag{29}$$

where $y_n = 1$ if player i won, and $y_n = 0$ if not.

D.3 Occupancy model

In occupancy models, we are interested in whether site i is occupied by species j. We model occupation as a binary latent variable y_{ij}, with probability Ψ_{ij} being the probability that the species is occupying the site. The logit of this probability is modeled as a linear function of environmental covariates, such as rainfall and temperature:

[14] `https://github.com/stan-dev/example-models/tree/master/ARM`

$$y_{ij} \sim \text{Bern}(\Psi_{ij}), \tag{30}$$

$$\text{logit}(\Psi_{ij}) = \boldsymbol{x}_i^{(\text{env})\intercal} \boldsymbol{\beta}_j^{(\text{env})} + \gamma_j, \tag{31}$$

$$\boldsymbol{\beta}_j^{(\text{env})} \overset{iid}{\sim} \mathcal{N}(0, I), \tag{32}$$

$$\gamma_j \overset{iid}{\sim} \mathcal{N}(0, 10^2). \tag{33}$$

However, y_{ij} is assumed not to be observed directly. Instead, we observe the binary outcome s_{ijk}, which equals one if species j was observed at site i on the k-th visit. If the species was observed, we know that it is present ($y_{ij} = 1$), assuming there are no false positives. If it was not, it may have been missed, and we model the probability that it would have been observed if it had been present, p_{ijk}. Mathematically speaking, these assumptions result in the following model:

$$p(s_{ijk} = 1 \mid y_{ij} = 1) = p_{ijk}, \tag{34}$$

$$p(s_{ijk} = 1 \mid y_{ij} = 0) = 0, \tag{35}$$

$$\text{logit}(p_{ijk}) = \boldsymbol{x}_{ik}^{(\text{obs})\intercal} \boldsymbol{\beta}_j^{(\text{obs})}, \tag{36}$$

where $\boldsymbol{x}_{ik}^{(\text{obs})\intercal}$ are a set of covariates assumed to be related to the probability of observing the species, and $\boldsymbol{\beta}_j^{(\text{obs})}$ are coefficients of a linear model relating these to the logit of the probability p_{ijk}.

As y_{ij} is not observed, it has to be marginalized out for ADVI models to be applicable. The resulting likelihood is given by:

$$p(s \mid \theta) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[(1 - \Psi_{ij}) \prod_{k=1}^{K_i} (1 - s_{ijk}) + \Psi_{ij} \prod_{k=1}^{K_i} (p_{ijk})^{s_{ijk}} (1 - p_{ijk})^{1 - s_{ijk}} \right]. \tag{37}$$

Its derivation can be found in the appendix of Ingram et al. [2022]. Here, K_i are the number of visits to site i, N is the total number of sites, J is the total number of species, and the rest of the variables are as defined previously.

E Preconditioning DADVI

As described in Section 3.3, in high-dimensional problems it is useful to use the conjugate gradient (CG) algorithm to compute both LR covariances and frequentist standard errors. The CG algorithm uses products of the form $\hat{\mathcal{H}} v$ to approximately solve $\hat{\mathcal{H}}^{-1} v$, and can be made more efficient with a preconditioning matrix M with $M \approx \hat{\mathcal{H}}^{-1}$ [Wright and Nocedal, 1999, Chapter 5].

The DADVI approximation itself provides an approximation to the upper left quadrant of $\hat{\mathcal{H}}^{-1}$, which can be used as a preconditioner. By the LR covariance formula Equation (11),

$$\underset{\mathcal{P}(\theta|y)(\theta)}{\text{Cov}}(\theta) \approx \underset{\mathcal{Q}(\theta|\hat{\eta})}{\widehat{\text{LRCov}}}(\theta) = \begin{pmatrix} I_{D_\theta} & 0_{D_\theta \times D_\theta} \end{pmatrix} \hat{\mathcal{H}}^{-1} \begin{pmatrix} I_{D_\theta} \\ 0_{D_\theta \times D_\theta}, \end{pmatrix},$$

which is just the upper-left quadrant of $\hat{\mathcal{H}}^{-1}$. Prior to computing the LR covariances, the best available approximation of $\underset{\mathcal{P}(\theta|y)(\theta)}{\text{Cov}}(\theta)$ — and, in turn, the upper-left quadrant of $\hat{\mathcal{H}}^{-1}$ — is the mean-field covariance estimate $\underset{\mathcal{Q}(\theta|\hat{\eta})}{\text{Cov}}(\theta) = \text{Diag}\left(\exp(\hat{\xi}_1), \ldots, \exp(\hat{\xi}_{D_\theta})\right)$. Therefore, whenever using CG on a DADVI optimum, we pre-condition with the matrix

$$\begin{pmatrix} \underset{\mathcal{Q}(\theta|\hat{\eta})}{\text{Cov}}(\theta) & 0_{D_\theta \times D_\theta} \\ 0_{D_\theta \times D_\theta} & I_{D_\theta} \end{pmatrix}.$$

Using the preceding preconditioner is formally similar to re-parameterizing the mean parameters into their natural parameters, as when taking a natural gradient in stochastic optimization [Hoffman et al., 2013].