Université Paris Dauphine – PSL

Master 1 in Applied Mathematics

Statistics

Testing New Variational Approximations for Bayesian Inference

JIN Ying & MESSELU Bethlehem

April 28th 2025

Supervised by:

Joshua J. Bon

Contents

1 Introduction

Bayesian statistics addresses inference problems by treating unknown quantities as random variables. This approach updates prior beliefs with observed data using Bayes' theorem, producing the posterior distribution used for prediction and decision-making. Bayesian theory has become a central pillar of modern machine learning (ML) models to quantify uncertainty, something point-estimate ML models cannot offer. Yet, computing the posterior is often computationally infeasible and rarely straightforward: The normalizing constant involves an integral that is often high-dimensional, analytically intractable, and typically impossible to evaluate with standard numerical methods. Consequently, finding practical ways to approximate these difficult-to-compute posteriors is a fundamental aspect of Bayesian statistics.

Markov-chain Monte Carlo (MCMC) tackles the bottleneck by replacing analytical integration with stochastic simulation. In MCMC, an ergodic Markov chain of parameters is constructed with stationary distribution equal to that of the posterior. After sampling this chain, expectations with respect to the posterior are approximated using these samples. The resulting estimates are consistent and—given enough time—approximately unbiased, making MCMC the gold-standard for accuracy. However, chain-based methods can be painfully slow in high dimensions, and demand diagnostic vigilance to assess convergence (Brooks & Roberts 1998), which limit their scalability for large or complex data sets.

Variational inference (VI) frames posterior approximation as an optimization problem. It involves approximating the posterior from a tractable family of distributions $\mathcal{Q}$ and aims to identify the member of this family that minimizes the Kullback-Leibler (KL) divergence from the exact posterior. A crucial aspect of variational inference is selecting $\mathcal{Q}$ to be sufficiently flexible to approximate the exact posterior closely, yet simple enough to allow efficient optimization.

VI trades accuracy for speed of inference. It can deliver good-enough estimates of the target values quickly when MCMC may take orders of magnitude more time to achieve similar results. Variational inference is well-suited for large datasets and situations requiring rapid model exploration. In contrast, MCMC is more appropriate for smaller datasets where the computational expense is justified by acquiring more precise samples (Blei et al. 2017).

Building on the optimisation view of Bayesian inference, recent approaches such as "black-box variational inference" (BBVI) methods further widen the reach of variational methods. BBVI removes model-specific analytical derivations by estimating gradients with Monte Carlo samples that work for any model with a tractable joint density (Ranganath et al. 2014). Automatic Differentiation Variational Inference (ADVI), a particularly widely used variant of BBVI, improves on this idea by re-parameterizing latent variables to obtain lower variance (Kucukelbir et al. 2017). Deterministic Automatic-differentiation Variational Inference (DADVI) does not try to approximate the gradient. Instead, it re-

lies on an approximation of the same objective function as ADVI, using a fixed sample throughout the whole process. This fixes the stochasticity beforehand, making the optimization deterministic and more stable (Giordano et al. 2024). Comparing DADVI with BBVI and ADVI will show whether this determinism translates more accurate posterior estimates, and greater robustness in the high-dimensional models, which is the motivation of this study.

In what follows, we start with setting up the variational inference problem, followed by a detailled explanation of ADVI and DADVI. The objectives of the study are then stated explicitly. Subsequent sections present the experimental methodology and results, analyze those findings, and summarize the principal conclusions. The report closes with a critical assessment of the study's limitations and a discussion of promising directions for future work.

2 Problem set up and relevant literature

We first set up the general problem. Let $x = x_{1:n}$ denote the observed data, and let $\theta \in \mathbb{R}^{D_\theta}$ represent the unknown parameters. The objective is to approximate the conditional density of unknown parameters given observed data, according to Bayes rule:

$$p(\theta \mid x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x \mid \theta)\, p(\theta)}{\int p(\theta, x)\, d\theta} \tag{1}$$

where:

- $p(\theta)$ is the prior distribution of latent variables,
- $p(x \mid \theta)$ is the likelihood function,
- $p(\theta \mid x)$ is the posterior distribution,
- $p(x) = \int p(\theta, x)\, d\theta$ is the normalizing constant.

As discussed, the normalizing constant is hard to compute. Variational inference addresses this by finding the member from a family of distributions that is closest to the true posterior using the Kullback–Leibler divergence:

$$q^*(\theta) = \arg\min_{q(\theta) \in \mathcal{Q}} \mathrm{KL}(q(\theta) \,\|\, p(\theta \mid x)) \tag{2}$$

where the KL divergence is given by:

$$\mathrm{KL}(q(\theta) \,\|\, p(\theta \mid x)) := \int q(\theta) \log\left(\frac{q(\theta)}{p(\theta \mid x)}\right) d\theta. \tag{3}$$

In our work, we consider a variational family $\mathcal{Q}$ composed of multivariate normal distributions under the **mean-field assumption**. This assumption implies

that the approximating distribution factorizes over dimensions, i.e., it is a product of independent univariate normal distributions. Specifically, the variational parameters are the mean $\mu = (\mu_1, \ldots, \mu_{D_\theta})^\top \in \mathbb{R}^{D_\theta}$ and standard deviation $\sigma = (\sigma_1, \ldots, \sigma_{D_\theta})^\top \in \mathbb{R}_+^{D_\theta}$,

$$\mathcal{Q} = \left\{ q(\theta) \in \mathcal{P}^{D_\theta} : q(\theta) = \prod_{d=1}^{D_\theta} \mathcal{N}(\theta_d \mid \mu_d, \sigma_d^2) \right\}. \tag{4}$$

where $\mathcal{P}^{D_\theta}$ is the class of density functions on $\mathbb{R}^{D_\theta}$.

If we decompose the KL divergence, we have:

$$\begin{aligned}
\mathrm{KL}(q(\theta) \,\|\, p(\theta \mid x)) &= \int q(\theta) \log \left(\frac{q(\theta)}{p(\theta \mid x)} \right) d\theta \\
&= \mathbb{E}_{q(\theta)} \left[\log \left(\frac{q(\theta)}{p(\theta \mid x)} \right) \right] \\
&= \mathbb{E}_{q(\theta)}[\log q(\theta)] - \mathbb{E}_{q(\theta)}[\log p(\theta \mid x)] \\
&= \mathbb{E}_{q(\theta)}[\log q(\theta)] - \mathbb{E}_{q(\theta)}[\log p(\theta, x)] + \log p(x).
\end{aligned} \tag{5}$$

Because $\log p(x)$ does not depend on θ, we can use the equivalent objective function:

$$\mathcal{L}(q) := \mathbb{E}_{q(\theta)}[\log q(\theta)] - \mathbb{E}_{q(\theta)}[\log p(\theta, x)], \tag{6}$$

which is commonly referred to as the Evidence Lower Bound (ELBO) when negated:

$$\mathrm{ELBO}(q) = -\mathcal{L}(q). \tag{7}$$

Thus, minimizing $\mathrm{KL}[q(\theta) \| p(\theta \mid x)]$ is equivalent to minimizing $\mathcal{L}(q)$ or maximizing $\mathrm{ELBO}(q)$, i.e.,

$$\arg \min_{q(\theta) \in \mathcal{Q}} \mathrm{KL}(q(\theta) \,\|\, p(\theta \mid x)) = \arg \min_{q(\theta) \in \mathcal{Q}} \mathcal{L}(q) = \arg \max_{q(\theta) \in \mathcal{Q}} \mathrm{ELBO}(q). \tag{8}$$

Using the mean field assumption, we can furthur simplify the objective function:

$$\mathcal{L}(q) = -\sum_{d=1}^{D_\theta} \log \sigma_d - \mathbb{E}_{q(\theta)} \left[\log p(\theta, x) \right]. \tag{9}$$

Given that examining the ELBO gives intuitions about the optimal variational density, in our experiments, we use the ELBO as a key performance metric to evaluate the effectiveness of ADVI and DADVI methods.

2.1 The Reparametrization Trick

In variational inference, $\mathcal{L}(q)$ is often hard to estimate due to the $\mathbb{E}_{q(\theta)}[\log p(\theta, x)]$ term. A common approach, used in both ADVI and DADVI, is to approximate this objective via an unbiased Monte Carlo estimator and to apply the reparametrization trick (Xu et al. 2019). This approximation provides a common framework for the ADVI and DADVI algorithms.

We define the Monte Carlo estimate of $\mathcal{L}(q)$ (negative ELBO) as:

$$\widehat{\mathcal{L}}(q \mid Z_{1:N}) = -\sum_{d=1}^{D_\theta} \log \sigma_d - \frac{1}{N} \sum_{n=1}^{N} \log p(\theta(Z_n), x) \tag{10}$$

where:

- $Z_n \sim \mathcal{N}(0_{D_\theta}, I_{D_\theta})$, for all $n \in \{1, \ldots, N\}$,
- $\theta(Z) = \mu + Z \odot \sigma$, for all $n \in \{1, \ldots, N\}$ (reparametrization trick),
- $\mu = (\mu_1, \ldots, \mu_{D_\theta})^\top$ represents the mean of the variational distribution,
- $\sigma = (\sigma_1, \ldots, \sigma_{D_\theta})^\top$ represents the standard deviation of the variational distribution,
- $\odot$ denotes the Hadamard (element-wise) product.

In the next sections, we will build upon this formulation to describe two distinct optimization approaches: ADVI, which makes use of stochastic gradient descent (SGD), and DADVI, which relies on a sample average approximation (SAA) strategy.

2.2 Automatic-differentiation variational inference

The ADVI algorithm relies on an optimization strategy based on stochastic gradient descent. This method requires several key components. First, the variational parameters to optimize at iteration k are defined as:

$$\boldsymbol{\phi}^{(k)} = \left(\mu_1^{(k)}, \ldots, \mu_{D_\theta}^{(k)}, \left(\sigma_1^2\right)^{(k)}, \ldots, \left(\sigma_{D_\theta}^2\right)^{(k)}\right),$$

The first D_θ components correspond to the mean parameters, and the remaining D_θ components correspond to the variances of the diagonal Gaussian variational distribution.

For simplicity, we use a fixed learning rate $\tau > 0$, which controls the step size of the parameter updates at each iteration. At each step, a random sample Z_k is independently drawn from the standard multivariate normal distribution $\mathcal{N}(0_{D_\theta}, I_{D_\theta})$, introducing stochasticity into the gradient estimation.

The gradient of the approximate negative ELBO (Eq. (10) with $N = 1$), denoted $\nabla_{\boldsymbol{\phi}^{(k)}} \widehat{\mathcal{L}}(q \mid Z_k)$, is then computed with respect to the current variational parameters. Finally, the parameters are updated according to the rule:

$$\boldsymbol{\phi}^{(k)} = \boldsymbol{\phi}^{(k-1)} - \tau \cdot \nabla_{\boldsymbol{\phi}^{(k)}} \widehat{\mathcal{L}}(q \mid Z_k).$$

Based on this optimization heuristic, the full algorithmic procedure for ADVI is presented in the following Algorithm (Giordano et al. 2024).

Algorithm 1: ADVI

Input: Initial parameters $\boldsymbol{\phi}_0$, max iterations M, step size τ, tolerance ε, number of samples $N = 1$ (default)
Output: Optimized variational parameters $\boldsymbol{\phi}^{(k)}$

$k \leftarrow 0$;
while $\|\boldsymbol{\phi}^{(k)} - \boldsymbol{\phi}^{(k-1)}\| \geq \varepsilon$ ***and*** $k < M$ **do**
  $k \leftarrow k + 1$;
  Draw Z_k ;
  Compute gradient $\nabla_{\boldsymbol{\phi}^{(k)}} \widehat{\mathcal{L}}(q \mid Z_k)$;
  Update parameters: $\boldsymbol{\phi}^{(k)} = \boldsymbol{\phi}^{(k-1)} - \tau \cdot \nabla_{\boldsymbol{\phi}^{(k)}} \widehat{\mathcal{L}}(q \mid Z_k)$;
return $\boldsymbol{\phi}^{(k)}$

2.3 Deterministic Automatic-differentiation variational inference

The DADVI method relies on sample average approximation (SAA) for the optimization over the variational parameters. This strategy transforms a stochastic gradient optimisation problem into a deterministic optimisation, conditional on the noise $Z_{1:N}$, since it approximates the objective function beforehand. The following paragraph outlines the components necessary to optimize with respect to the variational parameters using the DADVI method. We keep the same notation as in the previous section.

A sample average approximation of the objective function is used, which coincides with $\widehat{\mathcal{L}}(q \mid Z_{1:N})$. In addition, a deterministic optimization algorithm is used; in our case, we rely on gradient descent. Based on this optimization heuristic, the full algorithmic procedure for DADVI is presented in the following figure (Giordano et al. 2024).

Algorithm 2: DADVI

Input: Initial parameters $\boldsymbol{\phi}_0$, max iterations M, step size τ, tolerance ε, number of samples(chosen to match ADVI's convergence in our work)

Output: Optimized variational parameters $\boldsymbol{\phi}^{(k)}$

$k \leftarrow 0$;
Draw $Z_{1:N}$;
while $\|\boldsymbol{\phi}^{(k)} - \boldsymbol{\phi}^{(k-1)}\| \geq \varepsilon$ ***and*** $k < M$ **do**
    $k \leftarrow k + 1$;
    Compute gradient $\nabla_{\boldsymbol{\phi}^{(k)}} \widehat{\mathcal{L}}(q \mid Z_{1:N})$;
    Update parameters: $\boldsymbol{\phi}^{(k)} = \boldsymbol{\phi}^{(k-1)} - \tau \cdot \nabla_{\boldsymbol{\phi}^{(k)}} \widehat{\mathcal{L}}(q \mid Z_{1:N})$;
return $\boldsymbol{\phi}^{(k)}$

In practice, ADVI and DADVI differ mainly in how they treat the randomness that enters the gradient. ADVI redraws one standard-normal noise vector at every iteration, so each update is cheap but noisy; this can slow convergence unless the step size is tuned with care. DADVI instead samples a modest batch of noise vectors once at the start and keeps them fixed, turning the stochastic objective into a smooth, deterministic one. DADVI follows a far less variable gradient path. We anticipate that DADVI will reach a stable optimum in fewer iterations and yield a more accurate posterior approximation, while ADVI will remain attractive when memory is limited, and/or rough optimisation is sufficient.

Our work critically compares two methods, ADVI and DADVI, across synthetic and real-data experiments. It aims to address some key questions. First, in terms of accuracy, we assess how closely each algorithm approximates the reference posterior in cases where the exact posterior is available in closed form. Second, we investigate how the methods perform as the dimensionality of the problem increases. Third, regarding stability, we examine whether either method is prone to getting trapped in local optima or encountering numerical pathologies. Finally, in terms of generality, we assess whether any observed advantages persist when the posterior becomes multi-modal or when applied to real-world data.

3 Experiments

In this section, we conduct a series of experiments to compare the performance of ADVI and DADVI. We begin with cases where the true posterior is a Gaussian distribution, considering dimensions from 1 to 4 to investigate how both methods perform across different dimensional settings. We then extend the study to scenarios where the true posterior is a mixture of Gaussian distributions. Finally, we apply both methods to a real-world dataset examining the effectiveness of a beta-blocker through a meta-analysis.

To assess performance, when the true posterior is available, we directly compare it with the approximations produced by ADVI and DADVI to evaluate accuracy. In all cases, we also compare the ELBO values achieved by the two algorithms.

It is noted that both algorithms must be evaluated with equivalent computational complexity for the comparison to be meaningful. To achieve this, we first implement ADVI and record the number of iterations N required for convergence. We then use this number as the sample size for the Monte Carlo estimation in DADVI. We also report the estimated ELBO value with using N iterations.

For each experimental setup, we selected learning rates that yielded reasonable results—that is, variational approximations sufficiently close to the true posterior when it is available in closed form.

You can review and clone the analysis code here:
github.com/PineappleBlowsnow/Dauphine-M1-Memoire_ADVI-vs-DADVI.

3.1 One and two dimensional variational families

For the experiments on simple Gaussian distributions from different dimensional variational families, we will test both algorithms on an artificial dataset that we create based on the setup developed by Jacobs (2008).

3.1.1 Computation of the true posterior

We assume that the parameter of interest is μ, which has a prior distribution $\mu \sim \mathcal{N}(M, \tau^2)$, where $M \in \mathbb{R}$ is the prior mean, and $\tau^2 > 0$ is the prior variance. We further assume that the likelihood is independent and normally distributed identically $x_{1:n} \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$.

Given prior and likelihood, the true posterior is computed as follows:

$$
\begin{aligned}
p(\mu \mid x_{1:n}) &\propto p(\mu) \cdot p(x_{1:n} \mid \mu) \\
&= \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu - M)^2}{2\tau^2}\right) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{(\mu - M)^2}{2\tau^2} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\mu^2\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right) - 2\mu\left(\frac{M}{\tau^2} + \frac{n\bar{x}}{\sigma^2}\right) + \text{const}\right)\right) \\
&\propto \exp\left(-\frac{1}{2\left(\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)}\left(\mu - \frac{M\sigma^2 + n\bar{x}\tau^2}{\sigma^2 + n\tau^2}\right)^2\right).
\end{aligned}
\tag{11}
$$

Finally, we get that $\mu \mid x_{1:n} \sim \mathcal{N}\left(\frac{M\sigma^2+n\bar{x}\tau^2}{\sigma^2+n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\right)$. Since the posterior distribution is a univariate Gaussian, a Gaussian variational family is sufficient to recover this posterior.

3.1.2 Results

As a toy example, we consider a setting where the variational family is a Gaussian distribution with fixed and non-fixed variance. This allows us to evaluate the performance of ADVI and DADVI under two conditions: one where the variance is fixed and equal to the true posterior variance, and one where both the mean and variance are optimized.

We generated synthetic data from the model using the following parameters:

- Dataset size: $n = 1000$
- Prior mean: $M = 70$
- Prior variance: $\tau^2 = 1$
- Likelihood variance: $\sigma^2 = 100$

The dataset was generated as follows:

```
mu = np.random.normal(M, np.sqrt(tau2), 1)
X = np.random.normal(mu, np.sqrt(sigma2), n)
```

The comparison of the true posterior with the ADVI and DADVI approximations is presented in Table 1 and Figure 1.

In the **fixed variance setting**, it takes 2072 iterations for ADVI to converge for a precision of $\varepsilon = 10^{-5}$ and a learning rate of $\tau = 10^{-5}$. Both ADVI and DADVI recover posterior means that are close to the true value, indicating satisfactory performance. However, DADVI slightly outperforms ADVI in terms of the ELBO, suggesting a more accurate approximation of the true posterior distribution under this constraint.

In the **non-fixed variance setting**, it takes 7 iterations for ADVI to converge $\varepsilon = 10^{-5}$ and a learning rate of $\tau = 10^{-5}$. Here, the differences between the two methods become more pronounced. ADVI struggles to accurately estimate the variance and appears to converge to a local optimum. This results in an overestimated variance and a flatter posterior distribution. In contrast, DADVI provides a significantly better approximation, recovering both the mean and variance with high fidelity.

Figure 1 (right) further illustrates these observations. The DADVI approximation closely matches the shape and location of the true posterior, while the ADVI approximation visibly deviates, particularly in its dispersion.

These trends are reflected quantitatively in the ELBO values: DADVI achieves a higher ELBO (-3735.508) compared to ADVI (-3741.162), confirming the superiority of DADVI in more flexible variational settings.

Table 1: Comparison of ADVI and DADVI in fixed and non-fixed variance settings.

Setting	**Method**	**Posterior**	**ELBO**
Fixed Variance ($\bar{\sigma}^2 = 0.091$)	True	$\mathcal{N}(70.455, \bar{\sigma}^2)$	—
	ADVI	$\mathcal{N}(70.648, \bar{\sigma}^2)$	-515799.586
	DADVI	$\mathcal{N}(70.499, \bar{\sigma}^2)$	-515797.286
Non-Fixed Variance	True	$\mathcal{N}(70.455, \bar{\sigma}^2)$	—
	ADVI	$\mathcal{N}(70.000,\ 0.998)$	-3741.162
	DADVI	$\mathcal{N}(70.372,\ 0.037)$	-3735.508

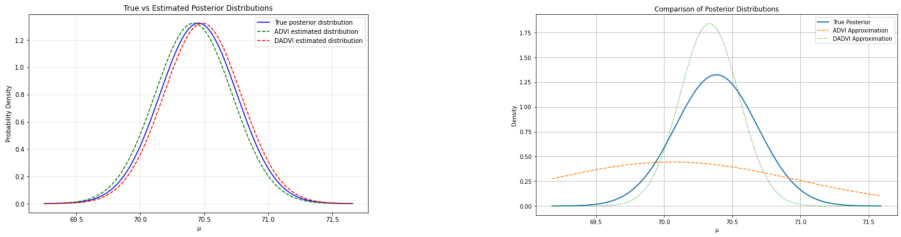

Figure 1: Left: Posterior approximations (ADVI vs. DADVI) with fixed variance. Right: Posterior approximations with both mean and variance optimized.

3.2 Three-Dimensional Gaussian variational family

3.2.1 Computation of the true posterior

In this setup, we consider a two-dimensional Gaussian model with independent likelihoods for each dimension, assuming equal variance across dimensions.

We assume that the variational parameter $\mu \sim \mathcal{N}(M, \tau^2 I_2)$, where $M \in \mathbb{R}$ is the prior mean, $\tau^2 > 0$ is the prior variance and I_2 is the identity matrix. We also assume that the likelihood is normally distributed: $x_{1:n} \mid \mu \sim \mathcal{N}(\mu, \sigma^2 I_2)$, where $\sigma^2 > 0$ is the variance and I_2 is the two-dimensional identity matrix.

According to Bayes' rule, we have

$$
\begin{aligned}
&p(\mu \mid x_{1:n}) \propto p(\mu) \cdot p(x_{1:n} \mid \mu) \\
&= \frac{1}{2\pi} \exp\left(-\frac{1}{2\tau^2} \|\mu - M\|^2\right) \cdot \frac{1}{(2\pi\sigma^2)^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \|x_i - \mu\|^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{\|\mu - M\|^2}{\tau^2} + \sum_{i=1}^{n} \frac{\|x_i - \mu\|^2}{\sigma^2}\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\|\mu\|^2 \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right) - 2\left\langle \mu, \frac{M}{\tau^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2} \right\rangle + \text{const}\right)\right) \\
&\propto \exp\left(-\frac{1}{2\left(\frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)} \left\| \mu - \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{M}{\tau^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right) \right\|^2\right)
\end{aligned}
\tag{12}
$$

Finally, we get that $\mu \mid x_{1:n} \sim \mathcal{N}\left(\left(\frac{1}{\tau^2} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{M}{\tau^2} + \frac{\sum_{i=1}^{n} x_i}{\sigma^2}\right),\ \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$.

3.2.2 Results

In this experiment, we generated synthetic data from a two-dimensional Gaussian model with independent likelihoods for each dimension, assuming the same variance across both dimensions. The parameters used to generate the data are as follows:

- **Dataset size:** $n = 1000$
- **Prior mean:** $M = [20, 30]$
- **Prior variance:** $\tau^2 = 1$
- **Likelihood variance:** $\sigma^2 = 100$

The data were generated as follows:

```
mu = np.random.multivariate_normal(M, tau2 * np.eye(2))
X = np.random.multivariate_normal(mu, sigma2 * np.eye(2), n)
```

Table 2: Posterior estimates and ELBO values for ADVI and DADVI.

Method	Posterior Mean	Posterior Variance	ELBO
True	[19.698, 28.590]	0.091	—
ADVI	[19.712, 28.578]	0.046	−102142.317
DADVI	[19.696, 28.591]	0.045	−102142.200

Table 2 compares the performance of ADVI and DADVI in approximating the posterior distribution based on a closed-form reference. Both methods yield

posterior means very close to the true values, suggesting good performance. The estimated variances are also similar, with DADVI providing a slightly more accurate match to the true posterior variance.

The results were obtained after 10000 iterations of the ADVI algorithm, using a precision threshold of $\varepsilon = 10^{-5}$ and a learning rate of $\tau = 10^{-3}$.

While the ELBO values are close, DADVI achieves a marginally better score (-102142.317 vs. -102142.200), indicating a slightly better optimization outcome and overall approximation.

3.3 Four-Dimensional Gaussian Variational family

3.3.1 Computation of the true posterior

In this experiment, we consider a two-dimensional Gaussian model with independent likelihoods for each dimension, assuming the different variances across both dimensions.

We assume that the variational parameter $\mu \sim \mathcal{N}(M, \tau^2 I_2)$, where $M \in \mathbb{R}^2$ is the prior mean, $\tau^2 > 0$ is the prior variance, and I_2 is the two-dimensional identity matrix. We further assume that the likelihood is normally distributed: $x_{1:n} \mid \mu \sim \mathcal{N}(\mu, \Sigma)$, where the covariance matrix is $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$, with $\sigma_1^2, \sigma_2^2 > 0$.

We have

$$
\begin{aligned}
p(\mu \mid x_{1:n}) &\propto p(\mu) \cdot p(x_{1:n} \mid \mu) \\
&= \exp\left(-\frac{1}{2}(\mu - M)^T \frac{1}{\tau^2} I_2 (\mu - M)\right) \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right) \\
&= \exp\left(-\frac{1}{2\tau^2}(\mu - M)^T(\mu - M) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right) \\
&\propto \exp\left(-\frac{1}{2\tau^2}\mu^T\mu + \frac{1}{\tau^2}\mu^T M - \frac{n}{2}\mu^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\sum_{i=1}^{n} x_i\right) \\
&\propto \exp\left(-\frac{1}{2}\mu^T\left(\frac{1}{\tau^2}I_2 + n\Sigma^{-1}\right)\mu + \mu^T\left(\frac{M}{\tau^2} + \Sigma^{-1}\sum_{i=1}^{n} x_i\right)\right) \\
&\propto \exp\left(-\frac{1}{2}(\mu - \bar{\mu})^T\left(\frac{1}{\tau^2}I_2 + n\Sigma^{-1}\right)(\mu - \bar{\mu})\right)
\end{aligned}
\tag{13}
$$

where $\bar{\mu} = \left(\frac{1}{\tau^2}I_2 + n\Sigma^{-1}\right)^{-1}\left(\frac{M}{\tau^2} + \Sigma^{-1}\sum_{i=1}^{n} x_i\right)$.

Finally, we get that $\mu \mid x_{1:n} \sim \mathcal{N}\left(\bar{\mu}, \left(\frac{1}{\tau^2}I_2 + n\Sigma^{-1}\right)^{-1}\right)$

3.3.2 Results

In this experiment, we consider a two-dimensional Gaussian model with independent likelihoods for each dimension, assuming different variances across both dimensions.

Data Generation: The synthetic data were generated using the following parameters:

- Dataset size: $n = 1000$
- Prior mean: $M = [20, 30]$
- Prior variance: $\tau^2 = 1$
- Covariance matrix of the likelihood: $\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 50 \end{pmatrix}$

The data were generated using the following code:

```
mu = np.random.multivariate_normal(M, tau2 * np.eye(2))
data = np.random.multivariate_normal(mu, np.array([[25, 0], [0, 50]]), n)
```

Table 3: Comparison of posterior approximations and ELBO values for ADVI and DADVI.

Method	Posterior	ELBO
True	$\mathcal{N}\left([19.815,\ 30.409], \begin{pmatrix} 0.024 & 0 \\ 0 & 0.048 \end{pmatrix}\right)$	—
ADVI	$\mathcal{N}\left([19.812,\ 30.448], \begin{pmatrix} 0.023 & 0 \\ 0 & 0.054 \end{pmatrix}\right)$	-6371.450
DADVI	$\mathcal{N}\left([19.815,\ 30.409], \begin{pmatrix} 0.024 & 0 \\ 0 & 0.048 \end{pmatrix}\right)$	-6371.444

The true posterior is a bivariate Gaussian with mean vector [19.815, 30.409] and diagonal covariance matrix with variances 0.024 and 0.048.

The results were obtained after 10000 iterations of the ADVI algorithm, which is the maximum number of iterations we set, using a precision threshold of $\varepsilon = 10^{-5}$ and a learning rate of 10^{-3}.

Both ADVI and DADVI closely approximate the true distribution. DADVI matches the posterior mean almost exactly, while ADVI shows only a small deviation in both components. Both methods recover the correct covariance structure.

The ELBO values reinforce this observation: ADVI and DADVI perform nearly identically in terms of evidence lower bound, with DADVI achieving a marginally better score (-6371.444 versus -6371.450). This suggests that while both algorithms are effective, DADVI demonstrates slightly improved precision in capturing the posterior's characteristics.

3.4 Model with mixture of Gaussians

We compute the true posterior to make comparisons for the ADVI and DADVI algorithms, but this is only possible for $n = 1$. For $n > 1$ we just consider the ELBO values from each algorithm.

3.4.1 Computation of true posterior for $n = 1$

We assume that:

- **Prior:** the variational parameter is $\mu \sim \mathcal{N}(0, \tau^2)$, where $\tau^2 \in \mathbb{R}^+$ is the prior variance.
- **Likelihood:** is normally distributed

$$X \mid \mu \sim \begin{cases} \mathcal{N}(\mu, \sigma^2) & \text{with probability } p \\ \mathcal{N}(-\mu, \sigma^2) & \text{with probability } 1-p \end{cases}$$

We can easily compute the exact posterior when dataset size $n = 1$ as shown below.

$$\begin{aligned}
p(\mu \mid X) &\propto p(\mu)\, p(X \mid \mu) \\
&= \mathcal{N}\big(\mu; 0, \tau^2\big)\Big[p\,\mathcal{N}\big(X; \mu, \sigma^2\big) + (1-p)\,\mathcal{N}\big(X; -\mu, \sigma^2\big)\Big] \\
&\propto p\,\exp\Big[-\tfrac{(X-\mu)^2}{2\sigma^2} - \tfrac{\mu^2}{2\tau^2}\Big] + (1-p)\,\exp\Big[-\tfrac{(X+\mu)^2}{2\sigma^2} - \tfrac{\mu^2}{2\tau^2}\Big] \\
&= p\,\exp\Big[-\tfrac{(X-\mu)^2\tau^2 + \mu^2\sigma^2}{2\sigma^2\tau^2}\Big] + (1-p)\,\exp\Big[-\tfrac{(X+\mu)^2\tau^2 + \mu^2\sigma^2}{2\sigma^2\tau^2}\Big] \\
&= p\,\exp\Big[-\tfrac{(\tau^2+\sigma^2)\mu^2 - 2X\tau^2\mu + X^2\tau^2}{2\sigma^2\tau^2}\Big] + (1-p)\,\exp\Big[-\tfrac{(\tau^2+\sigma^2)\mu^2 + 2X\tau^2\mu + X^2\tau^2}{2\sigma^2\tau^2}\Big] \\
&= p\,\exp\Big[-\tfrac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\Big(\mu - \tfrac{\tau^2}{\tau^2+\sigma^2}X\Big)^2\Big] + (1-p)\,\exp\Big[-\tfrac{\tau^2+\sigma^2}{2\sigma^2\tau^2}\Big(\mu + \tfrac{\tau^2}{\tau^2+\sigma^2}X\Big)^2\Big].
\end{aligned} \tag{14}$$

Hence the posterior is a two-component Gaussian mixture of the form

$$p(\mu \mid X) = p\,\mathcal{N}\Big(\mu\,;\ \tfrac{\tau^2}{\tau^2+\sigma^2}X,\ \tfrac{\sigma^2\tau^2}{\sigma^2+\tau^2}\Big) + (1-p)\,\mathcal{N}\Big(\mu\,;\ -\tfrac{\tau^2}{\tau^2+\sigma^2}X,\ \tfrac{\sigma^2\tau^2}{\sigma^2+\tau^2}\Big).$$

Although an exact posterior distribution can be written down for any sample size n, it is a 2^n-component Gaussian mixture, making it challenging to compute and visualize for large and moderate n.

3.4.2 Results

In this experiment, we generated synthetic data from a mixture of Gaussian models with symmetric means and identical variance.

The parameters used to generate the data were:

- Dataset size: $n = 1$ and $n = 1000$ (two cases)
- Mixture probability: $p = 0.5$
- Prior variance: $\tau^2 = 100$
- Likelihood variance: $\sigma^2 = 1$

The data were generated using the following procedure. First, a latent mean μ was sampled from the prior:

$$\mu \sim \mathcal{N}(0, \tau^2)$$

Then, a synthetic dataset was created using the Python-like function:

```
def sim(n, mu, sigma, p):
    X = np.random.normal(mu, sigma, n)
    Y = np.random.normal(-mu, sigma, n)
    Ber = np.random.binomial(1, p, n)
    Z = X * Ber + Y * (1 - Ber)
    return Z
```

This procedure generates data from a symmetric mixture model where each data point is sampled from $\mathcal{N}(\mu, \sigma^2)$ with probability p, and from $\mathcal{N}(-\mu, \sigma^2)$ with probability $1 - p$.

For $n = 1$, the exact posterior is analytically tractable. We used different random seeds to manipulate the sampled μ and produce posterior shapes with two bumps that were:

- **Far apart**
- **Close to each other**
- **Overlapping**

We then approximated the posterior distributions using both ADVI and DADVI methods and computed their respective ELBO values.

The outcomes, including the true posterior (when tractable), the ADVI and DADVI approximations, and the associated ELBO values, are compiled in a summary table for two cases ($n = 1$ and $n = 1000$) and for the three posterior configurations (far, close, overlapping). See Table 4 for a detailed comparison. For a visualization of the approximations posterior distributions of μ under different data configurations, see Figures 4 and 5 in the Appendix.

Table 4: Comparison of true and approximate posteriors under different configurations.

	Two Bumps Far n = 1	**Two Bumps Close n =1**	**Two Bumps Overlapping n = 1**	$n = 1000$
True Posterior	$\frac{1}{2}\mathcal{N}(11.003, 0.990) + \frac{1}{2}\mathcal{N}(-11.003, 0.990)$	$\frac{1}{2}\mathcal{N}(4.277, 0.990) + \frac{1}{2}\mathcal{N}(-4.277, 0.990)$	$\frac{1}{2}\mathcal{N}(1.904, 0.990) + \frac{1}{2}\mathcal{N}(-1.904, 0.990)$	Simulated (intractable)
DADVI Posterior	$\mathcal{N}(-1.163, 94.827)$	$\mathcal{N}(-1.516, 12.182)$	$\mathcal{N}(0.825, 3.174)$	$\mathcal{N}(-0.120, 9.242)$
ADVI Posterior	$\mathcal{N}(-0.007, 45.088)$	$\mathcal{N}(-0.013, 66.822)$	$\mathcal{N}(-0.008, 91.099)$	$\mathcal{N}(0.188, 5.088)$
ELBO (DADVI)	-18.936	-6.729	-4.921	-2861.938
ELBO (ADVI)	-23.871	-16.205	-40.329	-2874.620

The performance of ADVI and DADVI was evaluated across different posterior configurations generated by varying the separation between two modes in a synthetic Gaussian mixture model. When the two bumps of the true posterior were far apart (centered at ± 11), both ADVI and DADVI failed to capture the bimodal structure, returning single Gaussian approximations centered near zero. However, DADVI produced a slightly better approximation in terms of the ELBO, with a broader variance that better encompassed the two peaks, whereas ADVI's approximation remained narrower and less representative of the posterior's true shape.

As the two modes moved closer together (centered at ± 4.277), DADVI continued to outperform ADVI. Its approximation adjusted to the closer bimodal form with a narrower variance, whereas ADVI became overly diffuse, assigning excessive uncertainty. This is reflected in the ELBO values, where DADVI again achieved a substantially better score, indicating a more faithful posterior approximation.

In the case where the two bumps were overlapping (centered at ± 1.904), the posterior became nearly unimodal. Here, DADVI adapted effectively, producing a compact Gaussian centered near the true posterior's central mass. In contrast, ADVI failed to adjust and returned a variance of over 90, resulting in a near-uniform distribution and an extremely poor ELBO score. This stark contrast further highlights DADVI's superior capacity to adapt to posterior shape and scale.

Finally, for larger sample size ($n = 1000$), where the true posterior becomes harder to compute exactly but tends toward a unimodal form, DADVI performs slightly better than ADVI based on their ELBO values.

In summary, DADVI consistently outperforms ADVI across a variety of posterior shapes, particularly when the posterior is non-Gaussian or bimodal. It provides closer approximations to the true distribution and achieves higher ELBO scores, making it a more reliable choice in complex inference tasks.

3.5 Application of DADVI and ADVI to real medical dataset

We now apply ADVI and DADVI to a real-world scenario. Assume we have J independent clinical trials, each with a control arm (index 0) and a treatment arm (index 1). We assume that individuals in each clinical trial are drawn from the same population.

Notation (trial j):

$n_{0,j}$: number of participants in the control arm
$n_{1,j}$: number of participants in the treatment arm
$y_{0,j}$: number of deaths in the control arm
$y_{1,j}$: number of deaths in the treatment arm

We first define the model:

$$y_{0,j} \mid p_0 \sim \mathrm{Bin}(n_{0,j}, p_0), \quad y_{1,j} \mid p_1 \sim \mathrm{Bin}(n_{1,j}, p_1), \quad j = 1, \ldots, J \tag{15}$$

Here p_0 is the probability of death in the control group, and p_1 in the treatment group. We assume independent uniform priors (or Beta(1, 1)) for the probabilities:

$$p_0 \sim \mathcal{U}(0,1), \quad p_1 \sim \mathcal{U}(0,1). \tag{16}$$

Hence the joint prior density is

$$f_{p_0,p_1}(p_0, p_1) = f_{p_0}(p_0) f_{p_1}(p_1) \tag{17}$$

and the likelihood is

$$L(y \mid p_0, p_1) = \prod_{j=1}^{J} \binom{n_{0,j}}{y_{0,j}} p_0^{y_{0,j}} (1-p_0)^{n_{0,j}-y_{0,j}} \binom{n_{1,j}}{y_{1,j}} p_1^{y_{1,j}} (1-p_1)^{n_{1,j}-y_{1,j}} \tag{18}$$

Because the priors are independent and and the likelihood factorizes by arm, the posterior distributions of p_0, p_1 remain independent. Their exact posteriors are given by:

$$\begin{aligned} f(p_i \mid y) &= f(y \mid p_i)\, f(p_i) \\ &= \prod_{j=1}^{J} \left[p_i^{y_{i,j}} \, (1-p_i)^{\, n_{i,j}-y_{i,j}} \right] f(p_i) \\ &\propto p_i^{\sum_{j=1}^{J} y_{i,j}} \, (1-p_i)^{\sum_{j=1}^{J} (n_{i,j}-y_{i,j})} \\ &= p_i^{\left(\sum_{j=1}^{J} y_{i,j}+1\right)-1} \, (1-p_i)^{\left(\sum_{j=1}^{J} n_{i,j}-\sum_{j=1}^{J} y_{i,j}+1\right)-1} \end{aligned} \tag{19}$$

where $N = \sum_{j=1}^{J} n_j$ and $i \in \{0, 1\}$. Finally, our true posteriors are:

$$p_0 \mid y \sim \text{Beta}\left(\sum_{j=1}^{J} y_{0,j} + 1, \sum_{j=1}^{J}(n_{0,j} - y_{0,j}) + 1\right) \tag{20}$$

$$p_1 \mid y \sim \text{Beta}\left(\sum_{j=1}^{J} y_{1,j} + 1, \sum_{j=1}^{J}(n_{1,j} - y_{1,j}) + 1\right) \tag{21}$$

Since we wish to approximate the joint posterior with a bivariate normal, we map $(p_0, p_1) \in (0, 1)^2$ onto $\mathbb{R}^2$ via the logit:

$$\varphi : (0,1)^2 \to \mathbb{R}^2, \quad \varphi(p_0, p_1) = (\theta_0, \theta_1) = \left(\log \frac{p_0}{1 - p_0}, \log \frac{p_1}{1 - p_1}\right), \tag{22}$$

which has Jacobian determinant:

$$|J(\theta_0, \theta_1)| = \frac{e^{\theta_0}}{(1 + e^{\theta_0})^2} \cdot \frac{e^{\theta_1}}{(1 + e^{\theta_1})^2} \tag{23}$$

Hence the transformed density is:

$$f_{\theta_0,\theta_1}(\theta_0, \theta_1) = f_{p_0}\left(\frac{e^{\theta_0}}{1 + e^{\theta_0}}\right) f_{p_1}\left(\frac{e^{\theta_1}}{1 + e^{\theta_1}}\right) \cdot |J(\theta_0, \theta_1)| = |J(\theta_0, \theta_1)| \tag{24}$$

The ELBO for ADVI/DADVI is built from this transformed prior together with the log-likelihood expressed in (θ_0, θ_1) as follows.

$$\begin{aligned} L(y \mid \theta_0, \theta_1) = \prod_{j=1}^{J} &\binom{n_{0,j}}{y_{0,j}} \left(\tfrac{e^{\theta_0}}{1+e^{\theta_0}}\right)^{y_{0,j}} \left(\tfrac{1}{1+e^{\theta_0}}\right)^{n_{0,j}-y_{0,j}} \\ &\times \binom{n_{1,j}}{y_{1,j}} \left(\tfrac{e^{\theta_1}}{1+e^{\theta_1}}\right)^{y_{1,j}} \left(\tfrac{1}{1+e^{\theta_1}}\right)^{n_{1,j}-y_{1,j}}. \end{aligned} \tag{25}$$

3.5.1 Results

This analysis uses data from 22 clinical trials of beta-blockers to reduce mortality after myocardial infarction (Yusuf et al. 1985). In this example, our variational approximation assumes a mean-field Gaussian family.

The posterior distributions of the parameters p_0 and p_1 were approximated using both ADVI and DADVI. ADVI converged after 4,838 iterations, which was matched by using the same Monte Carlo samples in DADVI. It is noted that, according to the graphs, the posterior distributions obtained by DADVI

are already close to the true posteriors when using as few as 20 Monte Carlo samples.

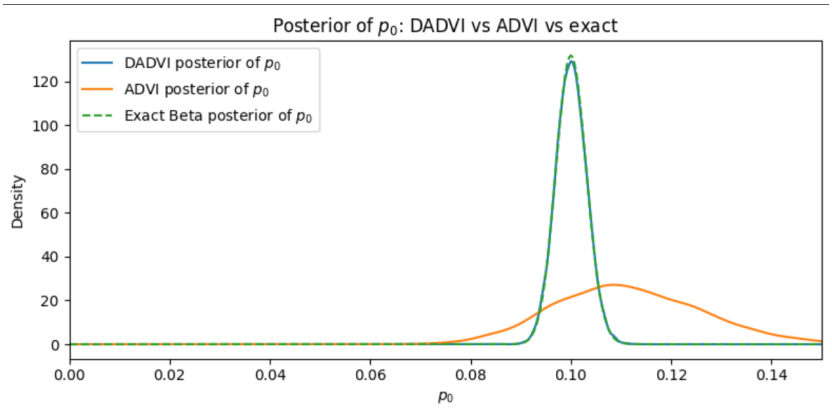

Figure 2: Posterior approximation for p_0 using ADVI and DADVI compared to the exact posterior.

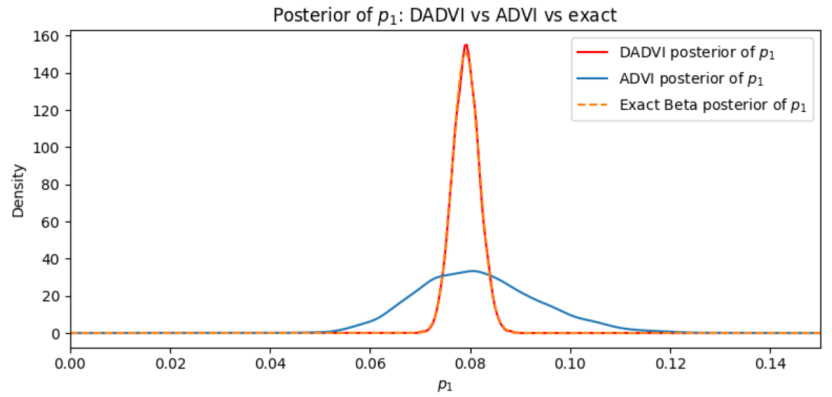

Figure 3: Posterior approximation for p_1 using ADVI and DADVI compared to the exact posterior.

To visualize these results, Figures 2 and 3 present comparisons of the approximate and exact posterior distributions for p_0 and p_1. The DADVI approximations are notably well-aligned with the exact posteriors in terms of both shape and central tendency, while ADVI yields much flatter, diffuse distribution.

Table 5: Comparison of posterior estimates for p_0 and p_1 in $\mathcal{N}(\text{mean},\ \text{std}^2)$ form, with ELBO values.

Method	$\boldsymbol{p_0} \sim \mathcal{N}(\cdot)$	$\boldsymbol{p_1} \sim \mathcal{N}(\cdot)$	**ELBO**
MLE	0.100	0.079	—
ADVI	$\mathcal{N}(0.112,\ 0.015^2)$	$\mathcal{N}(0.082,\ 0.012^2)$	-286.706
DADVI	$\mathcal{N}(0.100,\ 0.003^2)$	$\mathcal{N}(0.079,\ 0.003^2)$	-274.347

Table 5 presents a comparison of the posterior estimates for parameters p_0 and p_1 obtained via ADVI and DADVI, alongside the maximum likelihood estimates (MLEs) and corresponding ELBO values.

The ELBO values further support these observations: DADVI achieves a higher ELBO, which implies a tighter and more accurate approximation of the true posterior. Overall, DADVI demonstrates superior performance in both parameter recovery and variational optimization in this setting.

4 Conclusion

Taken together, these results indicate that DADVI is never inferior to ADVI and can sometimes substantially outperform it. In all scenarios, DADVI achieves a higher ELBO.

According to Table 6, both ADVI and DADVI consistently produce mean estimates that closely match the true values, regardless of increasing dimensionality or unknown variance parameters. This consistency in point estimation suggests that neither method exhibits systematic bias in identifying the high-probability region of the posterior.

However, DADVI provides better estimates of posterior variances. We observe that in the two-dimensional variational Gaussian with unfixed variance and in the one-dimensional mixture with two bumps and in the meta-analysis, ADVI's Gaussian approximation becomes noticeably "too flat" compared to both the DADVI approximation and the true posterior.

We conclude that DADVI's advantage becomes more pronounced as the complexity of the observed data increases.

Scenario	Key Feature Tested	Outcome
1-D Gaussian, fixed σ^2	Pure location inference	Both methods recover the mean; DADVI yields a slightly higher ELBO.
2-D Gaussian, unfixed σ^2	Must learn mean & variance	ADVI over-disperses; DADVI recovers close mean and relatively close variance with better ELBO.
3-D Gaussian	High dimension	Both methods recover the mean not the variance; DADVI yields a slightly higher ELBO.
4-D Gaussian	Different variances per axis	DADVI matches perfectly with the true posterior while ADVI also gives a good estimate.
1-D Gaussian mixture	Multi-modality	DADVI obtains better ELBO values than ADVI in all cases.
β-blocker meta-analysis	Real binomial data	DADVI matches exact Beta-posterior logits in mean & scale; ADVI is more diffuse.

Table 6: Condensed comparison of ADVI and DADVI across experimental settings.

5 Limitations and Future work

5.1 Limitations

The comparisons we explore are encouraging, but far from exhaustive. First, every synthetic example is low–dimensional. The largest model has just four unknowns, so it remains unclear how DADVI behaves in set ups with much more unknowns. Second, we confined both algorithms to diagonal-Gaussian (mean-field) variational families. A richer family might reduce or even erase the gap we have observed. Third, we chose not to report the running time or algorithmic complexity for a given level of accuracy. Although these numbers are easy to log, comparing them fairly would be challenging, as it is difficult to isolate the contributions of the intrinsic properties of ADVI and DADVI from external factors. Fourth, the real-data test bed is deliberately simple: a binomial meta-analysis with two logits. Other data types—time-series, spatial fields, deep-learning problems—could reveal different strengths or weaknesses. Finally, our evidence is purely empirical. We have not attempted formal comparisons, nor have we probed how sensitive each method is to hyperparameters such as learning rate.

5.2 Future work

Several directions follow naturally. First, we could test each method's sensitivity to Monte Carlo noise by running the ELBO estimation with different numbers of samples and comparing the results. Then, to see if DADVI's advantage scales up, we should repeat our experiments on larger models—like Bayesian neural networks or state-space models—using mini-batch stochastic training.

It would also be worthwhile to couple DADVI with more expressive variational families such as mixture distributions, so that multi-modal posteriors can be captured directly rather than through a unimodel Gaussian.

Beyond the ELBO, future evaluations could include posterior-predictive checks or cross-validations with log-likelihoods, which often catch problems that the ELBO can not diagnose.

6 Appendix: Visualization of the Approximation of the Mixture Posterior Distributions of μ when n = 1

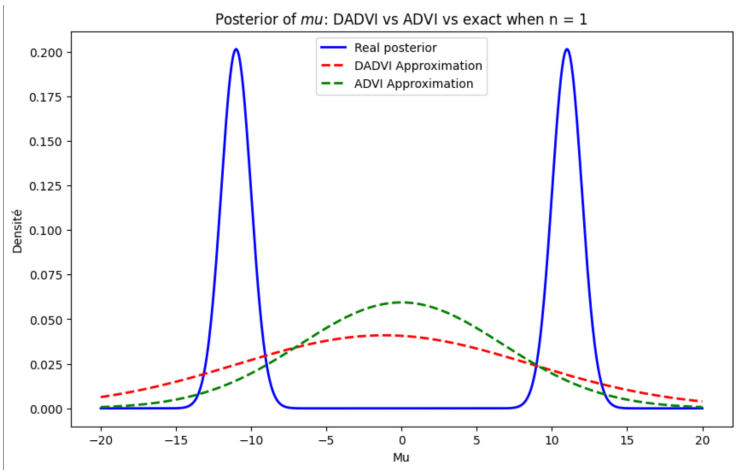

Figure 4: Posterior of μ — well-separated bumps, $n = 1$

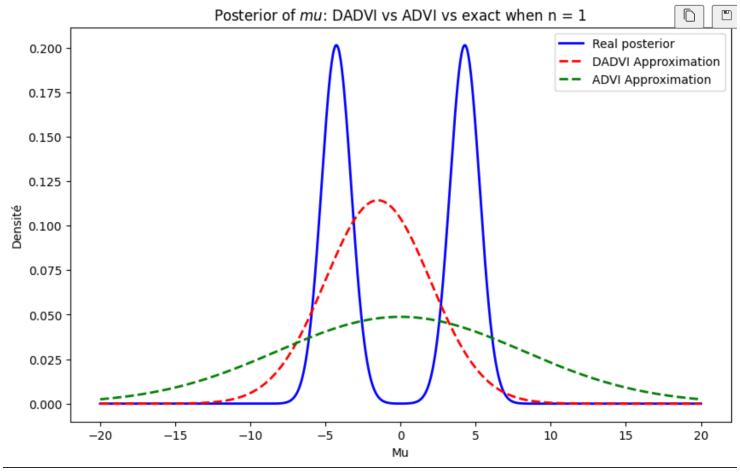

Figure 5: Posterior of μ — closely spaced bumps, $n = 1$

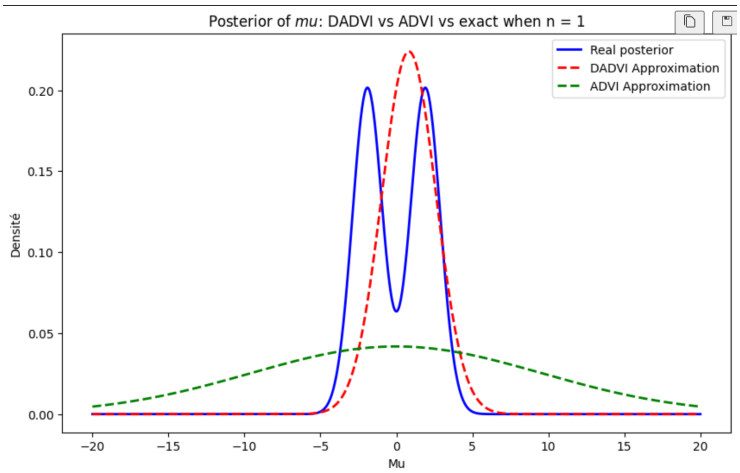

Figure 6: Posterior of μ — overlapping bumps, $n = 1$

References

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

Brooks, S. P. & Roberts, G. O. (1998), 'Convergence assessment techniques for markov chain monte carlo', *Statistics and Computing* **8**, 319–335.

Giordano, R., Ingram, M. & Broderick, T. (2024), 'Black box variational inference with a deterministic objective: Faster, more accurate, and even more black box', *Journal of Machine Learning Research* **25**(18), 1–39.

Jacobs, R. (2008), 'Bayesian statistics: Normal-normal model', *Department of Brain & Cognitive Sciences University of Rochester* .

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. (2017), 'Automatic differentiation variational inference', *Journal of Machine Learning Research* **18**(14), 1–45.

Ranganath, R., Gerrish, S. & Blei, D. M. (2014), Black box variational inference, *in* 'Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)', Vol. 33, JMLR Workshop and Conference Proceedings, pp. 814–822.

Xu, M., Quiroz, M., Kohn, R. & Sisson, S. A. (2019), Variance reduction properties of the reparameterization trick, *in* 'The 22nd international conference on artificial intelligence and statistics', PMLR, pp. 2711–2720.