

# Generalized Guarantees for Variational Inference in the Presence of Even and Elliptical Symmetry

Charles C. Margossian  
Department of Statistics  
University of British Columbia

Lawrence K. Saul  
Center for Computational Mathematics  
Flatiron Institute

## Abstract

We extend several recent results providing symmetry-based guarantees for variational inference (VI) with location-scale families. VI approximates a target density  $p$  by the best match  $q^*$  in a family  $\mathcal{Q}$  of tractable distributions that in general does not contain  $p$ . It is known that VI can recover key properties of  $p$ , such as its mean and correlation matrix, when  $p$  and  $\mathcal{Q}$  exhibit certain symmetries and  $q^*$  is found by minimizing the reverse Kullback-Leibler divergence. We extend these guarantees in two important directions. First, we provide symmetry-based guarantees for a broader family of divergences, highlighting the properties of variational objectives under which VI provably recovers the mean and correlation matrix. Second, we obtain further guarantees for VI when the target density  $p$  exhibits even and elliptical symmetries in *some but not all* of its coordinates. These partial symmetries arise naturally in Bayesian hierarchical models, where the prior induces a challenging geometry but still possesses axes of symmetry. We illustrate these theoretical results in a number of experimental settings.

this family the best approximation to a target distribution  $p$ . VI is typically presented as a scalable alternative to more classical algorithms, such as Markov chain Monte Carlo (MCMC, Robert and Casella, 2004). While VI can quickly obtain an approximation within a constrained family  $\mathcal{Q}$ , it is often unclear how well this solution approximates  $p$  (e.g., Yao et al., 2018; Gior-dano et al., 2018; Talts et al., 2018; Huggins et al., 2020; Dhaka et al., 2021). For this reason it is useful to understand the conditions under which VI returns approximations that are *provably* accurate. This paper contributes to a growing body of work on this subject (e.g., Wang and Blei, 2018; Katsevich and Rigollet, 2024; Margossian and Saul, 2025).

Most approaches to VI seek an approximation in  $\mathcal{Q}$  by minimizing the reverse Kullback-Leibler (KL) divergence. But VI can be formulated with other divergences, some of which possess attractive properties, such as the forward KL-divergence (Naesseth et al., 2020; Vehtari et al., 2020) and the Rényi divergences (Li and Turner, 2016; Dieng et al., 2017; Daudel et al., 2023). It is known that different divergences, when minimized, can return different solutions, even when the optimization is carried over the same family  $\mathcal{Q}$ . For this reason, it is of interest to understand which positive guarantees for VI hold across a *range* of divergences and are not sensitive to a particular choice of objective function. This paper considers this question for a family of divergences closely related to  $f$ -divergences (Rényi, 1961).

Our work extends recent guarantees when  $\mathcal{Q}$  is a family of location-scale distributions and the best variational approximation is found by minimizing the reverse KL divergence. Here it is known, under certain conditions, that VI recovers both the mean and correlation matrix of  $p$  whenever  $p$  is, respectively, even and elliptically symmetric (Margossian and Saul, 2025). We generalize these results in two directions. First, we obtain similar guarantees for VI with a broader class of divergences. Second, we show that if  $p$  exhibits symmetries in *some but not all* of its coordinates, then VI recovers

## 1 INTRODUCTION

Variational inference (VI) is a popular methodology for Bayesian inference and probabilistic machine learning (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017). The modus operandi of VI is to posit a family of tractable distributions  $\mathcal{Q}$  and find within

the partial mean and correlations along these coordinates. Of special note is that such partial symmetries arise in models with hierarchical priors.

**Related work.** Our work most closely builds on recent guarantees for VI in the presence of symmetries, despite misspecifications in the family  $\mathcal{Q}$  (Margossian and Saul, 2025). Earlier studies have also demonstrated VI’s ability to recover the mean, both empirically (e.g., MacKay, 2003; Giordano et al., 2018) and theoretically, in pre-asymptotic regimes (e.g., Katsevich and Rigollet, 2024), and others have obtained positive results when VI is used to maximize a marginal likelihood (Jordan et al., 1999; Li and Turner, 2016) or construct frequentist estimators (Wang and Blei, 2018; Alquier and Ridgway, 2020; Yang et al., 2020; Zhang and Gao, 2020). On the other hand, many studies have proven negative results for VI, particularly when it is used to quantify uncertainty in  $p$  (MacKay, 2003; Turner and Sahani, 2011; Giordano et al., 2018; Margossian and Saul, 2023; Margossian et al., 2025). A complementary line of work examines post-hoc diagnostics to assess the quality of VI—for example using importance sampling (Yao et al., 2018; Vehtari et al., 2024) or error bounds based on the Wasserstein distance (Huggins et al., 2020; Biswas and Mackey, 2023).

Previous work has also explored alternative objective functions for VI. Several studies have shown that when different divergences are minimized, the approximations from VI yield different quantifications of uncertainty and different estimators of the marginal likelihood (Li and Turner, 2016; Daudel et al., 2023; Margossian et al., 2025). Our results provide a counterpoint of sorts: we show that in the presence of certain symmetries, it may not matter which divergence is minimized, and that variational approximations from different divergences may all recover the mean and the correlation matrix of  $p$ , even when other properties of  $p$  are poorly estimated.

Finally, our study of partial symmetries relates to a large literature on the geometry of posteriors in hierarchical models and the interplay of this geometry with inference algorithms (e.g. Neal, 2001; Papaspiliopoulos et al., 2007; Betancourt and Girolami, 2015).

## 2 PRELIMINARIES

In this section, we provide formal definitions and identify the assumptions behind our theoretical analysis.

### 2.1 Objective functions for VI

VI minimizes a divergence between a target  $p(z)$  and an approximation  $q(z)$  over a family  $\mathcal{Q}$  of parameter-

ized distributions. We focus on the continuous case, with  $z \in \mathbb{R}^d$ , and assume  $p(z)$  and  $q(z)$  both admit a density with respect to a Lebesgue measure.

There are many choices of divergences which can, at least in theory, be optimized for VI. The most common choice is the reverse KL divergence,

$$\text{KL}(q(z)||p(z)) = \int (\log q(z) - \log p(z))q(z)dz. \quad (1)$$

In many applications, it is only possible to evaluate an unnormalized target density  $\tilde{p}$ , however, substituting  $p$  with  $\tilde{p}$  in the above equation does not change the underlying optimization problem. When the integral in eq. (1) is intractable, it can be approximated via Monte Carlo using draws from  $q$ .  $\text{KL}(q(z)||p(z))$  is then minimized by stochastic optimization.

There exist several alternatives to the reverse KL-divergence. Certain algorithms approximately minimize the forward KL-divergence (Naesseth et al., 2020; Vehtari et al., 2020). A generalization of the KL-divergences is provided by the  $\alpha$ -divergence (Li and Turner, 2016; Dieng et al., 2017; Daudel et al., 2023), which interpolates between the reverse and forward KL-divergences, respectively, in the limits  $\alpha \rightarrow 0$  and  $\alpha \rightarrow 1$  when the  $\alpha$ -divergence is defined as in Cichocki and Amari (2010). It is well known that different divergences yield different variational approximations when the family  $\mathcal{Q}$  is restricted and  $p \notin \mathcal{Q}$ .

Other measures of discrepancy between distributions include the Hellinger distance from information theory and the total variation distance, which is used to study MCMC (e.g. Roberts and Rosenthal, 2004). In practice, these objective functions are not used for VI because they are too difficult to compute in high dimensions. Even for idealized objective functions, however, it remains of theoretical interest to understand when they yield an approximation  $q$  that recovers statistical properties of  $p$ .

We consider a broad class of divergences that includes all the divergences (and distances) described above.

**Definition 1.** ( $\varphi$ -divergence) We refer to a divergence as a  $\varphi$ -divergence if it can be written as

$$D_\varphi(p||q) = \int \varphi \left( \log \frac{p(z)}{q(z)} \right) q(z) dz \quad (2)$$

where (i)  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ , (ii)  $\varphi(0) = 0$ , (iii)  $\varphi$  is convex, and (iv)  $\varphi$  is differentiable.

Table 1 provides examples of common divergences that satisfy this definition. The family of  $\varphi$ -divergences is closely related to the family of  $f$ -divergences (Rényi, 1961). We work with  $\varphi$ -divergences in this paper because they include the  $\alpha$ -divergences  $D_\alpha(p||q)$  for

Divergence	Notation	$\varphi(t)$	Decreasing	Linear
(Reverse) Kullback-Leibler	$\text{KL}(q  p)$	$-t$	✓	✓
Rényi of order $\alpha \in \mathbb{R}^+ \setminus \{0\}$	$D_\alpha(p  q)$	$\frac{e^{\alpha t} - 1}{\alpha(\alpha - 1)}$ (if $\alpha \in (0, 1)$ )	✓	✗
(Forward) Kullback-Leibler	$\text{KL}(p  q)$	$e^t t$	✗	✗
Squared Hellinger distance	$H(p  q)$	$\left(e^{\frac{1}{2}t} - 1\right)^2$	✗	✗
Total variation distance	$\text{TV}(p, q)$	$ e^t - 1 $	✗	✗

Table 1: *Examples of common  $\varphi$ -divergences. Our theoretical results require that  $\varphi$  is convex and differentiable. We obtain further guarantees when  $\varphi$  is decreasing, and even further ones when  $\varphi$  is linear.*

$\alpha \in (0, 1)$  and because they lead to clearer proofs in the setting where  $p$  is log concave. We examine the connection between  $\varphi$ -divergences and  $f$ -divergences further in Appendix A.

Our theoretical analysis will leverage certain key properties of  $\varphi$ -divergences. We require that the function  $\varphi$  is differentiable, an assumption that is tacitly made when practitioners use stochastic optimization in VI. We obtain our strongest results when  $\varphi$  is convex and, additionally, when it is monotone decreasing. When these properties hold, we are able to identify settings where the  $\varphi$ -divergence has a unique global minimizer with respect to the variational parameters of  $q$ . Finally, for one of our results on partial symmetries, we additionally assume that  $\varphi$  is linear; of all the divergences in Table 1, only the reverse KL divergence has this property. Table 1 summarizes, for each divergence, which of these additional regularity conditions are satisfied.

## 2.2 Even, elliptical, and partial symmetries

We focus on VI in settings where the target  $p$  and the approximation  $q$  exhibit certain symmetries.

**Definition 2.** (*Even/odd symmetry.*) We say a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is even (odd) symmetric about a point  $\nu \in \mathbb{R}^d$  if, for all  $\zeta \in \mathbb{R}^d$ , it satisfies

$$(\text{even}) \quad f(\nu + \zeta) = +f(\nu - \zeta), \quad (3)$$

$$(\text{odd}) \quad f(\nu + \zeta) = -f(\nu - \zeta). \quad (4)$$

**Remark 3.** If  $p(z)$  has a finite first moment and is even symmetric about  $\nu \in \mathbb{R}^d$ , then  $\mathbb{E}_p(z) = \nu$ .

Next we define elliptical symmetry. As shorthand, we use  $\|z - \nu\|_{M^{-1}}$  to denote the Mahalanobis distance  $\sqrt{(z - \nu)^T M^{-1} (z - \nu)}$ .

**Definition 4.** (*Elliptical symmetry.*) We say that  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is elliptically symmetric about  $\nu \in \mathbb{R}^d$  if there exists a positive-definite matrix  $M \in \mathbb{R}^{d \times d}$ , with

$\text{trace}(M) = d$ , such that for any pair  $z, z' \in \mathbb{R}^d$ , we have

$$f(z) = f(z') \text{ whenever } \|z - \nu\|_{M^{-1}} = \|z' - \nu\|_{M^{-1}}. \quad (5)$$

**Remark 5.** If  $p(z)$  is an elliptically symmetric density, then the matrix  $M$  in Definition 4 is proportional to its covariance matrix, and  $p$  has correlation matrix  $\text{Corr}_p[z_i, z_j] = M_{ij} / \sqrt{M_{ii} M_{jj}}$ . In this case we will refer to  $M$  as the normalized covariance matrix of  $p$ .

In some cases,  $p$  may only be symmetric along some set of coordinates  $\sigma$ . We formalize this notion below.

**Definition 6.** (*Symmetry along  $\sigma$* ) Consider a distribution  $p(z_\sigma, z_{\bar{\sigma}})$ . We say  $p$  is even symmetric along  $\sigma$  if for each  $z_{\bar{\sigma}}$ ,  $p(z_\sigma | z_{\bar{\sigma}})$  is even symmetric about some point  $m_\sigma(z_{\bar{\sigma}})$ . We also say  $p$  is elliptically symmetric along  $\sigma$  if, for each  $z_{\bar{\sigma}}$ ,  $p(z_\sigma | z_{\bar{\sigma}})$  is elliptically symmetric with some normalized covariance matrix  $M_\sigma(z_{\bar{\sigma}})$ .

We provide an illustrative example of partial symmetry. The elliptical funnel is the distribution over  $\tau \in \mathbb{R}$  and  $\theta \in \mathbb{R}^n$  generated by

$$\tau \sim \mathcal{N}(0, 1), \quad \theta \sim \mathcal{N}(0, e^{2\tau} C), \quad (6)$$

where  $C$  is a correlation matrix. This is an extension of the well-known funnel (Neal, 2001), in which  $C$  is diagonal. The geometry of the funnel is typical of hierarchical priors and prone to frustrate many inference algorithms. But the funnel also has certain partial symmetries: the conditional distribution  $p(\theta | \tau)$  is elliptically symmetric with a point of even symmetry and a normalized covariance matrix (proportional to  $C$ ) that do not depend on the variable  $\tau$ . Later we will show that VI provably recovers the mean and the correlations along  $\theta$  when  $\mathcal{Q}$  is a family of elliptical distributions—even when VI misestimates the mean for  $\tau$  (Figure 1).

## 2.3 Location-scale families

To match the symmetries of  $p$ , we need a family  $\mathcal{Q}$  whose distributions exhibit the same symmetries. One

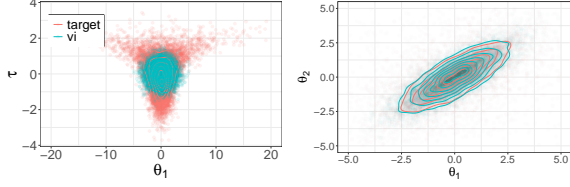


Figure 1: VI Gaussian approximation of an elliptical funnel, obtained by minimizing  $KL(q||p)$ . The funnel is asymmetric along  $\tau$  but symmetric along  $\theta$  and so VI provably recovers the mean and correlations of  $\theta$ .

natural choice for  $\mathcal{Q}$  is a location-scale family, which we define now in terms of the natural square-root  $S^{1/2}$  of a positive definite matrix  $S$ .

**Definition 7.** Let  $q_0$  be a density over  $\mathbb{R}^d$ . A location-scale family  $\mathcal{Q}$  is a two-parameter family  $\{q_{\nu,S}\}$  of densities over  $\mathbb{R}^d$  satisfying

$$q_{\nu,S}(z) = q_0(S^{-\frac{1}{2}}(z-\nu))|S|^{\frac{1}{2}}, \quad (7)$$

for all  $z, \nu \in \mathbb{R}^d$  and positive-definite matrices  $S \in \mathbb{R}^{d \times d}$ . We say that  $q_0$  is the base density of  $\mathcal{Q}$  and that  $\nu$  and  $S$  are its location and scale parameters.

**Definition 8.** A location family  $\mathcal{Q}$  is a one-parameter subfamily  $q_\nu$  of a location-scale family whose densities share the same scale.

Location-scale families are prominent in statistics and probabilistic modeling. Examples include the Gaussian, Laplace, Student-t, and Cauchy distributions. They are also popular choices for VI (e.g., Ranganath et al., 2014; Kingma and Welling, 2014; Kucukelbir et al., 2017; Cai et al., 2024b), where  $\mathcal{Q}$  is often taken to be the family of Gaussians. Throughout the paper, we assume that the base distribution  $q_0$  is spherically symmetric about the origin (and therefore also even symmetric about the origin).

## 2.4 Regularity conditions on $p$ and $q$

Our theoretical analysis is aided by imposing certain regularity conditions on  $p$  and  $q$ . First, for all of our results, we require that  $p$  is differentiable, and more generally, that for each  $q \in \mathcal{Q}$  we can interchange the order of differentiation and integration when differentiating  $D_\varphi(p||q)$ . The formal conditions for this assumption are provided by the dominated convergence theorem (e.g., Billingsley, 1995). This assumption is needed to minimize  $D_\varphi(p||q)$  via stochastic optimization, and in particular to compute Monte Carlo estimates of gradients with respect to the variational parameters.

Second, for some of our stronger results, we additionally assume that the density  $p(z)$  or  $p(z_\sigma|z_{\bar{\sigma}})$  is *somewhere-strictly log concave*, as defined below.

**Definition 9.** We say  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is *somewhere-strictly log concave* if  $f$  is log concave on all of  $\mathbb{R}^d$  and strictly log concave on some open set of  $\mathbb{R}^d$ .

This assumption is needed to ensure that  $D_\varphi(p||q)$  is strictly convex with respect to the variational parameters. With this assumption, we can relate stationary points of  $D_\varphi(p||q)$  to unique minimizers of  $D_\varphi(p||q)$ .

## 3 THEORETICAL GUARANTEES

In this section we provide generalized guarantees for VI that follow from different types of symmetry.

### 3.1 Guarantees from even symmetry

First we generalize earlier guarantees for VI with even symmetries (Margossian and Saul, 2025) to the broader family of  $\varphi$ -divergences.

**Theorem 10** (Exact Recovery of the Mean).

Let  $\mathcal{Q}$  be a location family and let  $D_\varphi$  be a  $\varphi$ -divergence. If  $p$  is even symmetric about  $\mu$ , then a stationary point of  $D_\varphi(p||q_\nu)$  occurs at  $\nu = \mu$ . Furthermore, if  $\varphi$  is strictly decreasing, and  $p$  somewhere-strictly log concave over  $\mathbb{R}^d$ , then  $\nu = \mu$  is a unique minimizer of  $D_\varphi(p||q_\nu)$ .

*Proof.* Let  $\zeta = z - \nu$ . By definition,  $q_\nu(z) = q_0(\zeta)$ , so that we can write

$$D_\varphi(p||q_\nu) = \int \varphi \left( \log \frac{p(\nu + \zeta)}{q_0(\zeta)} \right) q_0(\zeta) d\zeta. \quad (8)$$

We now differentiate with respect to  $\nu$  and carry the gradient through the integral.

$$\begin{aligned} \nabla_\nu D_\varphi(p||q_\nu) &= \int \nabla_\nu \varphi \left( \log \frac{p(\nu + \zeta)}{q_0(\zeta)} \right) q_0(\zeta) d\zeta \\ &= \int \frac{\nabla_\nu p(\nu + \zeta)}{q_0(\zeta)} \varphi' \left( \log \frac{p(\nu + \zeta)}{q_0(\zeta)} \right) q_0(\zeta) d\zeta \\ &= \int \frac{\nabla_\zeta p(\nu + \zeta)}{q_0(\zeta)} \varphi' \left( \log \frac{p(\nu + \zeta)}{q_0(\zeta)} \right) q_0(\zeta) d\zeta, \end{aligned} \quad (9)$$

where in the final line, we use the symmetry of  $\nu$  and  $\zeta$  in the argument of  $p$  to rewrite the gradient with respect to  $\zeta$  rather than  $\nu$ . If we set  $\nu = \mu$ , then  $p(\nu + \zeta)$  is even symmetric about the origin and  $\nabla_\zeta p(\nu + \zeta)$  is odd symmetric. All other terms in the integrand are even symmetric; hence the integral vanishes, indicating that a stationary point of  $D_\varphi(p||q_\nu)$  occurs at  $\nu = \mu$ .

Now suppose that  $\varphi$  is strictly monotone decreasing and  $p$  is somewhere-strictly log concave on  $\mathbb{R}^d$ . Then



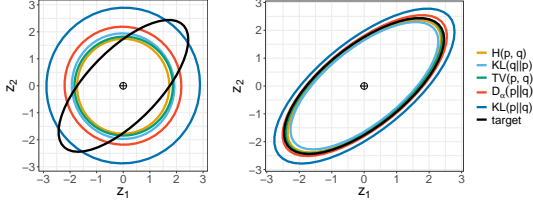


Figure 2: *Variational approximation of a multivariate student-t by a Gaussian. Empirically, for each divergence in Table 1, the mean of the student-t is recovered by a factorized Gaussian approximation (left), while its correlation matrix is recovered by a non-factorized Gaussian approximation (right). However, each divergence returns a different estimate of variance. For the  $\alpha$ -divergence, we use  $\alpha = 0.5$ .*

we show in Lemma 16 of Appendix B.1 that  $D_\varphi(p||q_\nu)$  is strictly convex in  $\nu$ , and hence the stationary point at  $\mu = \nu$  is a unique minimizer.  $\square$

Theorem 10 allows for several misspecifications in the variational approximation. For example,  $q$  may be factorized while  $p$  is not, or  $q$  and  $p$  may behave differently in their tails. Fig. 2 (left) illustrates VI’s ability to recover the mean when  $p(z)$  is a 2-dimensional student-t distribution with 5 degrees of freedom and correlation  $\rho=0.7$ , and  $\mathcal{Q}$  is a family of factorized Gaussians. Minimizing the  $\varphi$ -divergences in Table 1 (via grid search), we find that they all yield the same, exact estimate of the mean  $\nu$ . Interestingly, the exact mean is recovered for *all* five divergences; this is in fact a stronger result than what the theorem guarantees. While the theorem states that the mean-matching solution is a stationary point for all divergences, only the first two divergences in Table 1—the reverse KL and the  $\alpha$ -divergences with  $\alpha \in (0, 1)$ —satisfy the additional assumption that  $\varphi$  is monotone decreasing and which guarantees that  $\nu = \mu$  is a unique minimizer.

We conclude this section with an illustrative example of an *asymmetric* target  $p$ . Let

$$\begin{aligned} p(z) &= \text{skewed-}\mathcal{N}(0, 2^2, \kappa) \\ q(z) &= \text{Laplace}(\nu, 1), \end{aligned} \quad (10)$$

where  $\kappa \in \mathbb{R}^+$  controls the skewness of  $p$ . Here we chose  $\mathcal{Q}$  to be the family of Laplace distributions, so that the approximation remains misspecified even when  $\kappa = 0$  and the target is perfectly even symmetric. For  $\kappa = 0$ , all divergences yield a solution that recovers the mean of  $p$  (within some error in the stochastic grid search). As  $\kappa$  increases, however, the error in the mean increases for all divergences, though not to the same extent; see Fig. 3.

### 3.2 Guarantees from elliptical symmetry

Next we generalize earlier guarantees for VI with elliptical symmetries (Margossian and Saul, 2025) to the broader family of  $\varphi$ -divergences. To begin, we note that if  $p$  is elliptically symmetric, then it admits a point of even symmetry,  $\mu$ , and a normalized covariance matrix,  $M$ . In this case, there exists a spherically symmetric density  $p_0$  satisfying

$$p(z) = p_0(M^{-\frac{1}{2}}(z - \mu))|M|^{-\frac{1}{2}}. \quad (11)$$

Since  $p_0$  is spherically symmetric, we can define a function  $f : [0, \infty) \rightarrow \mathbb{R}$  such that  $f(|t|) = \log p_0(t)$ . With this definition we have the following theorem.

**Theorem 11** (Exact Recovery of the Mean and Correlation Matrix). *Let  $\mathcal{Q}$  be a location-scale family, and let  $D_\varphi$  be a  $\varphi$ -divergence with strictly decreasing  $\varphi$ . Suppose  $p$  is somewhere-strictly log concave over  $\mathbb{R}^d$  and elliptically symmetric about  $\mu$  with normalized covariance matrix  $M$ ; also, suppose  $f$  (as defined above) is everywhere continuously differentiable with  $|f'(0)| < \infty$ . Then  $D_\varphi(p||q)$  has a unique minimizer with respect to the location-scale parameters  $(\nu, S)$  of  $\mathcal{Q}$  at  $\nu = \mu$  and  $S = \gamma^2 M$  for some  $\gamma \in \mathbb{R}$ .*

A full proof is given in Appendix B.2. The main idea is to rewrite  $D_\varphi(p||q_{\nu,S})$  as an integral over spherically symmetric functions in the transformed variable  $\zeta = S^{-\frac{1}{2}}(z - \nu)$ . Then, with some technical machinery, it can be shown that a unique stationary point of  $D_\varphi(p||q_{\nu,S})$  occurs at  $\nu = \mu$  and  $S = \gamma^2 M$ , thus recovering the exact mean and correlation matrix.

As before, the theorem allows for misspecifications in  $\mathcal{Q}$ . Fig. 2 (right) illustrates VI’s ability to recover these statistics when  $p$  is a student-t distribution and  $\mathcal{Q}$  is the family of *multivariate* Gaussians. Though each divergence in Table 1 yields a different approximation, they all recover the mean and correlation of  $p$ . The theorem predicts this result for the reverse KL divergence and the Renyi-divergence with  $\alpha \in (0, 1)$ .

### 3.3 Guarantees from partial symmetry

Finally we provide guarantees under which  $q$  recovers the marginal mean and correlation matrix along coordinates of partial symmetry. We use  $\sigma$  to denote the coordinates along which  $p$  is symmetric and  $\bar{\sigma}$  to denote the remaining coordinates. In this notation, we decompose the variational parameters of  $q$  as

$$\nu = \begin{pmatrix} \nu_\sigma \\ \nu_{\bar{\sigma}} \end{pmatrix}, \quad S = \begin{pmatrix} S_{\sigma\sigma} & S_{\sigma\bar{\sigma}} \\ S_{\bar{\sigma}\sigma}^T & S_{\bar{\sigma}\bar{\sigma}} \end{pmatrix}, \quad (12)$$

and we use  $S_{\sigma|\bar{\sigma}}$  to denote the conditional scale matrix of  $q$ . We also decompose the mean  $\mu = (\mu_\sigma, \mu_{\bar{\sigma}})$  and

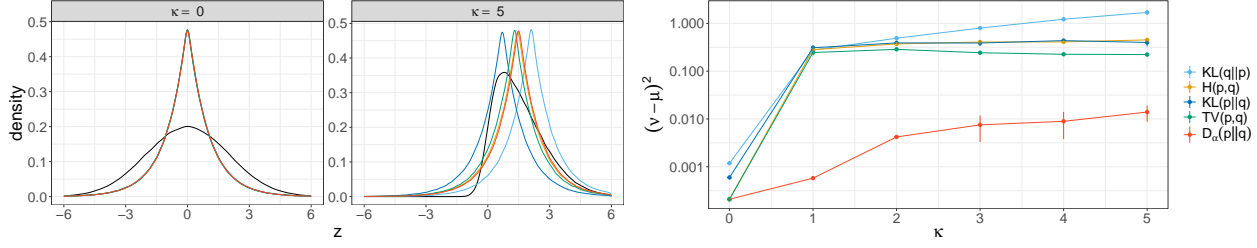


Figure 3: VI approximations to a skewed normal  $p$  with a Laplace distribution. (Left) When  $p$  has no skew ( $\kappa=0$ ), its mean is recovered by VI with all the divergences in Table 1; when  $p$  is largely skewed ( $\kappa=5$ ), the results disagree. (Right) The plot shows the error in the mean estimate (averaged over 10 stochastic optimizations).

covariance  $\Sigma$  of the target density  $p$  in a similar way.

**Theorem 12** (Exact Recovery of the Partial Mean along  $\sigma$ ). *Let  $\mathcal{Q}$  be a location family. If  $p$  is even symmetric along  $\sigma$  with a constant point of even symmetry, then  $\nu_{\sigma} = \mu_{\sigma}$  is a stationary point of  $D_{\varphi}(p||q_{\nu})$  for any fixed  $\nu_{\bar{\sigma}}$ . Also, if  $\varphi$  is strictly decreasing and  $p(z_{\sigma}|z_{\bar{\sigma}})$  is somewhere-strictly log concave over  $\mathbb{R}^{|\sigma|}$ , then  $\nu_{\sigma} = \mu_{\sigma}$  at all stationary points of  $D_{\varphi}(p||q_{\nu})$ .*

A complete proof is given in Appendix B.3, and its steps closely follow those for the proof of Theorem 10.

We now provide guarantees for VI when  $p$  is elliptically symmetric along  $\sigma$ . For these guarantees, we must make a number of additional assumptions—namely, that  $\mathcal{Q}$  is the family of multivariate Gaussians, that the variational approximation is found by minimizing  $\text{KL}(q||p)$  (as opposed to any  $\varphi$ -divergence), and also that the conditional mean and normalized covariance of  $p(z_{\sigma}|z_{\bar{\sigma}})$  do not depend on  $z_{\bar{\sigma}}$ . (The utility of these assumptions is made clear in the proof of the next theorem.) Despite these restrictions, the theorem still has a fairly broad scope: it covers the most common setting of black box VI (e.g. Kucukelbir et al., 2017), where  $\text{KL}(q||p)$  is minimized to find the best multivariate Gaussian approximation  $q$ , and it also covers non-trivial target densities such as the Rosenbrock distribution (Roberts et al., 1997) and the funnel of eq. (6) and some extensions we explore in Appendix B.5. The latter often characterizes hierarchical priors in Bayesian models.

To state the theorem, we introduce (as before) a spherically symmetric distribution  $p_0$ , this time constructed from the conditional distribution

$$p(z_{\sigma}|z_{\bar{\sigma}}) = p_0(\Sigma^{-\frac{1}{2}}(z_{\sigma} - \mu))|\Sigma|^{-\frac{1}{2}}. \quad (13)$$

We also define the function  $f : [0, \infty) \rightarrow \mathbb{R}$  given by  $f(|t|) = \log p_0(t)$ . Then we have the following:

**Theorem 13** (Exact Recovery of the Partial Correlations along  $\sigma$ ). *Let  $\mathcal{Q}$  be the family of multivariate Gaussians. Suppose  $p$  is elliptically symmetric along  $\sigma$  about a constant point of even symmetry and with a constant normalized covariance matrix, and suppose  $f$  is continuously differentiable and  $|f'(0)| < \infty$ . Also suppose  $p(z_{\sigma}|z_{\bar{\sigma}})$  is somewhere-strictly log concave over  $\mathbb{R}^{|\sigma|}$ . Then any minimizer of  $\text{KL}(q||p)$  satisfies  $\nu_{\sigma} = \mu_{\sigma}$  and  $S_{\sigma\sigma} = \gamma^2 \Sigma_{\sigma\sigma}$  for some unique  $\gamma \in \mathbb{R}$ .*

*Proof.* The variational approximation is found by minimizing  $\text{KL}(q||p)$ . We can decompose  $\text{KL}(q||p)$  as

$$\text{KL}(q||p) = \text{KL}(q_{\bar{\sigma}}||p_{\bar{\sigma}}) + \int_{\bar{\sigma}} \text{KL}(q_{\sigma|\bar{\sigma}}||p_{\sigma|\bar{\sigma}}) q(z_{\bar{\sigma}}) dz_{\bar{\sigma}}. \quad (14)$$

Note that the reverse KL divergence, for which  $\varphi(t)$  is linear in Table 1, is the only strictly decreasing  $\varphi$ -divergence that permits such a decomposition. The keystone of the proof is to find values for  $(\nu_{\sigma}, S_{\sigma\sigma}, S_{\sigma\bar{\sigma}})$  that uniformly minimize the integrand of eq. (14) for all values of  $z_{\bar{\sigma}}$ , and then to note that this minimum is realized for any values of  $(\nu_{\bar{\sigma}}, S_{\bar{\sigma}\bar{\sigma}})$  that fix  $q(z_{\bar{\sigma}})$  and hence  $\text{KL}(q_{\bar{\sigma}}||p_{\bar{\sigma}})$  in the remaining term.

We now show how to find these values. Per Theorem 11, the KL divergence in the integrand of eq. (14) is minimized if  $q(z_{\sigma}|z_{\bar{\sigma}})$  matches the mean and normalized covariance matrix of  $p(z_{\sigma}|z_{\bar{\sigma}})$ , or equivalently if  $E_q[z_{\sigma}|z_{\bar{\sigma}}] = E_p[z_{\sigma}|z_{\bar{\sigma}}]$  and  $\text{Cov}_q[z_{\sigma}|z_{\bar{\sigma}}] \propto \text{Cov}_p[z_{\sigma}|z_{\bar{\sigma}}]$ . From the properties of conditional Gaussians, we then require that

$$E_p[z_{\sigma}|z_{\bar{\sigma}}] = \nu_{\sigma} + S_{\sigma\bar{\sigma}} S_{\bar{\sigma}\bar{\sigma}}^{-1} (z_{\bar{\sigma}} - \nu_{\bar{\sigma}}), \quad (15)$$

$$\text{Cov}_p[z_{\sigma}|z_{\bar{\sigma}}] \propto (S_{\sigma\sigma} - S_{\sigma\bar{\sigma}} S_{\bar{\sigma}\bar{\sigma}}^{-1} S_{\bar{\sigma}\sigma}). \quad (16)$$

Since  $p$  has a constant point of even symmetry, the right side of eq. (15) cannot depend on  $z_{\bar{\sigma}}$ , and so we set  $S_{\sigma\bar{\sigma}} = \mathbf{0}$ . Then eq. (15) is solved by  $\nu_{\sigma} = E_p[z_{\sigma}|z_{\bar{\sigma}}]$  and eq. (16) by  $S_{\sigma\sigma} \propto \text{Cov}_p[z_{\sigma}|z_{\bar{\sigma}}]$ . Finally, we show in Lemma 18 of Appendix B.4 that  $\mu_{\sigma} = E_p[z_{\sigma}|z_{\bar{\sigma}}]$  and  $\Sigma_{\sigma\sigma} \propto \text{Cov}_p[z_{\sigma}|z_{\bar{\sigma}}]$  if  $p$  is elliptically symmetric along

Call name	$d$	Description	$\alpha_{90}$	$\alpha_{90}$ (nc)	$\alpha_{90}$ (marg)
student	2	Elliptical target with heavy tails.	0		
funnel	4	Elliptical funnel with partial symmetry (eq. 6).	0.061		
crescent	3	Elliptical Rosenbrock distribution.	29.36		
schools	10	Bayesian hierarchical model for education data.	287.79	1.74	1.03
disease	102	Gaussian process model for epidemiology data.	231,836	137.80	0.75
SKIM	305	Sparse kernel interaction model for genetic data.	220,485	138.29	NA

Table 2: Target distributions for experiments and estimated asymmetry  $\alpha_{90}$ ; see eq. (19). The hierarchical models (*schools*, *disease*, *SKIM*) were implemented using (i) a standard centered parameterization, (ii) a non-centered parameterization with less asymmetry, (iii) an approximate marginalization of the latent variables, yielding a collapsed (and even less asymmetric) posterior; see the discussion in section 4.1.

$\sigma$  with a constant point of even symmetry and a constant normalized covariance matrix. This completes the proof.  $\square$

Some of the conditions for this proof can be relaxed. In Appendix B.4, we consider the case where  $\mathbb{E}_p[z_\sigma|z_{\bar{\sigma}}]$  is a linear function of  $z_{\bar{\sigma}}$  and show that Gaussian VI provably recovers the conditional mean and correlations. In Appendix B.5, we apply these theoretical results to the funnel of eq. (6) and variations thereof.

## 4 NUMERICAL EXPERIMENTS

We investigate the performance of VI on a diverse set of target densities (Table 2). The first three are synthetic—a multivariate student-t, an elliptical funnel, and a crescent distribution—and the others are derived from Bayesian hierarchical models, including a model of education data (Gelman et al., 2013), a sparse kernel interaction model of gene microarray data (Agrawal et al., 2019), and a Gaussian process model of mortality counts with a Poisson likelihood (Vanhatalo et al., 2010). The full definition of each model is provided in Appendix C.1.

### 4.1 Are Bayesian posteriors symmetric?

We now discuss how approximate symmetries may arise in a Bayesian analysis. Given a latent variable  $z$  and observation  $x$ , the posterior is given by  $\pi(z|x) \propto \pi(z)\pi(x|z)$ , and the likelihood often factors as  $\pi(x|z) = \prod_i \pi(x_i|z)$ . In general this product may not have an even or elliptical symmetry even when one is found in each of its terms. However, if the product is dominated by a symmetric prior (for sparse data) or the likelihood (for rich data), then the posterior tends to be approximately symmetric (Margossian and Saul, 2025). In the latter regime, the Bernstein-von Mises theorem (van der Vaart, 1998) is often invoked to argue the posterior is approximately Gaussian.

One expects less symmetry in hierarchical models with

asymmetric priors, as in eq. (6). Further complexity arises when the prior mean and the correlation  $C$  for  $\theta$  depend on additional hyperparameters  $\mu, \rho \in \mathbb{R}$ ,

$$\theta \sim \mathcal{N}(\mu, e^{2\tau} C(\rho)), \quad (17)$$

as in Gaussian processes and, for example, the *disease* and *SKIM* targets in Table 2. But there are also strategies to mitigate these sources of asymmetry, without changing the generative model. One strategy is to use a *non-centered parameterization* (Papaspiliopoulos et al., 2007). Let  $L$  denote the Cholesky decomposition of the prior covariance matrix in eq. (17), such that  $LL^T = \exp(2\tau)C(\rho)$ . This strategy introduces an auxiliary variable  $\varepsilon \sim \mathcal{N}(0, 1)$  and recomputes the likelihood (equivalently) as  $\pi(x|\varepsilon) = \pi(x|\theta = L\varepsilon + \mu)$ . A Bayesian inference algorithm then approximates the posterior  $\pi(\varepsilon, \mu, \tau, \rho|x)$ , where the funnels in the prior have been removed due to the independence of  $\varepsilon$  and  $\tau$ . A caveat is that this non-centered parameterization can complicate the likelihood, especially in rich data regimes, leading to a challenging posterior geometry.

A second strategy is to marginalize out  $\theta$  and perform inference over the collapsed posterior,

$$\pi(\mu, \tau, \rho|x) \propto \pi(\mu, \tau, \rho) \int_{\Theta} \pi(x, \theta|\mu, \tau, \rho) d\theta. \quad (18)$$

Here the marginalization serves to remove the funnel over  $\theta$  and  $\tau$  in the prior. Inference on  $\theta$  is performed post-hoc by approximating  $\pi(\theta|\mu, \tau, \rho, x)$ . If the likelihood is normal, then the marginalization is tractable; otherwise it must be approximated—for example, via an integrated Laplace approximation (e.g., Rasmussen and Williams, 2006; Rue et al., 2009).

We implement each hierarchical model in three ways, via a standard (centered) parameterization, a non-centered parameterization, and a marginalized target (where the marginalization is performed exactly for *schools* and approximately for *disease* and *SKIM*). Empirically, we find the latter strategies to produce less asymmetric targets; see Table 2.

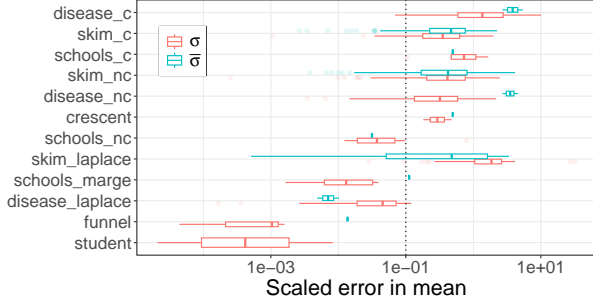


Figure 4: Absolute error in VI’s mean estimate scaled by the target’s standard deviation. The targets are ordered, bottom to top, from most to least symmetric. The dotted line is the standard error obtained with 100 independent draws. As a trend, VI returns better estimates of the mean for more symmetric targets. The mean is also better estimated in the *funnel*, *crescent*, and *disease* along the coordinates  $\sigma$  whose priors exhibit a partial symmetry.

## 4.2 VI algorithm

We use automatic differentiation VI (ADVI, Kucukelbir et al., 2017), as implemented in STAN (Carpenter et al., 2017; Stan Development Team, 2025). ADVI employs a Gaussian approximation and minimizes  $\text{KL}(q||p)$  via stochastic optimization. Details on the implementation of ADVI are described in Appendix C.2. While our theorems consider more general  $\varphi$ -divergences, we do not have a reliable way to minimize these divergences, and so we only give results for  $\text{KL}(q||p)$ . For models with a collapsed posterior, we use STAN’s prototype integrated Laplace approximation (Margossian et al., 2020). As a benchmark, we approximate  $\pi(z|x)$  with exact sampling of the synthetic targets and long runs (20000 iterations) of STAN’s Hamiltonian Monte Carlo sampler (Hoffman and Gelman, 2014; Betancourt, 2018) with a non-centered parameterization for the hierarchical models.

## 4.3 Results

Our first results investigate the ability of VI to recover the mean in the presence (or absence) of even symmetry. For each target density in Table 2, we stochastically measure its asymmetry. Given a sample  $z$  from  $p(z)$ , we compute a reflected sample  $z' = \hat{\mu} - z$ , where  $\hat{\mu}$  is the benchmark estimate of the mean of  $p$ . Then we measure the asymmetry of the target by computing

$$\alpha(z) = |\log \pi(z, x) - \log \pi(z', x)|. \quad (19)$$

If  $\pi(x, z)$  is even-symmetric about  $\hat{\mu}$ , then  $\alpha(z) = 0$  for all  $z$ . We evaluate  $\alpha(z)$  for 20,000 samples, and report its 90<sup>th</sup> quantile in Table 2. This procedure

works for all of the models except SKIM (where there are numerical instabilities in the reflected density due to the Laplace approximation).

From the above procedure, we know which targets are more or less (approximately) even symmetric. Next we report the absolute error in the means that are estimated by VI across all coordinates; see Fig. 4. Overall, we find that for more symmetric targets, VI yields better estimates of the mean. We also plot the errors separately along coordinates which are (a priori) even-symmetric versus those which are not. In the *funnel*, *crescent* and *disease* models, the mean is better estimated along the former, but in other models, the errors are comparable.

We find a similar trend for estimates of correlations; generally VI yields better estimates in models with more symmetry. This trend is clear for the synthetic targets, less so for the hierarchical models, perhaps because the latter have many nearly zero correlations. We provide these results in Appendix C.3.

## 5 Discussion

In this paper we derive novel symmetry-based guarantees for VI with a broad class of divergences and in cases where  $p$  exhibits partial symmetries. Our results provide not only theoretical insight, but also prescriptions for practitioners using VI to approximate the posteriors of Bayesian hierarchical models. They suggest, in particular, that VI can be improved by implementing these models in certain ways. These prescriptions are reminiscent of those known to improve the performance of MCMC samplers on challenging posteriors (e.g., Betancourt and Girolami, 2015; Gómez-Rubio and Rue, 2018; Margossian et al., 2020).

The above considerations suggest one way to improve VI in Bayesian hierarchical models. But a more common approach is simply to choose a richer variational family  $\mathcal{Q}$ —for example, one allowing skewed approximations (Tan and Chen, 2024), or even semi-parametric or non-parametric approximations (e.g., Agrawal et al., 2020; Xu et al., 2023; Cai et al., 2024a; Xu and Campbell, 2025). While this approach requires less bespoke manipulations of  $p$ , its computational expense grows quickly with the complexity of  $\mathcal{Q}$ . It is therefore of interest to understand how well simpler variational families can perform. Moving forward, we advocate a workflow in which VI proceeds first with a restricted family  $\mathcal{Q}$  (e.g., factorized or location-scale), the accuracy of the inference is checked (here, it would be interesting to develop a symmetry-based check), and then slight corrections are applied (e.g., Yao et al., 2018; Giordano et al., 2018) and  $\mathcal{Q}$  is progressively complexified as necessary.



## References

- Agrawal, A., Sheldon, D. R., and Domke, J. (2020). Advances in black-box vi: Normalizing flows, importance weighting, and optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 2192–2205.
- Agrawal, R., Huggins, J. H., Trippe, B., and Broderick, T. (2019). The Kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions. In *Proceedings of the 36th International Conference on Machine Learning*, pages 141–150.
- Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497.
- Betancourt, M. (2018). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434v1*.
- Betancourt, M. and Girolami, M. (2015). Hamiltonian monte carlo for hierarchical models. In *Current Trends in Bayesian Methodology with Applications*, page 24. Chapman and Hall/CRC.
- Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, New York, 3 edition.
- Biswas, N. and Mackey, L. (2023). Bounding Wasserstein distance with couplings. *Journal of the American Statistical Association*, 119(548):2947–2958.
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877.
- Cai, D., Modi, C., Margossian, C. C., Gower, R. M., Blei, D. M., and Saul, L. K. (2024a). EigenVI: score-based variational inference with orthogonal function expansions. In *Advances in Neural Information Processing Systems 37*, pages 132691–132721.
- Cai, D., Modi, C., Pillaud-Vivien, L., Margossian, C. C., Gower, R. M., Blei, D. M., and Saul, L. K. (2024b). Batch and match: black-box variational inference with a score-based divergence. *Proceedings of the 41st International Conference on Machine Learning*, page 5258–5297.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32.
- Cichocki, A. and Amari, S.-i. (2010). Families of alpha-beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568.
- Daudel, K., Benton, J., and Doucet, A. (2023). Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics. *Journal of Machine Learning Research*, 24(243):1–83.
- Dhaka, A. K., Catalina, A., Welandawe, M., Andersen, M. R., Huggins, J., and Vehtari, A. (2021). Challenges and opportunities in high dimensional variational inference. In *Advances in Neural Information Processing Systems 34*, pages 7787–7798.
- Dieng, A. B., Tran, D., Ranganath, R., Paisly, J., and Blei, D. M. (2017). Variational inference via  $\chi$  upper bound minimization. In *Advances in Neural Information Processing Systems 30*, pages 2732–2741.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science.
- Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 19(51):1–49.
- Gómez-Rubio, V. and Rue, H. (2018). Markov chain monte carlo with the integrated nested laplace approximation. *Statistics and Computing*, 28(5):1033–1051.
- Hoffman, M. D. and Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623.
- Huggins, J. H., Kasprzak, M., Campbell, T., and Broderick, T. (2020). Validated variational inference via practical posterior error bounds. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 1792–1802.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.
- Katsevich, A. and Rigollet, P. (2024). On the approximation accuracy of Gaussian variational inference. *The Annals of Statistics*, 52(4):1384–1409.
- Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(14):1–45.
- Li, Y. and Turner, R. E. (2016). Rényi divergence variational inference. In *Advances in Neural Information Processing Systems 29*, pages 1073–1081.

- MacKay, D. J. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press.
- Margossian, C. C., Pillaud-Vivien, L., and Saul, L. K. (2025). Variational inference for uncertainty quantification: an analysis of trade-offs. *Journal of Machine Learning Research*. (To appear).
- Margossian, C. C. and Saul, L. K. (2023). The shrinkage-delinkeage trade-off: An analysis of factorized Gaussian approximations for variational inference. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pages 1358–1367.
- Margossian, C. C. and Saul, L. K. (2025). Variational inference in location-scale families: Exact recovery of the mean and correlation matrix. In *Proceedings of the 28th Conference on Artificial Intelligence and Statistics*, pages 3466–3474.
- Margossian, C. C., Vehtari, A., Simpson, D., and Agrawal, R. (2020). Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent gaussian models and beyond. In *Advances in Neural Information Processing Systems 33*, pages 9086–9097.
- Naesseth, C. A., Lindsten, F., and Blei, D. M. (2020). Markovian score climbing: Variational inference with  $KL(p||q)$ . In *Advances in Neural Information Processing Systems 34*, pages 15499–15510.
- Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, 11:125–139.
- Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A general framework for the parametrization of hierarchical models. *Statistical Science*, 22(1):59–73.
- Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051.
- Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 814–822.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.
- Rényi, A. (1961). On measures of entropy and information. In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume I: Contributions to the theory of statistics*, pages 547–561. University of California Press, Berkeley, California.
- Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.
- Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.
- Roberts, G. O. and Rosenthal, J. S. (2004). General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71.
- Roualdes, E. A., Ward, B., Carpenter, B., Syboldt, A., and Axen, S. D. (2023). Bridgestan: Efficient in-memory access to the methods of a Stan model. *Journal of Open Source Software*, 8.
- Rubin, D. B. (1981). Estimation in parallelized randomized experiments. *Journal of Educational Statistics*, 6:377–400.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.
- Stan Development Team (2025). Stan modeling language users guide and reference manual.
- Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv:1804.06788*.
- Tan, L. S. L. and Chen, A. (2024). Variational inference based on a subclass of closed skew normals. *Journal of Computational and Graphical Statistics*, pages 1–15.
- Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press.
- van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Statistics in Medicine*, 29:1580–1607.
- Vehtari, A., Gelman, A., Sivula, T., Jylanki, P., Tran, D., Sahai, S., Blomstedt, P., Cunningham, J. P., Schiminovich, and Robert, C. P. (2020). Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *Journal of Machine Learning Research*, 21(17):1–53.
- Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2024). Pareto smoothed importance

- sampling. *Journal of Machine Learning Research*, 25(72):1–58.
- Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.
- Wang, Y. and Blei, D. M. (2018). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114:1147–1161.
- Xu, Z. and Campbell, T. (2025). Asymptotically exact variational flows via involutive MCMC kernels. In *Advances in Neural Information Process Systems 37*. (To appear).
- Xu, Z., Chen, N., and Campbell, T. (2023). MixFlows: principled variational inference via mixed flows. In *Proceedings of the 40th International Conference on Machine Learning*, pages 38342–38376.
- Yang, E., Pati, D., Jordan, M. I., and Wainwright, M. J. (2020).  $\alpha$ -variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905.
- Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Yes, but did it work?: Evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5577–5586.
- Zhang, F. and Gao, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207.

---

## Appendix

---

### A CONNECTION BETWEEN $\varphi$ -DIVERGENCES AND $f$ -DIVERGENCES

In this appendix, we review  $f$ -divergences and discuss their connection to  $\varphi$ -divergences. As in the rest of the paper, we assume  $p$  and  $q$  admit a density with respect to a Lebesgue measure.

**Definition 14.** ( *$f$ -divergence Rényi (1961)*) We refer to a divergence as an  $f$ -divergence, if it can be written as

$$D_f(p||q) = \int f\left(\frac{p(z)}{q(z)}\right) q(z) dz, \quad (20)$$

where (i)  $f(t)$  is convex, (ii)  $|f(t)| < \infty$  for all  $t > 0$ , (iii)  $f(1) = 0$ , and  $f(0) = \lim_{t \rightarrow 0^+} f(t)$ .

The  $f$ -divergence is not a generalization of the  $\varphi$ -divergence, nor is it a special case thereof. Rather, each definition imposes different restrictions on a function that acts on the ratio  $p(z)/q(z)$  inside the integral. There are many examples of divergences which are both  $f$ - and  $\varphi$ -divergences, including the divergences in Table 1, with one notable exception. The Rényi divergence of order  $\alpha$ ,

$$D_\alpha(p||q) = \int \frac{1}{\alpha(1-\alpha)} \left( \frac{p^\alpha(z)}{q^\alpha(z)} - 1 \right) dz, \quad (21)$$

is an  $f$ -divergence only if  $\alpha > 1$ . Indeed, for  $\alpha \in (0, 1)$ , the function  $f$  acting on  $t = p(z)/q(z)$ ,  $f(t) = (t^\alpha - 1)/(\alpha(\alpha - 1))$ , is not convex. However, one can verify that  $D_\alpha(p||q)$  is still a valid  $\varphi$ -divergence—not only for  $\alpha \in (0, 1)$  but for all values of  $\alpha \in \mathbb{R}^+ \setminus \{0, 1\}$ . The case where  $\alpha \in (0, 1)$  is important, because it interpolates between the reverse and forward KL-divergences. Hence, we chose to write our theoretical statements in terms of  $\varphi$ -divergences rather than  $f$ -divergences, in order to include all valid  $\alpha$ -divergences.

Another reason for working with the  $\varphi$ -divergences is because the regularity conditions in Section 2.4 concern the *log*-concavity of  $p$ . Hence it is more convenient to directly work with a function of  $\log p$ .

### B SUPPORTING PROOFS

In this appendix we provide additional Lemmas which support the proofs in the main body, as well as an extension of Theorem 13.



### B.1 Strict convexity of $D_\varphi(p||q_\nu)$ in $\nu$

In this section, we show that if  $p$  is somewhere-strictly log concave and  $\varphi$  is monotone decreasing, then  $D_\varphi(p||q_\nu)$  is strictly convex in  $\nu$ . We begin with a weaker result.

**Lemma 15.** *Let  $q_\nu$  be a location distribution with location parameter  $\nu \in \mathbb{R}^d$  and base distribution  $q_0$ . Suppose  $p$  is log concave over  $\mathbb{R}^d$  and  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  is monotone decreasing and convex. Then  $\varphi(\log p(z) - \log q_\nu(z))$  is convex in  $\nu$  for all  $z$ .*

*If in addition, (i)  $\varphi$  is strictly monotone decreasing and  $p$  is strictly log concave, OR (ii)  $\varphi$  is strictly convex, then  $\varphi(\log p(z) - \log q_\nu(z))$  is strictly convex in  $\nu$ .*

*Proof.* Let  $\zeta = z - \nu$ . Then

$$\log \frac{p(z)}{q_\nu(z)} = \log \frac{p(\zeta + \nu)}{q_0(\zeta)} =: g_\zeta(\nu), \quad (22)$$

where the function  $g_\zeta(\nu)$  is introduced for notational convenience. By assumption  $p(z)$  is log concave and therefore  $p(z + \nu)$  is log concave in  $\nu$ . Explicitly, for  $\lambda \in (0, 1)$  and  $\nu_0, \nu_1 \in \mathbb{R}$ ,

$$g_\zeta((1 - \lambda)\nu_0 + \lambda\nu_1) \geq (1 - \lambda)g_\zeta(\nu_0) + \lambda g_\zeta(\nu_1). \quad (23)$$

Then, from the decreasing monotonicity and convexity of  $\varphi$ ,

$$\begin{aligned} \varphi\left(\log \frac{p(\zeta + (1 - \lambda)\nu_0 + \lambda\nu_1)}{q_0(\zeta)}\right) &= \varphi(g_\zeta((1 - \lambda)\nu_0 + \lambda\nu_1)) \\ &\leq \varphi((1 - \lambda)g_\zeta(\nu_0) + \lambda g_\zeta(\nu_1)) \end{aligned} \quad (24)$$

$$\leq (1 - \lambda)\varphi(g_\zeta(\nu_0)) + \lambda\varphi(g_\zeta(\nu_1)) \quad (25)$$

$$= (1 - \lambda)\varphi\left(\log \frac{p(\zeta + \nu_0)}{q_0(\zeta)}\right) + \lambda\varphi\left(\log \frac{p(\zeta + \nu_1)}{q_0(\zeta)}\right). \quad (26)$$

Thus,  $\varphi(\log p(z) - \log q_\nu(z))$  is convex in  $\nu$ .

Furthermore:

- (i) The inequality in eq. (24) is strict if  $\varphi$  is strictly monotone decreasing and  $p$  is strictly log concave in  $z$ .
- (ii) The inequality in eq. (25) is strict if  $\varphi$  is strictly convex.

In either case, the inequality in eq. (26) becomes strict. □

Building on Lemma 15, we obtain a result about the strict convexity of  $D_\varphi(p||q)$ .

**Lemma 16.** *Assume the regularity conditions on  $p$  and  $q$  described in Section 2.4, and in particular that  $p$  is log concave over  $\mathbb{R}^d$ . Consider the  $\varphi$ -divergence  $D_\varphi(p||q)$  and suppose one of the following:*

- (i)  *$p$  is somewhere-strictly log concave and  $\varphi$  is strictly monotone decreasing, OR*
- (ii)  *$\varphi$  is monotone decreasing and strictly convex.*

*Then  $D_\varphi(p||q_\nu)$  is strictly convex in  $\nu$ .*

*Proof.* We will start by assuming condition (i).

Let  $\nu_0, \nu_1 \in \mathbb{R}^d$  and let  $\lambda \in (0, 1)$ . As a shorthand, we write  $\nu_\lambda = (1 - \lambda)\nu_0 + \lambda\nu_1$ . By assumption, there exists a set  $\Omega$  over which  $p$  is strictly log concave. Let

$$\Omega_\lambda = \{\zeta \in \mathbb{R}^d | \nu_\lambda + \zeta \in \Omega\}. \quad (27)$$

Consider the function  $g_\zeta : \mathbb{R}^d \rightarrow \mathbb{R}$ , with  $g_\zeta(\nu) = \log p(\zeta + \nu) - \log q_0(\zeta)$ . We will now derive an inequality on  $g_\zeta$ , using Proposition 11 of Margossian and Saul (2025). The latter only applies to functions with a univariate

input and so we introduce  $f_\zeta : \mathcal{R} \rightarrow \mathbb{R}$ , a function whose domain  $\mathcal{R}$  is the line that goes through  $\nu_0$  and  $\nu_1$ . For  $\nu \in \mathcal{R}$ ,  $f_\zeta(\nu) = g_\zeta(\nu)$ . Then, for  $\zeta \in \Omega_\lambda$  and applying Proposition 11 in [Margossian and Saul \(2025\)](#), we have

$$f_\zeta(\nu_\lambda) > (1 - \lambda)f_\zeta(\nu_0) + \lambda f_\zeta(\nu_1). \quad (28)$$

Then,

$$g_\zeta(\nu_\lambda) > (1 - \lambda)g_\zeta(\nu_0) + \lambda g_\zeta(\nu_1). \quad (29)$$

By assumption (condition (i)),  $\varphi$  is strictly monotone decreasing and so,

$$\varphi(g_\zeta(\nu_\lambda)) < (1 - \lambda)\varphi(g_\zeta(\nu_0)) + \lambda\varphi(g_\zeta(\nu_1)). \quad (30)$$

Next, denote  $\bar{\Omega}_\lambda$  the complement of  $\Omega_\lambda$ . Then

$$D_\varphi(p||q_\nu) = \int_{\Omega_\lambda} \varphi(g_\zeta(\nu))q_0(\zeta)d\zeta + \int_{\bar{\Omega}_\lambda} \varphi(g_\zeta(\nu))q_0(\zeta)d\zeta.$$

From eq. (30), we have

$$\int_{\Omega_\lambda} \varphi(g_\zeta(\nu_\lambda))q_0(\zeta)d\zeta < \int_{\Omega_\lambda} [(1 - \lambda)\varphi(g_\zeta(\nu_0)) + \lambda\varphi(g_\zeta(\nu_1))]q_0(\zeta)d\zeta. \quad (31)$$

Next,  $p$  is log concave and  $\varphi$  monotone decreasing by assumption. Then, by Lemma 15,

$$\int_{\bar{\Omega}_\lambda} \varphi(g_\zeta(\nu_\lambda))q_0(\zeta)d\zeta \leq \int_{\bar{\Omega}_\lambda} [(1 - \lambda)\varphi(g_\zeta(\nu_0)) + \lambda\varphi(g_\zeta(\nu_1))]q_0(\zeta)d\zeta. \quad (32)$$

Combining these two inequalities, we obtain that  $D_\varphi(p||q)$  is strictly convex.

We now consider assumption (ii). In that case, we apply Lemma 15 and obtain from the strict convexity of  $\varphi$  that  $\varphi(g_\zeta(\nu))$  is strictly convex, that is,

$$\varphi(g_\zeta(\nu_\lambda)) < (1 - \lambda)\varphi(g_\zeta(\nu_0)) + \lambda\varphi(g_\zeta(\nu_1)). \quad (33)$$

The wanted result follows.  $\square$

## B.2 Proof for Guarantees in the Presence of Elliptical Symmetry

In this appendix, we provide a proof for Theorem 11.

*Proof.* The proof proceeds in two steps. First, we show that  $D_\varphi(p||q_{\nu,S})$  is strictly convex in  $S^{1/2}$  and therefore admits a unique minimizer  $S$ . Second, we derive a stationary point which has the desired form  $S^{1/2} = \gamma M^{\frac{1}{2}}$  for some  $\gamma > 0$ .

Let  $\zeta = S^{-\frac{1}{2}}(z - \nu)$ . Then  $q(z) = q_0(\zeta)|S|^{-\frac{1}{2}}$  and the objective function becomes

$$\begin{aligned} D_\varphi(p||q) &= \int \varphi\left(\log \frac{p(z)}{q(z)}\right) q(z)dz \\ &= \int \varphi\left(\log \frac{p(S^{\frac{1}{2}}\zeta + \nu)}{|S|^{-\frac{1}{2}}q_0(\zeta)}\right) q_0(\zeta)d\zeta \end{aligned} \quad (34)$$

Next,

$$\begin{aligned} p(z) &= p_0\left(M^{-\frac{1}{2}}(z - \mu)\right) |M|^{-\frac{1}{2}} \\ &= p_0\left(M^{-\frac{1}{2}}(S^{\frac{1}{2}}\zeta + \nu - \mu)\right) |M|^{-\frac{1}{2}}. \end{aligned} \quad (35)$$

Since  $p$  and  $q$  are elliptically symmetric, they are also even-symmetric, and we can apply Theorem 10 to show that  $\nu = \mu$  is a unique minimizer of  $D_\varphi(p||q)$ . Then

$$\begin{aligned} D_\varphi(p||q) &= \int \varphi \left( \log \frac{p_0(M^{-\frac{1}{2}} S^{\frac{1}{2}} \zeta) |M|^{-\frac{1}{2}}}{q_0(\zeta) |S|^{-\frac{1}{2}}} \right) q_0(\zeta) d\zeta \\ &= \int \varphi \left( \log p_0(M^{-\frac{1}{2}} (S^{\frac{1}{2}} \zeta)) - \log q_0(\zeta) + \log |S^{\frac{1}{2}}| - \log |M^{\frac{1}{2}}| \right). \end{aligned} \quad (36)$$

By assumption on  $p$ ,  $p_0$  is log somewhere-strictly log concave in  $S^{1/2}$ . Furthermore, since  $S^{1/2}$  is positive-definite,  $\log |S^{1/2}|$  is concave in  $S^{1/2}$ . By definition of the  $\varphi$ -divergence,  $\varphi$  is convex and, by further assumption,  $\varphi$  is strictly decreasing. Applying the same reasoning as in the proof of Lemma 16, we have that  $D_\varphi(p||q)$  is strictly convex in  $S^{1/2}$ . This completes the first part of the proof.

We now find a stationary point of  $D_\varphi(p||q)$ . Let  $J = M^{-\frac{1}{2}} S^{\frac{1}{2}}$ . Since  $M^{-1/2}$  is fixed and invertible, the uniqueness of a solution for  $S^{1/2}$  implies there is a unique solution for  $J$ . To complete the proof, we must show that  $J = \gamma I$ —with  $I$  the identity matrix—is a stationary point for some  $\gamma \in \mathbb{R}$ . Since  $q_0$  is spherically symmetric, we can define  $g : \mathbb{R}^+ \rightarrow \mathbb{R}$  such that

$$g(||\zeta||) = q_0(\zeta). \quad (37)$$

Then, recalling that  $f(||J\zeta||) = \log p_0(J\zeta)$ , we have

$$\begin{aligned} D_\varphi(p||q) &= \int \varphi \left( \log p_0(J\zeta) + \frac{1}{2} \log |J| - \log q_0(\zeta) \right) q_0(\zeta) d\zeta \\ &= \int \varphi \left( f(||J\zeta||) - \log g(||\zeta||) + \frac{1}{2} \log |J| \right) g(||\zeta||) d\zeta. \end{aligned} \quad (38)$$

Differentiating with respect to  $J$ ,

$$\partial_J D_\varphi(p||q) = \int \left[ f'(||J\zeta||) \frac{J\zeta\zeta^T}{||J\zeta||} + \frac{1}{2} J^{-1} \right] \varphi' \left( \log \frac{f(||J\zeta||)}{g(||\zeta||)} |J|^{\frac{1}{2}} \right) g(||\zeta||) d\zeta. \quad (39)$$

We now plug into this expression  $J = \gamma I$  to obtain

$$\partial_J D_\varphi(p||q) = \int \left[ f'(\gamma ||\zeta||) \frac{\zeta\zeta^T}{||\zeta||} + \frac{1}{2} \gamma^{-1} I \right] \varphi' \left( \log \frac{f(\gamma ||\zeta||)}{g(||\zeta||)} \gamma^{\frac{1}{2}} \right) g(||\zeta||) d\zeta \quad (40)$$

It remains to show that there exists  $\gamma$  such that

$$\int \left[ f'(\gamma ||\zeta||) \frac{\zeta\zeta^T}{||\zeta||} \right] \varphi' \left( \log \frac{f(\gamma ||\zeta||)}{g(||\zeta||)} \gamma^{\frac{1}{2}} \right) g(||\zeta||) d\zeta = -\frac{1}{2} \gamma^{-1} I \int \varphi' \left( \log \frac{f(\gamma ||\zeta||)}{g(||\zeta||)} \gamma^{\frac{1}{2}} \right) g(||\zeta||) d\zeta. \quad (41)$$

The R.H.S of eq. (41) is a scalar product of the identity matrix. We now check that L.H.S of eq. (41) is also a scalar multiple of the identity. We obtain the  $(i, j)^{\text{th}}$  component of the L.H.S by noting that  $(\zeta\zeta^T)_{ij} = \zeta_i\zeta_j$ . Note that all other components in the integrand are spherically symmetric in  $\zeta$ . Then, the integrand is an even function if  $i = j$ , else it is an odd function. Therefore, all the non-diagonal components vanish when integrating. Finally, we note that for each coordinate  $(i, i)$ , the integral is the same and so the L.H.S is indeed a scalar product of the identity matrix.

Since both sides of eq. (41) are scalar multiples of the identity matrix, we can solve this equation by equating the traces on each side. For convenience, we first define

$$h(\gamma, ||\zeta||) = \varphi' \left( \log \frac{f(\gamma ||\zeta||)}{g(||\zeta||)} \gamma^{\frac{1}{2}} \right) g(||\zeta||) \quad (42)$$

Then, taking the trace on both sides of eq. (41), we have,

$$\int f'(\gamma ||\zeta||) ||\zeta|| h(\gamma, ||\zeta||) d\zeta = -\frac{d}{2} \gamma^{-1} \int h(\gamma, ||\zeta||) d\zeta. \quad (43)$$

Equivalently,

$$\int \left( f'(\gamma \|\zeta\|) \|\zeta\| + \frac{d}{2} \gamma^{-1} \right) h(\gamma, \|\zeta\|) d\zeta = 0. \quad (44)$$

By assumption,  $\varphi$  is monotone decreasing and so  $\varphi'(t) \leq 0$ . Also  $g(\|\zeta\|) \geq 0$  and so  $h(\gamma, \|\zeta\|) \leq 0$ . Next, it follows from the assumption of log-concavity and symmetry that  $p_0$  is maximized at the origin. From this, we deduce that  $f$  is monotone decreasing and  $f'(\gamma \|\zeta\|) \leq 0$ . We now examine the R.H.S of eq. (44) in the limits where  $\gamma \rightarrow 0$  and  $\gamma \rightarrow \infty$ :

- $\gamma \rightarrow 0$ : In this limit  $d\gamma^{-1}/2 \rightarrow \infty$ . By assumption,  $|f'(0)| < \infty$  and so the term in parenthesis must be positive. Since  $h(\gamma, \|\zeta\|) \leq 0$ , the integral is negative.
- $\gamma \rightarrow \infty$ : In this limit,  $d\gamma^{-1}/2 \rightarrow 0$ , while  $f'(\gamma \|\zeta\|) \rightarrow -\infty$  by concavity of  $f$ . We still have  $h(\gamma, \|\zeta\|) \leq 0$  and so the integral is positive.

To complete the proof, we show that the R.H.S of eq. (44) is continuous in  $\gamma$ . This follows from the assumption that  $f'$  and  $\varphi'$  are continuous. Therefore, there exists at least one value of  $\gamma \geq 0$  such that eq. (44) is verified. But since  $D_\varphi(p||q)$  has a unique minimizer with respect to  $J$ ,  $\gamma$  must be unique. Moreover,  $J = \gamma I$  and so  $S^{1/2} = \gamma M^{1/2}$ , with  $\gamma \geq 0$ .  $\square$

### B.3 Proof for Guarantees in the Presence of Partial Even Symmetry

In this appendix, we prove Theorem 12.

*Proof.* Let  $\zeta = z - \nu$ . We differentiate  $D_\varphi(p||q_\nu)$  and proceeding as in the proof for Theorem 10, we have,

$$\begin{aligned} \nabla_{\nu_\sigma} D_\varphi(p||q_\nu) &= \int \frac{\nabla_{\zeta_\sigma} p(\nu + \zeta)}{q_0(\zeta)} \varphi' \left( \log \frac{p(\nu + \zeta)}{q_0(\zeta)} \right) q_0(\zeta) d\zeta \\ &= \int_{\bar{\sigma}} \int_{\sigma} \frac{\nabla_{\zeta_\sigma} p(\nu + \zeta)}{q_0(\zeta)} \varphi' \left( \log \frac{p(\nu + \zeta)}{q_0(\zeta)} \right) q_0(\zeta) d\zeta_\sigma d\zeta_{\bar{\sigma}}. \end{aligned} \quad (45)$$

Setting  $\nu_\sigma = \mu_\sigma$ , we obtain that  $p(\nu + \zeta)$  is even symmetric in  $\zeta_\sigma$  about the origin, for any value of  $\zeta_{\bar{\sigma}}$ . Therefore  $\nabla_{\zeta_\sigma} p(\nu + \zeta)$  is odd symmetric in  $\zeta_\sigma$  and the inner integral vanishes.

Assume now that  $\varphi$  is strictly decreasing and that  $p(z_\sigma|z_{\bar{\sigma}})$  is somewhere-strictly log concave over  $\mathbb{R}^{|\sigma|}$ . Since

$$p(z) = p(z_\sigma, z_{\bar{\sigma}}) = p(z_\sigma|z_{\bar{\sigma}})p(z_{\bar{\sigma}}), \quad (46)$$

we have that  $p(z)$  is somewhere-strictly log concave in  $z_\sigma$  over  $\mathbb{R}^{|\sigma|}$  and applying Lemma 16 from Appendix B.1, we have that  $D_\varphi(p||q_\nu)$  is strictly convex in  $\nu_\sigma$ . Hence, any stationary point  $\nu^*$  of  $D_\varphi(p||q_\nu)$  must verify  $\nu_\sigma^* = \mu_\sigma$ .  $\square$

### B.4 Proofs for Guarantees in the Presence of Partial Elliptical Symmetry

In this appendix, we provide supporting lemmas for the proof of Theorem 13 and an extension of this theorem.

We begin by rederiving a somewhat standard result of probability, which states that the covariance between two variables can be rewritten as a covariance between a variable and an expectation value.

**Lemma 17.**

$$\text{Cov}(z_{\bar{\sigma}}, z_\sigma) = \text{Cov}(z_{\bar{\sigma}}, \mathbb{E}[z_\sigma|z_{\bar{\sigma}}]). \quad (47)$$

*Proof.* We begin by applying Tower's law to the definition of the covariance,

$$\begin{aligned} \text{Cov}(z_\sigma, z_{\bar{\sigma}}) &= \mathbb{E}[(z_\sigma - \mathbb{E}[z_\sigma])(z_{\bar{\sigma}} - \mathbb{E}[z_{\bar{\sigma}}])] \\ &= \mathbb{E}[\mathbb{E}[(z_\sigma - \mathbb{E}[z_\sigma])(z_{\bar{\sigma}} - \mathbb{E}[z_{\bar{\sigma}}])|z_{\bar{\sigma}}]]. \end{aligned}$$



Notice that the second term in parenthesis does not depend on  $z_\sigma$ , and so it can be pulled out of the inner conditional expectation. Then,

$$\begin{aligned}\text{Cov}(z_\sigma, z_{\bar{\sigma}}) &= \mathbb{E}[(z_{\bar{\sigma}} - \mathbb{E}[z_{\bar{\sigma}}])\mathbb{E}[(z_\sigma - \mathbb{E}[z_\sigma])|z_{\bar{\sigma}}]] \\ &= \mathbb{E}(z_{\bar{\sigma}} - \mathbb{E}z_{\bar{\sigma}})(\mathbb{E}[z_\sigma|z_{\bar{\sigma}}] - \mathbb{E}[z_\sigma]) \\ &= \text{Cov}(\mathbb{E}(z_{\bar{\sigma}}), \mathbb{E}[z_\sigma|z_{\bar{\sigma}}]).\end{aligned}$$

□

We now derive a Lemma which characterizes the marginal and conditional mean and covariances of  $p$ , when  $p$  is elliptically symmetric.

**Lemma 18.** *Suppose  $p$  is elliptically symmetric along  $\sigma$  with a constant point of even symmetry  $m$  and a constant normalized covariance matrix  $M$ . Then  $\mu_\sigma = \mu_{\sigma|\bar{\sigma}}$ ,  $\Sigma_{\sigma\sigma} \propto \Sigma_{\sigma|\bar{\sigma}}$  and  $\Sigma_{\sigma\bar{\sigma}} = \mathbf{0}$ .*

*Proof.* We show that  $p(z_\sigma)$  and  $p(z_\sigma|z_{\bar{\sigma}})$  have matching elliptical symmetry. This follows from

$$p(z_\sigma) = \int_{\bar{\sigma}} p(z_\sigma|z_{\bar{\sigma}})q(z_{\bar{\sigma}})dz_{\bar{\sigma}}, \quad (48)$$

and the fact the elliptical symmetry of the integrand along  $\sigma$  is constant with respect to  $z_{\bar{\sigma}}$ . Since the point of even symmetry is found at the mean, we therefore have  $\mu_\sigma = \mu_{\sigma|\bar{\sigma}}$ . Similarly, from the matching elliptical symmetry, we have  $\Sigma_{\sigma\sigma} \propto \Sigma_{\sigma|\bar{\sigma}}$ .

Combining Lemma 17 and the fact  $\nu_\sigma = \nu_{\sigma|\bar{\sigma}}$ , we obtain that

$$\text{Cov}(z_\sigma, z_{\bar{\sigma}}) = \text{Cov}(\mathbb{E}[z_\sigma], z_{\bar{\sigma}}) = \text{Cov}(\nu_\sigma, z_{\bar{\sigma}}) = 0, \quad (49)$$

since  $\nu_\sigma$  does not depend on  $z_{\bar{\sigma}}$ . Therefore  $\Sigma_{\sigma\bar{\sigma}} = \mathbf{0}$ , which completes the proof. □

**Remark 19.** *Lemma 18 does not establish equality between the marginal covariance  $\Sigma_{\sigma\sigma}$  and the conditional covariance  $\Sigma_{\sigma|\bar{\sigma}}$ , rather it establishes a relationship of proportionality. Neal's funnel (eq. 6) illustrates this point: there, the covariance of  $p(\theta|\tau)$  is  $\exp(2\tau)C$  and depends on  $\tau$ , and so it cannot be equal to the marginal covariance of  $p(\theta)$ . Still, Lemma 18 teaches us the two covariances are equal up to a multiplicative scalar. We do not have an explicit expression for the covariance of  $p(\theta)$ , however, we note that  $\tau$  merely acts as a scalar multiplier on the conditional covariance.*

We end this appendix with an extension of Theorem 13, for the case where the point of even symmetry varies linearly with  $z_{\bar{\sigma}}$ .

**Theorem 20.** *Let  $\mathcal{Q}$  be the family of Gaussians. Suppose  $p$  is elliptically symmetric along  $\sigma$  about a point of even symmetry, which is linear in  $z_{\bar{\sigma}}$ , and with a constant normalized covariance matrix, and suppose  $f$  is continuously differentiable and  $|f'(0)| < \infty$ . In addition, suppose  $p(z_\sigma|z_{\bar{\sigma}})$  is somewhere-strictly log concave over  $\mathbb{R}^{|\sigma|}$ .*

*Then any minimizer of  $KL(q||p)$  must verify  $\nu_{\sigma|\bar{\sigma}} = \mu_{\sigma|\bar{\sigma}}$  and  $S_{\sigma|\bar{\sigma}} = \gamma^2 \Sigma_{\sigma|\bar{\sigma}}$  for some unique  $\gamma \in \mathbb{R}$ .*

*Proof.* The proof follows the same argument as the proof of Theorem 13. Once again, minimizing the conditional divergence,  $KL(q(z_\sigma|z_{\bar{\sigma}})||p(z_\sigma|z_{\bar{\sigma}}))$  is achieved by solving eqs. (15–16), restated here for convenience,

$$\begin{aligned}\mu_{\sigma|\bar{\sigma}} &= \nu_\sigma + S_{\sigma\bar{\sigma}}S_{\bar{\sigma}\bar{\sigma}}^{-1}(z_{\bar{\sigma}} - \nu_{\bar{\sigma}}), \\ \Sigma_{\sigma|\bar{\sigma}} &= \gamma^2(S_{\sigma\sigma} - S_{\sigma\bar{\sigma}}S_{\bar{\sigma}\bar{\sigma}}^{-1}S_{\bar{\sigma}\sigma}),\end{aligned}$$

where  $\gamma$  is uniquely defined (non-constructively) in the proof of Theorem 11. Now, the key difference with Theorem 13 is that  $\mu_{\sigma|\bar{\sigma}}$  depends linearly on  $z_{\bar{\sigma}}$ , specifically for some  $\alpha \in \mathbb{R}^{|\sigma|}$  and  $A \in \mathbb{R}^{|\sigma| \times |\bar{\sigma}|}$ ,

$$\mu_{\sigma|\bar{\sigma}} = \alpha + Az_{\bar{\sigma}}. \quad (50)$$

Matching the coefficients, we require,

$$S_{\bar{\sigma}\sigma} = AS_{\bar{\sigma}\bar{\sigma}} \quad (51)$$

$$\nu_{\sigma} = \mu_{\sigma|\bar{\sigma}} + S_{\sigma\bar{\sigma}}S_{\bar{\sigma}\bar{\sigma}}^{-1}\nu_{\bar{\sigma}} \quad (52)$$

$$S_{\sigma\sigma} = \gamma^2\Sigma_{\sigma|\bar{\sigma}} + S_{\sigma\bar{\sigma}}S_{\bar{\sigma}\bar{\sigma}}^{-1}S_{\bar{\sigma}\sigma}. \quad (53)$$

We sequentially solve eqs. (51–53) to complete the proof.  $\square$

## B.5 Analysis of the Elliptical Funnel

We now apply our theoretical results to the elliptical funnel and variations thereof. We begin with eq. (6), restated here for convenience,

$$\tau \sim \mathcal{N}(0, 1); \theta \sim \mathcal{N}(0, e^{2\tau}C).$$

The joint distribution,  $p(\tau, \theta) = p(\tau)p(\theta|\tau)$  is elliptically symmetric along  $\theta$  with a constant point of even symmetry at 0 and a constant normalized covariance matrix. It may seem surprising that the normalized covariance matrix is constant, since the covariance of  $\theta$  in  $p(\theta|\tau)$  depends on  $\tau$ . However,  $\tau$  only controls the scale of the covariances and does not alter the elliptical symmetry itself. In particular, the conditional correlation matrix is  $C$  for any value of  $\tau$ . One can further show that  $\tau$  and  $\theta$  are uncorrelated, and that the marginal correlation matrix of  $\theta$  is still  $C$ , per Lemma 17. Moreover, we have from Theorem 12 that VI provably recovers the mean of  $\theta$ . We also have from Theorem 13 that VI recovers the *marginal* correlation matrix of  $\theta$ , under certain regularity conditions.

Consider now the elliptical funnel with a varying mean  $\mu \in \mathbb{R}$ ,

$$\mu \sim \mathcal{N}(0, 1); \tau \sim \mathcal{N}(0, 1); \theta \sim \mathcal{N}(\mu, e^{2\tau}C). \quad (54)$$

$p(\mu, \tau, \theta)$  remains even symmetric along  $\theta$ , however the point of even symmetry is now given by  $\mu$  and is no longer constant. Since the point of even symmetry depends linearly on  $\mu$ , we obtain from Theorem 20 that VI recovers the *conditional* mean and correlation for  $\theta$ . We can make a stronger statement, by recognizing that  $p(\mu, \theta|\tau)$  is Gaussian and therefore  $p$  is elliptically symmetric along  $(\mu, \theta)$  with a constant point of even symmetry at 0. Then, applying Theorem 12, we have that VI recovers the *marginal* mean along  $(\mu, \theta)$ . However, the normalized covariance matrix of  $p(\mu, \theta|\tau)$  is not constant. This can be seen by examining the correlation between  $\mu$  and  $\theta$ , which depends on  $\tau$  (and goes to 1 as  $\tau \rightarrow -\infty$ ). We can therefore not apply our theoretical results and do not have guarantees on how well VI estimates the correlation matrix of  $(\mu, \theta)$ .

Finally, we consider the more general funnel from eq. (17), where  $C$  is allowed to vary with a hyperparameter  $\rho \in \mathbb{R}$ ,

$$\mu \sim \mathcal{N}(0, 1); \tau \sim \mathcal{N}(0, 1); \rho \sim p(\rho); \theta \sim \mathcal{N}(\mu, e^{2\tau}C(\rho)).$$

In this setting, we still have that  $p(\mu, \tau, \rho, \theta)$  is elliptically symmetric in  $(\mu, \theta)$  with a constant point of even symmetry, and so VI provably recovers the mean of  $\mu$  and  $\rho$  per Theorem 12. On the other hand, the normalized covariance matrix is no longer constant along  $\theta$ , let alone  $(\mu, \theta)$ , and therefore, we do not have guarantees for VI's ability to recover the conditional or marginal correlations of  $\theta$ .

## C EXPERIMENTAL DETAILS

In this appendix, we provide additional details for the numerical experiments in Section 4.

The code to reproduce all experimental results and figures in the paper is provided in the Supplemental Material. We use R as a scripting language and STAN (Carpenter et al., 2017; Stan Development Team, 2025) as a probabilistic programming language to specify models and run VI and MCMC. Our work with STAN is greatly facilitated by the packages BRIDGESTAN (Roualdes et al., 2023). All experiments are run on CPU using a 2.8 GHz Quad Core Intel Core i7 processor.

### C.1 Targets

Here, we provide details on the targets in Table 2. We specify the coordinates  $\sigma$  along which the target is even symmetric for the synthetic targets (**student-t**, **funnel**, and **crescent**) and *a priori* even symmetric for the Bayesian models (**schools**, **disease**, SKIM). Symmetry in the prior can manifest as *approximate* symmetry in the posterior.

**student-t** ( $d = 2$ ). A multivariate student-t distribution with correlation 0.5. The target is elliptically symmetric and  $\sigma = (z_1, z_2)$ .

**funnel** ( $d = 4$ ). The elliptical funnel with varying mean, specified by eq. (54). The correlation matrix  $C$  in  $p(\theta|\mu, \tau)$  has off-diagonal element  $C_{12} = 0.5$ . Here  $\sigma = (\mu, \theta)$ .

**crescent** ( $d = 3$ ). The elliptical Rosenbrock distribution is comprised of a two-dimensional Gaussian and a third coordinate whose depends quadratically on the first two components. When plotted, the joint density between the third component and any of the first two components has the shape of a crescent. In details, for  $x \in \mathbb{R}^2$  and  $y \in \mathbb{R}$ ,

$$x \sim \mathcal{N}(0, \Sigma); \quad y \sim \mathcal{N}(a(\|x\|_{\Sigma^{-1}} - b), c^2). \quad (55)$$

In this experiment, we set

$$\Sigma = 10^2 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad a = 0.03, \quad b = 100, \quad c = 0.02. \quad (56)$$

This distribution is even and elliptically symmetric along its first two coordinates,  $\sigma = x$ . This is an extension of the two-dimensional Rosenbrock distribution by Roberts et al. (1997). The two-dimensional Rosenbrock distribution is even-symmetric along its first coordinate. We add an additional dimension in order to obtain a non-trivial partial elliptical symmetry.

**schools** ( $d = 10$ ). A Bayesian hierarchical model of the effects of a preparation program for a standardized test across  $N = 8$  schools (Rubin, 1981; Gelman et al., 2013). We observe  $y_i$ , the average change in test scores, and  $\eta_i$ , the empirical standard deviation across students, for each school. The model is then

$$\mu \sim \mathcal{N}(5, 3^2); \quad \tau \sim \mathcal{N}^+(0, 5^2); \quad \theta_i \sim \mathcal{N}(\mu, \tau^2); \quad y_i \sim \mathcal{N}(\theta_i, \eta_i^2). \quad (57)$$

The prior  $p(\mu, \tau, \theta)$  is even symmetric along  $\sigma = (\mu, \theta)$ . We can implement this model either using the standard (centered) parameterization (eq. 57), the non-centered parameterization described in Section 4.1, or by marginalizing out  $\theta$ —here he exploit the fact the prior  $p(\theta|\mu, \tau)$  and likelihood  $p(y_i|\theta_i)$  are Gaussians, and so,

$$p(y_i|\mu, \tau) = \mathcal{N}(0, \sigma_i^2 + \tau^2); \quad p(\theta_i|y_i, \mu, \tau) = \mathcal{N}\left(\frac{y_i/\eta_i^2 + \mu/\tau^2}{1/\eta_i^2 + 1/\tau^2}, \frac{1}{1/\eta_i^2 + 1/\tau^2}\right). \quad (58)$$

**disease** ( $d = 102$ ). A model for the mortality counts across counties in Finland, due to alcoholism (Vanhatalo et al., 2010). For each county, we observe the mortality count,  $y_i$ , the standardized expected number of deaths,  $y_{e,i}$ , and the two-dimensional location of the country,  $x_i$ . The original model considers 911 counties, however we consider a random subset of 100 counties to reduce the computational cost of the experiment. The model uses a Gaussian process prior with a squared exponential kernel. Specifically, the prior covariance matrix  $K$  is defined by,

$$K_{ij} = \alpha^2 \exp\left(-\frac{\|x_i - x_j\|^2}{\rho^2}\right), \quad (59)$$

and the full model is,

$$\begin{aligned} \rho &\sim \text{inv-Gamma}(2.42, 14.81); \quad \alpha \sim \text{inv-Gamma}(10, 10); \\ \theta &\sim \mathcal{N}(0, K(\alpha, \rho, x)); \quad y_i \sim \text{Poisson}(y_{e,i} \exp(\theta_i)). \end{aligned} \quad (60)$$

The prior is even symmetric along  $\sigma = \theta$ . As before, this model can be implemented using a centered or non-centered parameterization. Exact marginalization is not possible, but can be achieved using an integrated Laplace approximation.

**SKIM** ( $d = 305$ ). A sparse kernel interaction model (SKIM) (Agrawal et al., 2019). This model is a regularized regression model that accounts for interaction effect between covariates. Covariates are probabilistically selected using a horseshoe prior (Piironen and Vehtari, 2017). As in Margossian et al. (2020), we apply the model to a genetic microarray classification data set on prostate cancer. We observe  $N = 102$  patients with  $p = 200$  pre-selected genetic covariates (out of a total of 5966 covariates) and denote  $X \in \mathbb{R}^{N \times p}$  the design matrix. We observe for each patient  $y_i$ , a binary variable that indicates whether the patient has cancer.

To specify the Bayesian model, we first set the following hyperparameters:

$$p_0 = 5; s_{\text{global}} = \frac{p_0}{\sqrt{N}(p - p_0)}; \nu_{\text{local}} = 1; \nu_{\text{global}} = 1; s_{\text{slab}} = 2 \quad s_{\text{df}} = 100 \quad c_0 = 5. \quad (61)$$

Then, a standard parameterization of the model is,

$$\begin{aligned} \lambda_i &\sim \text{Student}_t(\nu_{\text{local}}, 0, 1); \tau \sim \text{Student}_t(\nu_{\text{global}}, 0, s_{\text{global}}); c_{\text{aux}} \sim \text{invGamma}(s_{\text{df}}/2, s_{\text{df}}/2); \\ \chi &\sim \text{InverseGamma}(s_{\text{df}}/2, s_{\text{df}}/2); c = \sqrt{c_{\text{aux}} s_{\text{slab}}}; \tilde{\lambda}_i^2 = \frac{c^2 \lambda_i^2}{c^2 + \tau^2 \lambda_i^2}; \eta_2 = \tau^2 \chi / c^2 \\ \beta_0 &\sim \mathcal{N}(0, c_0^2); \beta_i \sim \mathcal{N}(0, \tau^2 \tilde{\lambda}_i^2); \beta_{ij} \sim \mathcal{N}(0, \eta_2^2 \tilde{\lambda}_i^2 \tilde{\lambda}_j^2); y \sim \text{Bernoulli}(\text{logit}^{-1}(\beta_0 + X\beta)). \end{aligned} \quad (62)$$

Following Agrawal et al. (2019), we marginalize out  $\beta_i$  and  $\beta_{ij}$  using a Gaussian process reparameterization. To define the Gaussian process' covariance matrix  $K$ , we first introduce the matrices:

$$\begin{aligned} K_1 &= X \text{diag}(\tilde{\lambda}^2) X^T \\ K_2 &= [X \circ X] \text{diag}(\tilde{\lambda}^2) [X \circ X]^T, \end{aligned} \quad (63)$$

where “ $\circ$ ” denotes the element-wise Hadamard product. Finally,

$$K = \frac{1}{2} \eta_2^2 (K_1 + 1) \circ (K_1 + 1) - \frac{1}{2} \eta_2^2 K_2 - (\tau^2 - \eta_2^2) K_1 + c_0^2 - \frac{1}{2} \eta_2^2. \quad (64)$$

Then, the the Gaussian process prior and the likelihood are

$$f \sim \mathcal{N}(0, K); y \sim \text{Bernoulli}(\text{logit}^{-1}(f)). \quad (65)$$

Once again, this model admits three implementations: a centered parameterization, a non-centered parameterization, and an implementation where  $f$  is approximately marginalized out with an integrated Laplace approximation.

## C.2 VI algorithm

For the experiment in Section 4, we specify targets in STAN and fit them with ADVI (Kucukelbir et al., 2017). ADVI employs a Gaussian approximation over the unconstrained scale. Constrained variables are automatically transformed to an unconstrained scale by STAN. For example, a variable  $z \in \mathbb{R}^+$  is replaced by  $\log z \in \mathbb{R}$ . In our experiments, we report estimates of the mean on the unconstrained scale, since our theoretical analysis applies to variables defined over  $\mathbb{R}$ . This choice allows us to test how predictive/illustrative our theory is, however practitioners may be more interested in estimates of the mean over the original scale.

ADVI minimizes  $\text{KL}(q||p)$  via stochastic optimization. We warm-start VI using a factorized (mean-field) approximation, then switch to a Gaussian with a full covariance matrix. We use a large batch size ( $B \geq 50$ ) to better estimate the ELBO and its gradient, and improve our chances of finding an optimal solution. All other tuning parameters use the default options in STAN.

## C.3 Additional results on correlation

Figure 5 plots the error in estimates of the correlations across the models in Table 2. As before, the models are ordered according to their even asymmetry (eq. 19). In the synthetic examples, we find that more symmetric targets yield better estimates of the correlation. Furthermore, better estimates of the correlations are obtained along symmetric coordinates, when the target is partially symmetric.



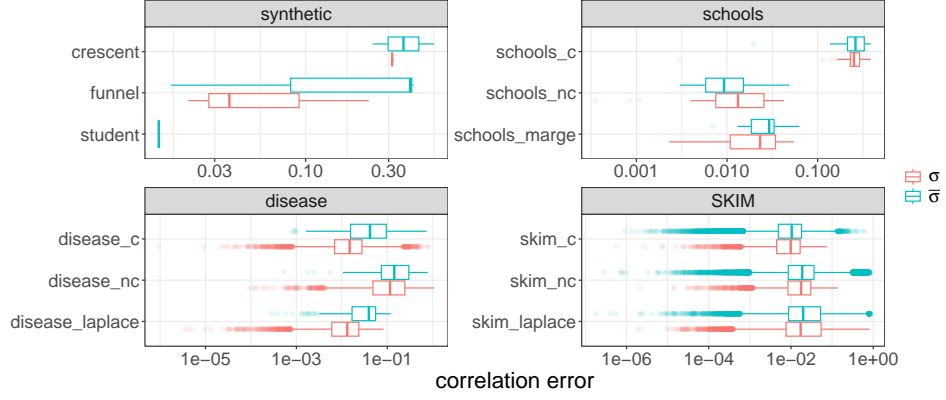


Figure 5: *Error in VI estimates of the correlation.* We split the targets into four groups: *synthetic* targets and implementations of *schools*, *disease*, and *SKIM*. Within each panel, the models are ordered bottom to top from most symmetric to least symmetric according to eq. (19). For the synthetic targets, we obtain better estimates of the correlation for more symmetric targets. There is no clear pattern for other targets.

These patterns are not clear in the non-synthetic targets. In particular, for **disease** and **SKIM**, we see no difference in the quality of correlation estimates between implementations with varying degrees of symmetry, and between *a priori* symmetric and asymmetric coordinates. We suspect this is because the correlation matrices for these models are sparse. Overall, the error in estimates of the correlations tends to be small.