Journal of Machine Learning Research 26 (2025) 1-41 Submitted 6/24; Revised 5/25; Published 9/25

Variational Inference for Uncertainty Quantification: an Analysis of Trade-offs

Charles C. Margossian CHARLES.MARGOSSIAN@UBC.CA
Department of Statistics
University of British Columbia
Vancouver, BC, Canada

Loucas Pillaud-Vivien LOUCAS.PILLAUD-VIVIEN@ENPC.FR
CERMICS Laboratory
Ecole des Ponts ParisTech
Champs-sur-Marne, France

Lawrence K. Saul LSAUL@FLATIRONINSTITUTE.ORG
Center for Computational Mathematics
Flatiron Institute
New York, NY, USA

Editor: Barbara Engelhardt

Abstract

Given an intractable distribution p, the problem of variational inference (VI) is to find the best approximation from some more tractable family $\mathcal{Q}$. Commonly, one chooses $\mathcal{Q}$ to be a family of factorized distributions (i.e., the mean-field assumption), even though p itself does not factorize. We show that this mismatch leads to an *impossibility theorem*: if p does not factorize, then any factorized approximation $q \in \mathcal{Q}$ can correctly estimate at most one of the following three measures of uncertainty: (i) the marginal variances, (ii) the marginal precisions, or (iii) the generalized variance (which for elliptical distributions is closely related to the entropy). In practice, the best variational approximation in $\mathcal{Q}$ is found by minimizing some divergence $D(q, p)$ between distributions, and so we ask: how does the choice of divergence determine which measure of uncertainty, if any, is correctly estimated by VI? We consider the classic Kullback-Leibler divergences, the more general α-divergences, and a score-based divergence which compares $\nabla \log p$ and $\nabla \log q$. We thoroughly analyze the case where p is a Gaussian and q is a (factorized) Gaussian. In this setting, we show that all the considered divergences can be *ordered* based on the estimates of uncertainty they yield as objective functions for VI. Finally, we empirically evaluate the validity of this ordering when the target distribution p is not Gaussian.

Keywords: variational inference, Kullback-Leibler divergence, α-divergence, Fisher divergence, score matching

1. Introduction

In many problems, it is useful to approximate an intractable distribution, p, by a more tractable distribution, q. This problem can be posed as an optimization, one whose goal is to discover the best-matching distribution q within some larger, parameterized family $\mathcal{Q}$ of

approximating distributions. Optimizations of this sort are the subject of a large literature on variational inference (VI; Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017).

Formally, suppose we wish to approximate a *target* distribution p over latent variables $\mathbf{z} = z_{1:n} \in \mathcal{Z}$, and let $\mathcal{Q}$ be the family of approximating distributions. VI attempts to solve an optimization of the form

$$q^* = \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\ D(q, p), \tag{1}$$

where D is a *divergence* satisfying (i) $D(q, p) \geq 0$ for all $q \in \mathcal{Q}$ and (ii) $D(q, p) = 0$ if and only if $p = q$. In most applications, $\mathcal{Q}$ is not sufficiently rich to capture all the features of p—that is $p \notin \mathcal{Q}$—and so, even after minimizing $D(q, p)$, the optimal approximation q must be compromised in some way. A crucial question is then whether q can still capture the features of p that are important for a given application, even if other (less relevant) features are poorly estimated. In this paper, we study certain trade-offs that arise from the choice of $\mathcal{Q}$, and we analyze how these trade-offs are resolved by the particular choice of divergence in eq. (1).

Perhaps the most common choice for $\mathcal{Q}$ is the family of *factorized* approximations,

$$q(\mathbf{z}) = \prod_{i=1}^{n} q_i(z_i). \tag{2}$$

We refer to this setting as F-VI. Variational families of this form originated in the mean-field approximation for physical models of ferromagnetism (Parisi, 1988). Factorized approximations are computationally convenient for many applications of machine learning because the number of variational parameters scales linearly with the dimensionality of $\mathbf{z}$ (e.g Bishop et al., 2002; Blei, 2012; Giordano et al., 2023); in addition better scaling can be achieved for large data sets via amortization, which further restricts the form of $\mathcal{Q}$ (Kingma and Welling, 2014; Margossian and Blei, 2024).

Our paper begins by formalizing the inherent trade-offs that arise in F-VI. Consider a target p which does *not* factorize but admits a finite covariance matrix $\mathbf{\Sigma}$. By its very nature, the factorized approximation in eq. (2) cannot estimate the correlations between different components of $\mathbf{z}$. Nonetheless, a factorized approximation may still capture useful features of p, such as its mean and some quantification of uncertainty. There are many ways to characterize the uncertainty of a multivariate distribution p. In this paper, we specifically consider:

(i) the marginal variances, Σ_{ii}, equal to the diagonal elements of the covariance matrix,

(ii) the marginal precisions, Σ_{ii}^{-1}, equal to the diagonal elements[1] of the precision matrix (i.e., the inverse covariance matrix), and

(iii) the generalized variance, $|\mathbf{\Sigma}|$, defined as the determinant of the covariance matrix.

Which measure of uncertainty to pursue depends on the application. It is especially common in Bayesian statistics to report the marginal variances of interpretable variables (Gelman

1. Throughout the paper, for any square matrix $\mathbf{A}$, we use A^p_{ij} to denote $(A^p)_{ij}$, the ij^{th} element of the p^{th} matrix power of A, and we use $(A_{ij})^p$ to denote the p^{th} scalar power of the matrix element A_{ij}.

Divergence	Notation	Definition
(Reverse) Kullback-Leibler	$\mathrm{KL}(q\|\|p)$	$\mathbb{E}_q\left[\log \frac{q}{p}\right]$
(Forward) Kullback-Leibler	$\mathrm{KL}(p\|\|q)$	$\mathbb{E}_p\left[\log \frac{p}{q}\right]$
α-family with $\alpha \in \mathbb{R}^+\backslash\{0,1\}$	$\mathrm{D}_\alpha(p\|\|q)$	$\frac{1}{\alpha(\alpha-1)}\,\mathbb{E}_q\left[\frac{p^\alpha}{q^\alpha} - 1\right]$
(Reverse) Score-based	$\mathrm{S}(q\|\|p)$	$\mathbb{E}_q\left[\left\|\nabla \log \frac{q}{p}\right\|^2_{\mathrm{Cov}(q)}\right]$
(Forward) Score-based	$\mathrm{S}(p\|\|q)$	$\mathbb{E}_p\left[\left\|\nabla \log \frac{q}{p}\right\|^2_{\mathrm{Cov}(p)}\right]$

Table 1: Divergences that we analyze in this paper for FG-VI. The score-based divergences (Cai et al., 2024b) are defined here for the special case that the base distribution is Gaussian.

et al., 2013), and in some models, the marginal precisions (Bernardo and Smith, 2000). On the other hand, sometimes it is useful to report a single scalar measure of uncertainty; this is provided by the generalized variance, a measure of uncertainty that for elliptical distributions is closely related to the differential entropy in information theory (MacKay, 2003).

The first main result of this paper is an impossibility theorem: it states that any factorized approximation, of the form in eq. (2), can match *at most one* of the above three measures of uncertainty. An immediate consequence is that there exists no universally "best" factorized approximation of p; what is best for one application may be of little use for another.

This ambiguity is reflected by the fact that different divergences return different solutions to eq. (1). This, in turn, begs the following question: which divergence is best suited for a particular application? By far the most popular choice in practice is the Kullback-Leibler divergence, $\mathrm{KL}(q||p)$, which measures the disagreement between $\log p$ and $\log q$. Other valid choices include the "forward" $\mathrm{KL}(p||q)$ (e.g., Naesseth et al., 2020; Vehtari et al., 2020), the α-divergences $\mathrm{D}_\alpha(p||q)$ (Minka, 2005; Li and Turner, 2016) which interpolate between $\mathrm{KL}(q||p)$ and $\mathrm{KL}(p||q)$ when $\alpha \in (0,1)$, and score-based (or Fisher) divergences (Courtade, 2016; Cai et al., 2024b,a; Modi et al., 2025), which measure the discrepancy between $\nabla \log p$ and $\nabla \log q$. Table 1 lists the divergences that we consider in this paper.

We investigate how each divergence in Table 1 resolves the constraints imposed by the impossibility theorem. To make progress, we identify a particular setting of VI where these questions can be fully analyzed. This is the setting where q belongs to a the family of Gaussians with a diagonal covariance matrix—a common choice in black box VI (Kucukelbir et al., 2017), which we call FG-VI—and where p is a Gaussian with a non-diagonal covariance matrix:

$$p = \mathrm{normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

$$q = \mathrm{normal}(\boldsymbol{\nu}, \boldsymbol{\Psi}), \tag{4}$$

with $\Psi_{ij} = 0$ whenever $i \neq j$. This setting, though relatively simple, is already rich enough to provide several counter-intuitive results, and it reveals in a stark way that different inferential goals are best served by different divergences.

Our work in this setting builds on earlier analyses of F-VI based on minimizing KL-divergences (Margossian and Saul, 2023). In particular, we also analyze F-VI when q is chosen to minimize the α-divergences or score-based divergences in Table 1; this requires more technical machinery because there do not exist closed-form solutions (even in this simple setting) for the variational approximation that minimizes these divergences. Thus we fill a gap in the VI literature for these non-classical divergences by further characterizing their variational approximations.

The second main result of the paper is an *ordering* of all the divergences in Table 1 based on the estimates they yield of the marginal variances, marginal precisions, and generalized variance when p and q are Gaussian. This ordering provides a natural and interpretable way of comparing divergences based on a downstream inferential task. While there exist many ways to compare divergences (e.g., Gibbs and Su, 2002), we believe that this ordering should be particularly helpful to practitioners. The ordering not only illuminates which divergence may yield the best estimate of some measure of uncertainty; it also cautions us that certain divergences can fail spectacularly—for instance, producing solutions for F-VI with variances that collapse to 0 or blow up to infinity.

Outline. In section 2, we state the main results of the paper—the impossibility theorem, and the ordering theorem—and we show how and where they fit into the existing literature. In section 3, we prove the impossibility theorem by analyzing the pairwise trade-offs that arise in F-VI between different measures of uncertainty. In section 4, we calculate the divergences in Table 1 and characterize the solutions to eq. (1) when p and q are Gaussian, as in eqs. (3–4). In section 5, we prove the ordering theorem by comparing the estimates of uncertainty obtained by minimizing these divergences. In section 6, we empirically investigate the validity of this ordering, on a variety of targets and data sets, when q is Gaussian but p is not. Finally, in section 7, we discuss our results in light of various applications which require uncertainty quantification (e.g., Bayesian statistics, pre-conditioning of Markov chain Monte Carlo, and information theory), and we explore directions for future work.

2. Summary of Contributions

The main results of the paper are an impossibility theorem for F-VI, starting from eq. (2), and an ordering theorem for the divergences in Table 1, starting from eqs. (3–4).

2.1 Main Results

We begin with the impossibility theorem. To state the result, the following notation is useful. For any square matrix, we use $\mathrm{diag}(\cdot)$ to denote the vector formed from its diagonal elements, and we use vector inequalities, such as $\mathrm{diag}(\mathbf{\Psi}) \leq \mathrm{diag}(\mathbf{\Sigma})$, to denote that $\Psi_{ii} \leq \Sigma_{ii}$ for all i. Furthermore, we use denote A^{-1}_{ij} to denote the $(ij)^{\text{th}}$ element of $\mathbf{A}^{-1}$. Note that for a non-diagonal matrix, $A^{-1}_{ij} \neq 1/A_{ij}$.

Theorem 1 (Impossibility theorem for F-VI) *Let p and q be distributions with covariances $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, respectively, where $\boldsymbol{\Psi}$ is diagonal but $\boldsymbol{\Sigma}$ is not. Then*

1. *(Variance matching)*
   If $\mathrm{diag}(\boldsymbol{\Psi}) = \mathrm{diag}(\boldsymbol{\Sigma})$, then $|\boldsymbol{\Psi}| > |\boldsymbol{\Sigma}|$ and $\mathrm{diag}(\boldsymbol{\Psi}^{-1}) \leq \mathrm{diag}(\boldsymbol{\Sigma}^{-1})$, and this last inequality is strict for at least one component along the diagonal.

2. *(Precision matching)*
   If $\mathrm{diag}(\boldsymbol{\Psi}^{-1}) = \mathrm{diag}(\boldsymbol{\Sigma}^{-1})$, then $|\boldsymbol{\Psi}| < |\boldsymbol{\Sigma}|$ and $\mathrm{diag}(\boldsymbol{\Psi}) \leq \mathrm{diag}(\boldsymbol{\Sigma})$, and this last inequality is strict for at least one component along the diagonal.

3. *(Generalized variance matching)*
   If $|\boldsymbol{\Psi}| = |\boldsymbol{\Sigma}|$, then $\Psi_{ii} < \Sigma_{ii}$ and $\Psi_{jj}^{-1} < \Sigma_{jj}^{-1}$ for at least one i and j.

The theorem shows, for instance, that a factorized approximation with the correct marginal variances will overestimate the generalized variance and underestimate at least one of the marginal precisions. It also highlights similar trade-offs when the factorized approximation matches the marginal precisions or the generalized variance.

Our goal is to understand how different choices of divergences resolve the trade-offs imposed by Theorem 1. To do so, we consider the case where p and q are Gaussian, as in eqs. (3–4). Note that in this setting p and q have the same generalized variance if and only if they have the same entropy. All of the divergences in Table 1 yield a correct estimate for the mean of p, but they yield different variational approximations for the diagonal covariance matrix $\boldsymbol{\Psi}$, and therefore different estimates of the marginal variances, marginal precisions, and entropy of p. Moving forward, we say that one divergence *dominates* another if it always returns a larger estimate of the marginal variances of p.

Definition 2 (Ordering of divergences) *Let $\mathcal{P}$ and $\mathcal{Q}$ denote the families of multivariate Gaussian distributions in eqs. (3–4), where $\boldsymbol{\Psi}$ is diagonal but $\boldsymbol{\Sigma}$ is not. Consider two divergences D_a and D_b, and for any $p \in \mathcal{P}$, let $\boldsymbol{\Psi}(a)$ and $\boldsymbol{\Psi}(b)$ denote the covariances of the factorized approximations obtained, respectively, by minimizing $D_a(q, p)$ and $D_b(q, p)$ over all $q \in \mathcal{Q}$. We say D_a dominates D_b, written*

$$D_a \succ D_b,$$

if for any $p \in \mathcal{P}$ we have that $\mathrm{diag}(\boldsymbol{\Psi}(a)) \geq \mathrm{diag}(\boldsymbol{\Psi}(b))$ and also that $\Psi(a)_{ii} > \Psi(b)_{ii}$ for some element along the diagonal of these matrices.

The above definition[2] is based on an ordering of marginal variances, but it is also possible (because $\boldsymbol{\Psi}$ is diagonal) to define an equivalent ordering based on the marginal precisions. These orderings also imply an ordering of the generalized variances, though the converse is not true.

We now state the second main result of the paper, which is an ordering over the divergences listed in Table 1.

2. This definition can be further simplified if we impose the additional constraint that every row and column of $\boldsymbol{\Sigma}$ has nonzero off-diagonal elements. In this case, we can adopt the simpler definition that $\mathcal{D}_a \succ \mathcal{D}_b$ if $\boldsymbol{\Psi}_a \succ \boldsymbol{\Psi}_b$, and all subsequent results hold with this more stringent characterization.

Theorem 3 (Ordering theorem for FG-VI) *The divergences in Table 1 are ordered in the sense given above. In particular, for any* (α_1, α_2) *satisfying* $0 < \alpha_1 < \alpha_2 < 1$,

$$S(q||p) \prec KL(q||p) \prec D_{\alpha_1}(p||q) \prec D_{\alpha_2}(p||q) \prec KL(p||q) \prec S(p||q). \tag{5}$$

The theorem shows, for instance, that the (reverse) score-based divergence yields smaller estimates of marginal variance than the (reverse) KL divergence, which in turn yields smaller estimates than all of the α-divergences. We prove Theorem 3 in Section 5. The main challenge here arises from the fact that the variational approximations from the score-based and α-divergences cannot be computed in closed form, and thus it is necessary (as done in Section 4) to characterize these solutions in other ways that can be further analyzed.

The proof of Theorem 3 also yields several intermediate results of interest. Most notably, we obtain the following positive results for VI-based uncertainty quantification:

(i) q matches the marginal variances of p when minimizing $\mathrm{KL}(p||q)$ (MacKay, 2003);
(ii) q matches the marginal precisions of p when minimizing $\mathrm{KL}(q||p)$ (Turner and Sahani, 2011);
(iii) q matches the generalized variance (or entropy) of p when minimizing $\mathrm{D}_\alpha(p||q)$, for some $\alpha \in (0, 1)$, which depends on $\mathbf{\Sigma}$.

We also obtain certain cautionary results. For example, we show that q can wildly misestimate uncertainty when minimizing a score-based divergence; in fact, the estimated variance may collapse to 0 when optimizing $S(q||p)$ or blow up to infinity when optimizing $S(p||q)$. We refer to such cases as instances of *variational collapse*, where the variational approximation is not well-defined because $\operatorname{arginf}_{q \in \mathcal{Q}} D(q, p) \notin \mathcal{Q}$.

We extend some of our results for α-divergences to the case where $\alpha > 1$. In this regime, we prove the following: (i) $\mathrm{D}_\alpha(p||q) \succ \mathrm{KL}(p||q)$, (ii) there does *not* exist an ordering of $\mathrm{D}_\alpha(p||q)$ and $S(p||q)$, and (iii) $\mathrm{D}_\alpha(p||q)$ yields strictly positive, bounded estimates of the variance and so (unlike the score-based divergences) it does not suffer from variational collapse. We are *not* able to show for $\alpha > 1$ that the α-divergences are ordered amongst themselves, satisfying $\mathrm{D}_{\alpha_2}(p||q) \succ \mathrm{D}_{\alpha_1}(p||q)$ when $\alpha_2 > \alpha_1 > 1$. As we discuss later, there are technical reasons why our proof for $\alpha \in (0, 1)$ does not extend to this case where $\alpha > 1$. Based on our numerical experiments, however, we conjecture that the α-divergences are indeed ordered amongst themselves for $\alpha > 1$, and we leave the proof (or disproof) of this conjecture for future work.

For convenience, Table 2 provides a summary of the notation used in the paper.

2.2 Related Work

Our work is in line with several efforts to evaluate VI through the lens of downstream inferential tasks. The mean and variances of p are of most interest for Bayesian statistics, but Huggins et al. (2020) caution that a low $\mathrm{KL}(q||p)$ divergence does not guarantee that q accurately estimates these quantities. Instead they recommend to estimate the Wasserstein distance, which can be used to bound the difference in moments of q and p; see also Biswas and Mackey (2024). In latent variable models, VI is used to bound a marginal likelihood which is then maximized with respect to the model parameters, e.g the weights of a neural

Notation	Definition
F-VI	Factorized Variational Inference
FG-VI	Factorized Gaussian Variational Inference
diag($\mathbf{A}$)	The vector formed from the diagonal of the squared matrix $\mathbf{A}$.
$\mathbf{v} \geq \mathbf{u}$	For two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $v_i \geq u_i, \forall i$.
A_{ij}^{-1}	The $(ij)^{\text{th}}$ element of $\mathbf{A}^{-1}$. In general $A_{ij}^{-1} \neq 1/A_{ij}$.
$\mathbf{A} \succ \mathbf{B}$	The matrix $\mathbf{A}-\mathbf{B}$ is positive definite.
$\|\|\mathbf{v}\|\|_{\mathbf{A}}^2$	For $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{n \times n}$, the weighted inner-product $\mathbf{v}^T\mathbf{A}\mathbf{v}$.
$D_a \succ D_b$	Divergence D_a *dominates* divergence D_b (Definition 2).
$\mathbf{\Sigma}$	The covariance matrix of the target p (eq. 3).
$\mathbf{\Psi}$	The covariance matrix of the approximation q (eq. 4).

Table 2: Summary of notation. For the notation on divergences, see Table 1.

network (Tomczak, 2022). Several works examine the tightness of this bound (e.g Li and Turner, 2016; Dieng et al., 2017; Daudel et al., 2023) and the statistical properties of the resulting maximum likelihood estimator (Wang and Blei, 2018). VI is also used in tandem with other inference schemes—for example, to initialize Markov chain Monte Carlo (Zhang et al., 2022) or to estimate a proposal distribution for importance sampling. For the latter, the quality of VI can be assessed from the distribution of importance weights (Yao et al., 2018; Vehtari et al., 2024). In sum, there are many different ways to evaluate VI, and one lesson of our paper is that different inferential goals may be incompatible—may, in fact, compete against one another.

Many researchers have studied, both theoretically and empirically, how well F-VI estimates uncertainty (and, in particular, the marginal variances) when minimizing KL($q||p$) (e.g., MacKay, 2003; Wang and Titterington, 2005; Bishop, 2006; Turner and Sahani, 2011; Giordano et al., 2018). Several works also examine the more specific setting in Section 4 where p and q are Gaussian with (respectively) a non-diagonal and diagonal covariance matrix. MacKay (2003) reports that q matches the marginal variances when minimizing KL($p||q$), while Turner and Sahani (2011) find that q matches the marginal precisions when minimizing KL($q||p$). In a precursor to this paper, Margossian and Saul (2023) prove that a precision-matching solution underestimates both the entropy and the marginal variances; here we provide a new and more intuitive proof of this result. We also note that Theorem 1 provides a more general framework for understanding these sorts of trade-offs, as it assumes neither that p and q are Gaussian nor that q is obtained by minimizing a particular divergence, such as KL($q||p$).

Previous studies in VI have investigated a large class of divergences for the optimization in eq. (1). A number of authors have considered the α-divergence (Li and Turner, 2016; Dieng et al., 2017; Daudel et al., 2021), for instance, as it provides a non-trivial general-

ization of the KL divergence. More recently, Cai et al. (2024b) introduced the score-based divergence $S(q||p)$ in Table 1 and showed that it can be efficiently optimized for Gaussian variational families; see also Modi et al. (2025). It is known that the α-divergence reduces in certain limits to the KL divergence (either from q to p, or vice versa); however, there is no such correspondence for the score-based divergences, and thus it is not obvious how the variational approximations obtained by minimizing $S(q||p)$ or $S(p||q)$ might compare to those obtained by minimizing $\mathrm{KL}(q||p)$ or $\mathrm{KL}(p||q)$. Our analysis, notably the ordering in Theorem 3, sheds light on the matter.

Finally, we highlight a recent contribution by Chen et al. (2025) who build on the ideas in this paper to analyze the weighted Fisher divergence, defined as

$$S_M(q||p) = \mathbb{E}_q \left[\nabla \left\| \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \right\|_{\mathbf{M}}^2 \right] \tag{6}$$

for $\mathbf{M} \succ 0$. If $\mathbf{M} = \mathbf{\Psi}$, then this divergence reduces to the reverse score-based divergence $S(q||p)$; on the other hand, if $\mathbf{M} = \mathbf{I}$ (the identity matrix), then $S_M(q||p)$ reduces to the Fisher divergence, which we denote by $F(q||p)$. Chen et al. (2025) notably show that (i) $F(q||p) \prec \mathrm{KL}(q||p)$, (ii) $F(q||p)$ and $S(q||p)$ cannot be ordered, and (iii) unlike the score-based divergence $S(q||p)$, the Fisher divergence $F(q||p)$ is not subject to variational collapse.

3. Proof of Impossibility Theorem

We begin by restating the impossibility theorem of the previous section as a collection of trade-offs.

Theorem 4 (Restatement of impossibility theorem) *Let p and q be distributions over $\mathbb{R}^N$ with covariances $\mathbf{\Sigma}$ and $\mathbf{\Psi}$, respectively, where $\mathbf{\Psi}$ is diagonal but $\mathbf{\Sigma}$ is not.*

1. ***Trade-off between marginal variances and generalized variance:***
   If $\mathrm{diag}(\mathbf{\Psi}) = \mathrm{diag}(\mathbf{\Sigma})$, then $|\mathbf{\Psi}| > |\mathbf{\Sigma}|$, and if $|\mathbf{\Psi}| = |\mathbf{\Sigma}|$, then $\Psi_{ii} < \Sigma_{ii}$ for some i.

2. ***Trade-off between marginal precisions and generalized variance:***
   If $\mathrm{diag}(\mathbf{\Psi}^{-1}) = \mathrm{diag}(\mathbf{\Sigma}^{-1})$, then $|\mathbf{\Psi}| < |\mathbf{\Sigma}|$, and if $|\mathbf{\Psi}| = |\mathbf{\Sigma}|$, then $\Psi_{ii}^{-1} < \Sigma_{ii}^{-1}$ for some i.

3. ***Variance-precision trade-off:***
   If $\mathrm{diag}(\mathbf{\Psi}) = \mathrm{diag}(\mathbf{\Sigma})$, then $\mathrm{diag}(\mathbf{\Psi}^{-1}) \leq \mathrm{diag}(\mathbf{\Sigma}^{-1})$ and $\Psi_{ii}^{-1} < \Sigma_{ii}^{-1}$ for some i.
   If $\mathrm{diag}(\mathbf{\Psi}^{-1}) = \mathrm{diag}(\mathbf{\Sigma}^{-1})$, then $\mathrm{diag}(\mathbf{\Psi}) \leq \mathrm{diag}(\mathbf{\Sigma})$ and $\Psi_{ii} < \Sigma_{ii}$ for some i.

The above restatement makes exactly the same claims as Theorem 1, but the proof is easier to organize along these lines.

We emphasize that the trade-offs listed above are not tied to any particular choice of the divergence $D(q, p)$ in eq. (1). Rather, these trade-offs are inherent to the use of a factorized approximation, and they arise regardless of how q is chosen. Indeed, we will prove each trade-off by appealing only to the properties of $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ as positive-definite matrices. However, for each trade-off, we will also provide additional intuitions (and in one case, an additional proof) that reflect the statistical setting in which these matrices arise.

Proof of trade-off between marginal variances and generalized variance. We begin by defining two auxiliary matrices whose elements are dimensionless (unlike those of $\mathbf{\Sigma}$ and $\mathbf{\Psi}$). The first of these is the *correlation* matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$ of the target distribution p, with elements

$$C_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}, \tag{7}$$

which is positive-definite and has unit diagonal entries, $C_{ii} = 1$. The second is the diagonal matrix $\mathbf{R} \in \mathbb{R}^{n \times n}$ that records the *ratios* of marginal variances in q and p; i.e.,

$$R_{ii} = \frac{\Psi_{ii}}{\Sigma_{ii}}. \tag{8}$$

In what follows, it will be convenient to consider the generalized variances on a log scale. Taking differences, we have

$$\log|\mathbf{\Sigma}| - \log|\mathbf{\Psi}| = \log\left|\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{\Psi}^{-\frac{1}{2}}\right| = \log\left|\mathbf{R}^{-\frac{1}{2}}\mathbf{C}\mathbf{R}^{-\frac{1}{2}}\right| = \log|\mathbf{R}|^{-1} - \tfrac{1}{2}\log|\mathbf{C}|^{-1}. \tag{9}$$

The first determinant on the right side of eq. (9) is easy to compute because the matrix $\mathbf{R}$, as defined in eq. (8), is diagonal. We can bound the second determinant by the Hadamard inequality (Horn and Johnson, 2012, Theorem 7.8.1), which states that

$$|\mathbf{C}| < \prod_{i=1}^{n} C_{ii} = 1. \tag{10}$$

The inequality in eq. (10) is strict under the assumption that $\mathbf{\Sigma}$ (and hence also $\mathbf{C}$) is not diagonal. Combining the last two results, we see that

$$\log|\mathbf{\Sigma}| - \log|\mathbf{\Psi}| \;<\; \sum_i \log\left(\frac{\Sigma_{ii}}{\Psi_{ii}}\right). \tag{11}$$

Now suppose that $\mathrm{diag}(\mathbf{\Psi}) = \mathrm{diag}(\mathbf{\Sigma})$; then from eq. (11) we see that $|\mathbf{\Sigma}| < |\mathbf{\Psi}|$, proving the first part of the trade-off. Conversely, suppose that $|\mathbf{\Sigma}| = |\mathbf{\Psi}|$; then from eq. (11) we see that $\sum_i \log \Sigma_{ii} > \sum_i \log \Psi_{ii}$, and this can only be true if $\Sigma_{ii} > \Psi_{ii}$ for some i. This proves the second part of the trade-off. ■

The above trade-off has a particularly interesting interpretation in the case where p and q are Gaussian, as in eqs. (3–4). In this case, the generalized variance of q, given by $|\mathbf{\Psi}|$, provides effectively the same measure of uncertainty as the *entropy* of q, which is equal to $\frac{1}{2}\log|\mathbf{\Psi}|$ plus an additive constant. Thus when q correctly estimates the marginal variances, it overestimates the entropy, and when it correctly estimates the entropy, it underestimates at least one marginal variance. The entropy of q in this setting is governed by a *shrinkage-delinkage trade-off* (Margossian and Saul, 2023): the entropy of q is decreased when q underestimates the marginal variances of p (the shrinkage effect), but it is increased when q ignores the correlations in p between off-diagonal components (the delinkage effect).

Proof of trade-off between marginal precisions and generalized variance. This proof follows the same structure as the preceding one. We define auxiliary matrices $\tilde{\mathbf{C}}$

and $\tilde{\mathbf{R}}$ with the dimensionless elements

$$\tilde{C}_{ij} = \frac{\Sigma^{-1}_{ij}}{\sqrt{\Sigma^{-1}_{ii}\Sigma^{-1}_{jj}}}, \qquad \tilde{R}_{ii} = \frac{\Psi^{-1}_{ii}}{\Sigma^{-1}_{ii}}, \tag{12}$$

where $\tilde{\mathbf{C}}$ has all ones along the diagonal and $\tilde{\mathbf{R}}$ has all zeroes off the diagonal. In terms of these matrices, in analogy to eq. (9), we can write

$$\log|\mathbf{\Sigma}| - \log|\mathbf{\Psi}| = \log\left|\mathbf{R}^{-\frac{1}{2}}\mathbf{C}\mathbf{R}^{-\frac{1}{2}}\right| = \log\left|\tilde{\mathbf{R}}^{\frac{1}{2}}\tilde{\mathbf{C}}^{-1}\tilde{\mathbf{R}}^{\frac{1}{2}}\right| = \log|\tilde{\mathbf{R}}| - \tfrac{1}{2}\log|\tilde{\mathbf{C}}|. \tag{13}$$

As before, the determinant of $\tilde{\mathbf{R}}$ is easy to compute because it is a diagonal matrix, and again we deduce from the Hadamard inequality that $|\tilde{\mathbf{C}}| < \prod_i \tilde{\mathbf{C}}_{ii} = 1$. Thus we find

$$\log|\mathbf{\Sigma}| - \log|\mathbf{\Psi}| \;>\; \sum_i \log\left(\frac{\Psi^{-1}_{ii}}{\Sigma^{-1}_{ii}}\right). \tag{14}$$

Note that the direction of inequality in eq. (14) is reversed from that in eq. (11). Now suppose that $\mathrm{diag}(\mathbf{\Psi}^{-1}) = \mathrm{diag}(\mathbf{\Sigma}^{-1})$; then from eq. (14) we see that $|\mathbf{\Sigma}| > |\mathbf{\Psi}|$, proving the first part of the trade-off. Conversely, if $|\mathbf{\Sigma}| = |\mathbf{\Psi}|$, then from eq. (14) we see that $\sum_i \log\Sigma^{-1}_{ii} > \sum_i \log\Psi^{-1}_{ii}$, and this can only be true if $\Sigma^{-1}_{ii} > \Psi^{-1}_{ii}$ for some i. This proves the second part of the trade-off. ■

We can also interpret this trade-off in the special case when p and q are Gaussian, as in eqs. (3–4). In this case, it states the following: when q correctly estimates the marginal precisions, it underestimates the entropy, and when it correctly estimates the entropy, it underestimates at least one marginal precision.

Next we prove the final trade-off between marginal variances and precisions. A short proof of this trade-off was given in Margossian and Saul (2023). Here instead we appeal to a stronger result from linear algebra (Horn and Johnson, 2012, Theorem 7.17), which will also prove useful in Section 5. This result is given by the following lemma.

Lemma 5 *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$. *If* $\mathbf{A} \succ \mathbf{0}$, *then*

$$A_{ii}A^{-1}_{ii} \geq 1 \quad \textit{for all } i, \tag{15}$$

$$A_{jj}A^{-1}_{jj} > 1 \quad \textit{whenever } \textstyle\sum_{k\neq j}|A_{jk}| > 0. \tag{16}$$

Any non-diagonal element A_{jk} in this lemma generates strict inequalities of the form in eq. (16) for the diagonal elements A_{jj} and A_{kk}. From this lemma, we obtain the variance-precision trade-off by substituting either the covariance matrix $\mathbf{\Sigma}$ or the precision matrix $\mathbf{\Sigma}^{-1}$ in Theorem 4 for the matrix $\mathbf{A}$ in eqs. (15–16). Unfortunately, neither the above lemma nor the proof in Margossian and Saul (2023) offer much intuition. Here, we propose an alternative proof rooted in statistical notions.

Proof of variance-precision trade-off. Let $\xi \sim \mathcal{N}(0, \mathbf{\Sigma})$, and let $\boldsymbol{\xi}_{-i}$ denote all the elements of $\boldsymbol{\xi}$ other than ξ_i. Also denote the density of ξ by π. From the conditional distribution

$$\pi(\xi_i \mid \boldsymbol{\xi}_{-i}) \propto \exp\left(-\frac{1}{2}\xi_i\Sigma^{-1}_{ii}\xi_i - \xi_i\sum_{j\neq i}\Sigma^{-1}_{ij}\xi_j\right), \tag{17}$$

we can identify (from the first quadratic term in the exponent) that $\mathrm{Var}[\xi_i|\boldsymbol{\xi}_{-i}] = 1/\Sigma^{-1}_{ii}$; in other words, the conditional variance of this variable is given by the reciprocal of its marginal precision. Now by the law of total variance, we have that

$$\begin{aligned}
\Sigma_{ii} &= \mathrm{Var}[\xi_i], && (18)\\
&= \mathbb{E}[\mathrm{Var}[\xi_i|\boldsymbol{\xi}_{-i}]] + \mathrm{Var}[\mathbb{E}[\xi_i|\boldsymbol{\xi}_{-i}]], && (19)\\
&\geq \mathbb{E}[\mathrm{Var}[\xi_i|\boldsymbol{\xi}_{-i}]], && (20)\\
&= \mathbb{E}[1/\Sigma^{-1}_{ii}], && (21)\\
&= 1/\Sigma^{-1}_{ii}. && (22)
\end{aligned}$$

Thus we have shown $\Sigma_{ii}\Sigma^{-1}_{ii} \geq 1$ for all i. Moreover, since by assumption $\boldsymbol{\Sigma}$ is not diagonal, there must exist some coordinate such that $\mathbb{E}[\xi_i|\boldsymbol{\xi}_{-i}] \neq \mathbb{E}[\xi_i]$ and $\mathrm{Var}[\mathbb{E}[\xi_i|\boldsymbol{\xi}_{-i}]] > 0$; for this coordinate, the inequality is strict with $\Sigma_{ii}\Sigma^{-1}_{ii} > 1$. Suppose now that q matches the marginal precisions of p; then for all elements along the diagonal, we have

$$\Psi_{ii} = (\Psi^{-1}_{ii})^{-1} = (\Sigma^{-1}_{ii})^{-1} \leq \Sigma_{ii}, \tag{23}$$

and this inequality must be strict for at least one of the marginal variances. This proves one part of the trade-off. Likewise, if q matches the marginal variances of p, then for all elements along the diagonal we have

$$\Psi^{-1}_{ii} = (\Psi_{ii})^{-1} = (\Sigma_{ii})^{-1} \leq \Sigma^{-1}_{ii}, \tag{24}$$

and this inequality must be strict for at least one of the marginal precisions. This proves the other part of the trade-off. ■

4. Divergences for FG-VI

In this section we derive the solutions for FG-VI that are obtained from eq. (1–4) by minimizing the divergences in Table 1. We also highlight the properties of solutions that are obtained in this way. As a useful reference, we summarize the results for these solutions in Table 3. We emphasize that all of the results in this section are based on the further assumption that p and q are Gaussian, as in eqs. (3–4).

4.1 KL Divergences

The KL divergence is not a symmetric function of its arguments, and thus the solutions for FG-VI depend on the direction of the divergence that is minimized. We consider each direction in turn.

Reverse direction: KL($q||p$). Most VI is based on minimizing the reverse KL divergence in Table 1. The implications of this choice have been extensively studied for FG-VI. For completeness, we review certain key results in the notation of this paper. When q is chosen to minimize $\mathrm{KL}(q||p)$, it matches the mean of p (that is, $\boldsymbol{\nu} = \boldsymbol{\mu}$), and its covariance is given by

$$\mathrm{diag}(\boldsymbol{\Psi}^{-1}) = \mathrm{diag}(\boldsymbol{\Sigma}^{-1}). \tag{25}$$

	Reduced divergence	**(Implicit) Solution**
$\mathrm{KL}(q\|\|p)$	$\frac{1}{2}\mathrm{trace}\left(\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right) - \log\left\|\mathbf{\Psi}\mathbf{\Sigma}^{-1}\right\| + c$	$\Psi_{ii} = 1/\Sigma^{-1}_{ii}$
$\mathrm{KL}(p\|\|q)$	$\frac{1}{2}\sum_{i=1}^{n}(\log\frac{\Psi_{ii}}{\Sigma_{ii}} + \frac{\Sigma_{ii}}{\Psi_{ii}}) + c$	$\Psi_{ii} = \Sigma_{ii}$
$\mathrm{D}_\alpha(p\|\|q)$	$\frac{1}{\alpha(\alpha-1)}\left[\left\|\mathbf{\Sigma}_\alpha\right\|^{-\frac{1}{2}}\left\|\mathbf{\Psi}\right\|^{\frac{\alpha}{2}}\left\|\mathbf{\Sigma}\right\|^{\frac{1-\alpha}{2}} - 1\right]$	$\Psi_{ii} = \left[\alpha\mathbf{\Sigma}^{-1} + (1-\alpha)\mathbf{\Psi}^{-1}\right]^{-1}_{ii}$
$\mathrm{S}(q\|\|p)$	$\mathrm{trace}\left(\mathbf{I} - \mathbf{\Psi}\mathbf{\Sigma}^{-1}\right)$	$\underset{\mathbf{s}\geq 0}{\mathrm{argmin}}\left[\frac{1}{2}\mathbf{s}^T\mathbf{H}\mathbf{s} - \mathbf{1}^T\mathbf{s}\right]$, $s_{ii} = \Psi_{ii}\Sigma^{-1}_{ii}$
$\mathrm{S}(p\|\|q)$	$\mathrm{trace}\left(\mathbf{I} - \mathbf{\Sigma}\mathbf{\Psi}^{-1}\right)$	$\underset{\mathbf{t}\geq 0}{\mathrm{argmin}}\left[\frac{1}{2}\mathbf{t}^T\mathbf{J}\mathbf{t} - \mathbf{1}^T\mathbf{t}\right]$, $t_{ii} = \Psi^{-1}_{ii}\Sigma_{ii}$

Table 3: Divergences when p and q are Gaussian (eq 3-4). We report the reduced divergence, obtained after matching the means of p and q. For the Rényi divergence, the optimal covariance can be found by solving a fixed-point equation. For the score-based divergences, we need to solve a non-negative quadratic program. The auxiliary matrix $\mathbf{\Sigma}_\alpha$ is defined in eq. (30), $\mathbf{H}$ in eq. (60), and $\mathbf{J}$ in eq. (63).

Further details of these calculations can be found in the references (Turner and Sahani, 2011; Margossian and Saul, 2023). We summarize the main properties of this solution for uncertainty quantification, all of which follow from the results of the previous section:

(i) The marginal precisions of q match those of p.
(ii) The entropy of q underestimates the entropy of p.
(iii) The marginal variances of q underestimate the marginal variances of p.
(iv) The marginal variances of q match the conditional variances of p when each Gaussian random variable is conditioned on all the others.

It has been widely observed that VI based on the reverse KL divergence tends to underestimate uncertainty (MacKay, 2003; Minka, 2005; Turner and Sahani, 2011; Blei et al., 2017; Giordano et al., 2018). The above results show that the "uncertainty deficit" of FG-VI depends on the measure that is used to quantify uncertainty. Indeed, when p is Gaussian, the marginal precisions are correctly estimated by minimizing $\mathrm{KL}(q||p)$; moreover, it has been shown that when $\mathbf{\Sigma}$ has *constant* off-diagonal entries, the entropy gap between p and q only grows as $\mathcal{O}(\log n)$, while the entropy itself grows as $\mathcal{O}(n)$, meaning that the fractional entropy gap tends to 0 as $n \to \infty$ (Margossian and Saul, 2023, Theorem 3.6). Crucially, for q to estimate the entropy well, it must necessarily underestimate the marginal variances, as prescribed by eq. (9). Other asymptotically correct results are obtained in the thermodynamic limit of Ising models with constant, long-range interactions, where the mean-field approximation has its roots (Parisi, 1988).

Forward direction: KL($p||q$). Next we consider the forward KL divergence in Table 1. This divergence is not generally minimized for VI because it involves an expectation, namely $\mathbb{E}_p[\log(p/q)]$ with respect to the target distribution p. When p is not tractable, it is not

possible to compute this expectation analytically, and it may be too expensive to estimate this expectation by sampling. Still, it is illuminating to contrast the properties of the reverse and forward KL divergences for FG-VI. The latter is similarly minimized by setting $\boldsymbol{\nu}=\boldsymbol{\mu}$, but now we obtain a different solution when the remaining terms are minimized with respect to $\boldsymbol{\Psi}$. In particular, $\mathrm{KL}(p||q)$ is minimized by setting

$$\mathrm{diag}(\boldsymbol{\Psi}) = \mathrm{diag}(\boldsymbol{\Sigma}), \tag{26}$$

thus matching the marginal distributions of p and q (MacKay, 2003). We summarize the main properties of this solution for uncertainty quantification, all of which follow directly from the Impossibility Theorem:

(i) The marginal variances of q match those of p.
(ii) The entropy of q overestimates the entropy of p.
(iii) The marginal precisions of q underestimate the marginal precisions of p.

Comparing the results for $\mathrm{KL}(p||q)$ and $\mathrm{KL}(q||p)$, we see the important effects of the divergence when VI is used for uncertainty quantification. These effects motivate our subsequent study of additional divergences.

4.2 α-divergence

Next we consider a one-parameter family of divergences that includes the KL divergences in the previous section as limiting cases. The α-divergence is given by

$$\mathrm{D}_\alpha(p||q) = \frac{1}{\alpha(\alpha-1)} \int \left(\frac{p^\alpha(\mathbf{z})}{q^\alpha(\mathbf{z})} - 1 \right) q(\mathbf{z})\, \mathrm{d}\mathbf{z}, \tag{27}$$

where it is assumed that $\alpha>0$ and $\alpha\neq 1$. Like the KL divergence, the α-divergence has the property that $\mathrm{D}_\alpha(p||q) \geq 0$, with equality holding if and only $p=q$.

The α-divergence in eq. (27) has been studied in the context of VI (Li and Turner, 2016). The divergence can be defined in various ways, but these various definitions[3] have in common that they all yield the same underlying optimization for VI. We use the above definition (Cichocki and Amari, 2010) to recover the KL divergences in the previous sections as limiting cases:

$$\lim_{\alpha\to 0} \mathrm{D}_\alpha(p||q) = \mathrm{KL}(q||p), \tag{28}$$

$$\lim_{\alpha\to 1} \mathrm{D}_\alpha(p||q) = \mathrm{KL}(p||q). \tag{29}$$

Thus for $\alpha \in (0,1)$, the α-divergences provide a one-parameter family of divergences interpolating between $\mathrm{KL}(q||p)$ and $\mathrm{KL}(p||q)$. Likewise, for $\alpha = 2$, eq. (27) recovers the χ^2-divergence, which can also be minimized for approximate inference (Dieng et al., 2017).

The α-divergence can be computed exactly between two multivariate Gaussians (Burbea, 1984; Liese and Vajda, 1987; Hobza et al., 2009; Gil et al., 2013) such as the ones in eqs. (3–4). To do so, it is convenient to define the matrices

$$\boldsymbol{\Sigma}_\alpha = \alpha\boldsymbol{\Psi} + (1-\alpha)\boldsymbol{\Sigma}, \tag{30}$$

$$\boldsymbol{\Phi}_\alpha = \alpha\boldsymbol{\Sigma}^{-1} + (1-\alpha)\boldsymbol{\Psi}^{-1}, \tag{31}$$

3. Closely related is the Rényi divergence, given by $R_\alpha(p||q) = (\alpha-1)^{-1} \log \int p(\mathbf{z})^\alpha q(\mathbf{z})^{1-\alpha}$.

which are also related by the identity $\boldsymbol{\Sigma}_\alpha = \boldsymbol{\Sigma}\,\boldsymbol{\Phi}_\alpha\boldsymbol{\Psi}$. It is known that the integral for $\mathrm{D}_\alpha(p||q)$ in eq. (27) only exists when $\boldsymbol{\Phi}_\alpha \succ 0$. In this case, the α-divergence is given by

$$\mathrm{D}_\alpha(p||q) = \frac{1}{\alpha(\alpha-1)}\left[e^{-\frac{\alpha(1-\alpha)}{2}(\boldsymbol{\mu}-\boldsymbol{\nu})^\top\boldsymbol{\Sigma}_\alpha^{-1}(\boldsymbol{\mu}-\boldsymbol{\nu})}\,|\boldsymbol{\Sigma}_\alpha|^{-\frac{1}{2}}|\boldsymbol{\Psi}|^{\frac{\alpha}{2}}|\boldsymbol{\Sigma}|^{\frac{1-\alpha}{2}} - 1\right]. \tag{32}$$

This expression vanishes if $\boldsymbol{\mu}=\boldsymbol{\nu}$ and $\boldsymbol{\Psi}=\boldsymbol{\Sigma}$ (in which case $\boldsymbol{\Sigma}_\alpha=\boldsymbol{\Sigma}$), but not otherwise.

To analyze FG-VI with this divergence, we must minimize eq. (32) with respect to the variational mean $\boldsymbol{\nu}$ and diagonal covariance $\boldsymbol{\Psi}$ of q. The next propositions establish useful properties of the variational approximations that are found in this way. In particular, the first shows that q matches the mean of p.

Proposition 6 (Mean matching) *Let $\alpha \in \mathbb{R}^+\backslash\{0,1\}$, and let p and q be given by eqs. (3–4) where $\boldsymbol{\Psi}$ is diagonal. If q minimizes the α-divergence in eq. (27), then it matches the mean of p; that is, $\boldsymbol{\nu}=\boldsymbol{\mu}$.*

Proof We proceed by optimizing the right side of eq. (32). First, we consider the case where $\alpha\in(0,1)$. In this case, the prefactor $1/\alpha(\alpha-1)$ is *negative*, and the expression as a whole is minimized by *maximizing* the first term in brackets. Upon taking logarithms, we see equivalently that

$$\operatorname*{argmin}_{q\in\mathcal{Q}}\Big[\mathrm{D}_\alpha(p||q)\Big] = \operatorname*{argmin}_{\boldsymbol{\nu},\boldsymbol{\Psi}}\Big[\alpha(1-\alpha)(\boldsymbol{\mu}-\boldsymbol{\nu})^\top\boldsymbol{\Sigma}_\alpha^{-1}(\boldsymbol{\mu}-\boldsymbol{\nu}) + \log|\boldsymbol{\Sigma}_\alpha| - \alpha\log|\boldsymbol{\Psi}|\Big]. \tag{33}$$

The first term on the right is minimized (and zeroed) by setting $\boldsymbol{\nu}=\boldsymbol{\mu}$, thus matching the means of q and p.

A similar proof holds when $\alpha > 1$. ■

The next proposition shows that q estimates finite, non-zero variances, or equivalently, that $\boldsymbol{\Psi}\succ 0$ and $\boldsymbol{\Psi}^{-1}\succ 0$ for any $\boldsymbol{\Psi}$ minimizing the α-divergence in eq. (32).

Proposition 7 (Variance bounds) *Let $\alpha \in \mathbb{R}^+\backslash\{0,1\}$, and let p and q be given by eqs. (3–4) where $\boldsymbol{\Psi}$ is diagonal. If q minimizes the α-divergence in eq. (27), then its variances are strictly positive and finite; that is, $0<\Psi_{ii}<\infty$ for all i.*

Proof Again we proceed by minimizing the right side of eq. (32). We first consider the case $\alpha\in(0,1)$. By the previous proposition, the minima occurs at $\boldsymbol{\nu}=\boldsymbol{\mu}$. Making this substitution, we find

$$\operatorname*{argmin}_{q\in\mathcal{Q}}\left[\mathrm{D}_\alpha(p||q)\right] = \operatorname*{argmin}_{\boldsymbol{\Psi}}\frac{1}{\alpha(\alpha-1)}\,|\boldsymbol{\Sigma}_\alpha|^{-\frac{1}{2}}|\boldsymbol{\Psi}|^{\frac{\alpha}{2}} \tag{34}$$

$$= \operatorname*{argmax}_{\boldsymbol{\Psi}}|\boldsymbol{\Sigma}_\alpha|^{-\frac{1}{2}}\,|\boldsymbol{\Psi}|^{\frac{\alpha}{2}} \tag{35}$$

$$= \operatorname*{argmax}_{\boldsymbol{\Psi}}\left|\alpha\boldsymbol{\Psi}+(1-\alpha)\boldsymbol{\Sigma}\right|^{-\frac{1}{2}}\left|\boldsymbol{\Psi}^{-\alpha}\right|^{-\frac{1}{2}} \tag{36}$$

$$= \operatorname*{argmin}_{\boldsymbol{\Psi}}\left|\alpha\boldsymbol{\Psi}^{1-\alpha}+(1-\alpha)\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-\alpha}\right|. \tag{37}$$

The determinant in eq. (37) is finite for any diagonal covariance $\mathbf{\Psi}$ satisfying $\Psi \succ 0$ and $\Psi^{-1} \succ 0$. But for $\alpha \in (0,1)$, this determinant diverges if any $\Psi_{ii} \to \infty$ due to the first term $\alpha \mathbf{\Psi}^{1-\alpha}$, and likewise it diverges if any $\Psi_{ii} \to 0$ due to the second term $(1-\alpha)\mathbf{\Sigma}\mathbf{\Psi}^{-\alpha}$. This proves the proposition for the case $\alpha \in (0,1)$.

Now suppose $\alpha > 1$. Recall that the α-divergence in eq. (32) is only defined for $\mathbf{\Phi}_\alpha \succ 0$ where $\mathbf{\Phi}_\alpha = \alpha\mathbf{\Sigma}^{-1} + (1-\alpha)\mathbf{\Psi}^{-1}$. But when $\alpha > 1$, the condition $\mathbf{\Phi}_\alpha \succ 0$ can only be satisfied if $\Psi_{ii} > 0$ for all i. To see in addition that $\Psi_{ii} < \infty$, we proceed as before:

$$\begin{aligned}
\underset{q \in \mathcal{Q}}{\operatorname{argmin}} \left[\mathrm{D}_\alpha(p||q)\right] &= \underset{\mathbf{\Psi}}{\operatorname{argmin}} \frac{1}{\alpha(\alpha-1)} |\mathbf{\Sigma}_\alpha|^{-\frac{1}{2}} |\mathbf{\Psi}|^{\frac{\alpha}{2}} && (38)\\
&= \underset{\mathbf{\Psi}}{\operatorname{argmin}} \left|\mathbf{\Sigma}\mathbf{\Phi}_\alpha\mathbf{\Psi}\right|^{-\frac{1}{2}} \left|\mathbf{\Psi}\right|^{\frac{\alpha}{2}} && (39)\\
&= \underset{\mathbf{\Psi}}{\operatorname{argmax}} \left|\mathbf{\Phi}_\alpha \mathbf{\Psi}^{1-\alpha}\right| && (40)\\
&= \underset{\mathbf{\Psi}}{\operatorname{argmax}} \left|\alpha\mathbf{\Sigma}^{-1}\mathbf{\Psi}^{1-\alpha} + (1-\alpha)\mathbf{\Psi}^{-\alpha}\right|. && (41)
\end{aligned}$$

The determinant in eq. (40) is strictly positive for any covariance $\mathbf{\Psi}$ with bounded elements satisfying $\mathbf{\Phi}_\alpha \succ 0$. For $\alpha > 1$ this same determinant, rewritten in eq. (41), vanishes if any $\Psi_{ii} \to \infty$ because in this limit the matrices $\mathbf{\Psi}^{1-\alpha}$ and $\mathbf{\Psi}^{-\alpha}$ become low-rank. Thus the maximum must occur where $0 < \Psi_{ii} < \infty$ for all i. This proves the proposition for $\alpha > 1$. ■

We can prove further properties of the variational approximation q that minimizes $\mathrm{D}_\alpha(q||p)$ in eq. (32). Though the minimizer of $\mathrm{D}_\alpha(q||p)$ does not admit an analytical expression, it is possible to derive fixed-point equations that are satisfied by the variances Ψ_{ii} at this minimizer. As we shall see later, these fixed-point equations encode a great deal of information about variational approximations with the α-divergence.

Proposition 8 (Fixed-point equations) *Let $\alpha \in \mathbb{R}^+ \backslash \{0,1\}$, and let p and q be given by eqs. (3–4) where $\mathbf{\Psi}$ is diagonal. Then the α-divergence in eq. (27) is minimized when $\boldsymbol{\nu} = \boldsymbol{\mu}$ and the estimated precisions from $\mathbf{\Psi}$ satisfy the fixed-point equation*

$$\operatorname{diag}\left(\mathbf{\Psi}^{-1}\right) = \operatorname{diag}\left(\mathbf{\Sigma}_\alpha^{-1}\right), \tag{42}$$

or equivalently when the estimated variances from $\mathbf{\Psi}$ satisfy the fixed-point equation

$$\operatorname{diag}(\mathbf{\Psi}) = \operatorname{diag}\left(\mathbf{\Phi}_\alpha^{-1}\right). \tag{43}$$

Proof We derive the fixed-point equations by attempting to minimize the α-divergence in eq. (32) with respect to the variational parameters. First, we consider the case where $\alpha \in (0,1)$. In this case, from eq. (37), we have

$$\underset{q \in \mathcal{Q}}{\operatorname{argmin}} \Big[\mathrm{D}_\alpha(p||q)\Big] = \underset{\mathbf{\Psi}}{\operatorname{argmin}} \Big[\log|\alpha\mathbf{\Psi} + (1-\alpha)\mathbf{\Sigma}| - \alpha \log|\mathbf{\Psi}|\Big]. \tag{44}$$

From the previous propositions, we know that the minimum of eq. (44) occurs where $0<\Psi_{ii}<\infty$. Solving for the minimum, we find

$$0 = \frac{\partial}{\partial \Psi_{ii}}\Big[\log|\alpha\mathbf{\Psi} + (1-\alpha)\mathbf{\Sigma}| - \alpha\log|\mathbf{\Psi}|\Big], \tag{45}$$

$$= \alpha[\alpha\mathbf{\Psi} + (1-\alpha)\mathbf{\Sigma}]^{-1}_{ii} - \alpha\Psi^{-1}_{ii}, \tag{46}$$

$$= \alpha\left[\mathbf{\Sigma}_\alpha^{-1} - \mathbf{\Psi}^{-1}\right]_{ii}, \tag{47}$$

thus proving eq. (42).

To derive the fixed-point equation in eq. (43), we rewrite eq. (44) as

$$\underset{q\in\mathcal{Q}}{\operatorname{argmin}}\Big[\mathrm{D}_\alpha(p||q)\Big] = \underset{\mathbf{\Psi}}{\operatorname{argmin}}\Big[\log|\alpha\mathbf{\Psi} + (1-\alpha)\mathbf{\Sigma}| - \alpha\log|\mathbf{\Psi}| + \log|\mathbf{\Psi}| - \log|\mathbf{\Psi}|\Big], \tag{48}$$

$$= \underset{\mathbf{\Psi}}{\operatorname{argmin}}\Big[\log\left|\alpha + (1-\alpha)\mathbf{\Sigma}\mathbf{\Psi}^{-1}\right| - (1-\alpha)\log\left|\mathbf{\Psi}^{-1}\right|\Big], \tag{49}$$

$$= \underset{\mathbf{\Psi}^{-1}}{\operatorname{argmin}}\Big[\log\left|\alpha\mathbf{\Sigma}^{-1} + (1-\alpha)\mathbf{\Psi}^{-1}\right| - (1-\alpha)\log\left|\mathbf{\Psi}^{-1}\right|\Big], \tag{50}$$

where in the last step we have exploited the fact that additive constants, such as $\log\left|\mathbf{\Sigma}^{-1}\right|$, do not change the location of the minimum with respect to $\mathbf{\Psi}$ (or equivalently, with respect to $\mathbf{\Psi}^{-1}$). Solving for the minimum, we find

$$0 = \frac{\partial}{\partial \Psi^{-1}_{ii}}\Big[\log\left|\alpha\mathbf{\Sigma}^{-1} + (1-\alpha)\mathbf{\Psi}^{-1}\right| - (1-\alpha)\log\left|\mathbf{\Psi}^{-1}\right|\Big] = (1-\alpha)\left[\mathbf{\Phi}_\alpha^{-1} - \mathbf{\Psi}\right]_{ii}, \tag{51}$$

which is equivalent to the claim that $\mathrm{diag}(\mathbf{\Psi}) = \mathrm{diag}\left(\mathbf{\Phi}_\alpha^{-1}\right)$ in eq. (43).

A similar procedure can be used to derive the fixed-point equations when $\alpha>1$. In this case, from eq. (41) we have

$$\underset{q\in\mathcal{Q}}{\operatorname{argmin}}\Big[\mathrm{D}_\alpha(p||q)\Big] = \underset{\mathbf{\Psi}}{\operatorname{argmax}}\big[\log|\alpha\mathbf{\Psi} + (1-\alpha)\mathbf{\Sigma}| - \alpha\log|\mathbf{\Psi}|\big]. \tag{52}$$

As before, from the previous propositions, we can exclude the edge cases $\Psi_{ii}=0$ and $\Psi_{ii}=\infty$ as potential solutions, and therefore the maximum of the right side in eq. (52) is determined by the vanishing of its gradient with respect to $\mathbf{\Psi}$. But the gradients of eqs. (44) and (52) are identical and therefore yield the same fixed-point equations. ■

Proposition 8 provides two (equivalent) fixed-point equations for the variational parameter $\mathbf{\Psi}$, and it will prove useful to work with both in Section 5. In componentwise notation, these fixed-point equations take the form

$$\Psi^{-1}_{ii} = \big[\alpha\mathbf{\Psi} + (1-\alpha)\mathbf{\Sigma}\big]^{-1}_{ii}, \tag{53}$$

$$\Psi_{ii} = \big[\alpha\mathbf{\Sigma}^{-1} + (1-\alpha)\mathbf{\Psi}^{-1}\big]^{-1}_{ii}, \tag{54}$$

and we see that eq. (53) reduces to the precision-matching solution $\Psi^{-1}_{ii} = \Sigma^{-1}_{ii}$ as $\alpha\to 0$ and eq. (54) reduces to the variance-matching solution $\Psi_{ii} = \Sigma_{ii}$ as $\alpha\to 1$.

4.3 Score-based Divergence

Next we consider the score-based divergences presented in Table 1. These divergences are similar to the relative Fisher information (Courtade, 2016, eq. 7) except that they use a *weighted* norm (Cai et al., 2024b, Appendix A) to measure the difference between the gradients of $\log p$ and $\log q$. When p and q are Gaussian, these score-based divergences are given by

$$S(q||p) = \int \mathrm{d}\mathbf{z}\, q(\mathbf{z}) \left\|\nabla \log q(\mathbf{z}) - \nabla \log p(\mathbf{z})\right\|^2_{\mathrm{Cov}(q)}, \tag{55}$$

$$S(p||q) = \int \mathrm{d}\mathbf{z}\, q(\mathbf{z}) \left\|\nabla \log q(\mathbf{z}) - \nabla \log p(\mathbf{z})\right\|^2_{\mathrm{Cov}(p)}, \tag{56}$$

where we use $\|\mathbf{v}\|^2_{\mathbf{A}} = \mathbf{v}^\top \mathbf{A}\mathbf{v}$ to denote the weighted norm for any $\mathbf{A} \succ \mathbf{0}$. The score-based divergences in eqs. (55-56) are dimensionless measures of the discrepancy between p and q, and in particular, like the KL and α-divergences in the previous sections, they are invariant to affine reparameterizations of the support of these distributions (Cai et al., 2024b, Theorem A.4). For p and q in eqs. (3–4), these divergences are given by

$$S(q||p) = \mathrm{tr}\left[\left(\mathbf{I} - \boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right)^2\right] + (\boldsymbol{\nu}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\nu}-\boldsymbol{\mu}), \tag{57}$$

$$S(p||q) = \mathrm{tr}\left[\left(\mathbf{I} - \boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}\right)^2\right] + (\boldsymbol{\mu}-\boldsymbol{\nu})^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}(\boldsymbol{\mu}-\boldsymbol{\nu}), \tag{58}$$

as also shown in Cai et al. (2024b). The rest of this section examines the solutions that minimize these expressions with respect to the variational mean $\boldsymbol{\nu}$ and diagonal covariance $\boldsymbol{\Psi}$.

First we consider the minimization of the *reverse* score-based divergence, $S(q||p)$. This minimization can be formulated as a nonnegative quadratic program (NQP)—that is, a convex instance of quadratic programming with nonnegativity constraints.

Proposition 9 (NQP for minimizing $S(q||p)$) *Let p and q be given by eqs. (3–4) where $\boldsymbol{\Psi}$ is diagonal. Then the reverse score-based divergence in eq. (57) is minimized by setting $\boldsymbol{\nu} = \boldsymbol{\mu}$ and solving the quadratic program*

$$\min_{\mathbf{s}\geq 0} \left[\tfrac{1}{2}\mathbf{s}^\top \mathbf{H}\,\mathbf{s} - \mathbf{1}^\top \mathbf{s}\right], \tag{59}$$

where $\mathbf{s}$ is constrained to lie in the nonnegative orthant, $\mathbf{1}$ is the vector of all ones, $\mathbf{H}$ is the positive-definite matrix with elements

$$H_{ij} = \frac{\left(\Sigma^{-1}_{ij}\right)^2}{\Sigma^{-1}_{ii}\Sigma^{-1}_{jj}}, \tag{60}$$

and $\boldsymbol{\Psi}$ is obtained from the solution of eq. (59) by identifying $s_i = \Psi_{ii}\Sigma^{-1}_{ii}$.

Proof It is clear that eq. (57) is minimized by setting $\boldsymbol{\nu} = \boldsymbol{\mu}$ in the rightmost term. For the remaining term, we note that

$$\tfrac{1}{2}\,\mathrm{tr}\left[\left(\mathbf{I} - \boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right)^2\right] \;=\; \tfrac{n}{2} \;-\; \sum_i \Psi_{ii}\Sigma^{-1}_{ii} \;+\; \tfrac{1}{2}\sum_{ij}\Psi_{ii}\Psi_{jj}\left(\Sigma^{-1}_{ij}\right)^2, \tag{61}$$

and the NQP in eq. (59) is obtained by defining $s_i = \Psi_{ii}\Sigma_{ii}^{-1}$. Finally, we observe that the matrix $\mathbf{H}$ in eq. (60) is always positive-definite. Indeed, $\mathbf{H}$ is the Hadamard (i.e., component-wise) square of the correlation matrix built from the precision matrix $\mathbf{\Sigma}^{-1}$, and we note from the Schur product theorem that the Hadamard product of two positive-definite matrices is also positive-definite (Horn and Johnson, 2012). ■

Given the symmetry of eqs. (57-58), it is not surprising that we obtain an analogous optimization when minimizing the *forward* score-based divergence, $S(p||q)$, in terms of the variational parameters $\boldsymbol{\nu}$ and $\mathbf{\Psi}$. In particular, we have the following result.

Proposition 10 (NQP for minimizing $S(q||p)$) *Let p and q be given by eqs. (3–4) where $\mathbf{\Psi}$ is diagonal. Then the forward score-based divergence in eq. (57) is minimized by setting $\boldsymbol{\nu} = \boldsymbol{\mu}$ and solving the quadratic program*

$$\min_{\mathbf{t} \geq 0} \left[\tfrac{1}{2} \mathbf{t}^\top \mathbf{J}\, \mathbf{t} - \mathbf{1}^\top \mathbf{t} \right], \tag{62}$$

where $\mathbf{t}$ is constrained to lie in the nonnegative orthant, $\mathbf{J}$ is the positive-definite matrix with elements

$$J_{ij} = \frac{(\Sigma_{ij})^2}{\Sigma_{ii}\Sigma_{jj}}, \tag{63}$$

and $\mathbf{\Psi}$ is obtained from the solution of eq. (62) by identifying $t_i = \Psi_{ii}^{-1}\Sigma_{ii}$.

Proof The proof follows the same steps as in the previous proposition, but with the matrices $\mathbf{\Psi}$ and $\mathbf{\Sigma}$ playing the reverse roles as they did in eq. (61). ■

We conclude this section by noting a peculiar property of the score-based divergences in Table 1. It is possible for the solutions of the NQPs in Propositions 9 and 10 to lie on the boundary of the nonnegative orthant. That is, there may exist some i^{th} component along the diagonal of $\mathbf{\Psi}$ such that $S(q||p)$ is minimized by setting $\Psi_{ii} = 0$ (in Proposition 9) or such that $S(p||q)$ is minimized by setting $\Psi_{ii}^{-1} = 0$ (in Proposition 10). Strictly speaking, *solutions of this form—with zero or infinite marginal variances—do not define proper distributions that lie in the family $\mathcal{Q}$ of factorized Gaussian approximations.* Such solutions do not arise with FG-VI based on the KL or α-divergences. This phenomenon gives rise to the following definition.

Definition 11 *Let $\mathcal{Q}$ denote the family of factorized Gaussian approximations in eq. (2) that are proper distributions (i.e., with $0 < \mathcal{H}(q) < \infty$). For some target distribution p and divergence D, we say that FG-VI undergoes **variational collapse** when* $\operatorname{arginf}_{q \in \mathcal{Q}} D(q, p) \notin \mathcal{Q}$.

Fig. 1 illustrates this collapse when FG-VI is used to approximate a multivariate Gaussian distribution in three dimensions with varying degrees of correlation; the amount of correlation is determined by the magnitude of the off-diagonal elements C_{12}, C_{23}, and C_{13} in eq. (7). These off-diagonal elements are constrained by the fact that the correlation matrix

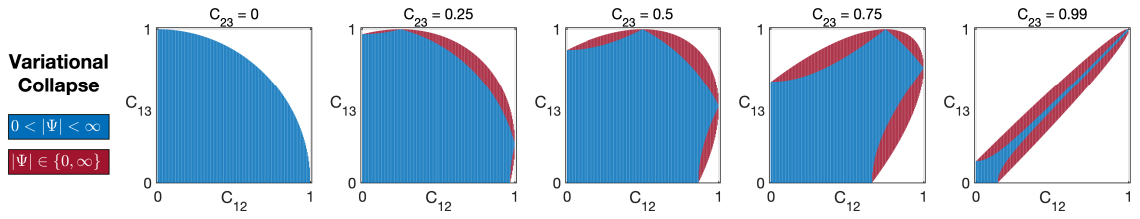

Figure 1: When FG-VI is based on minimizing the score-based divergences in Table 1, it may estimate zero or infinite values for the marginal variances. The red areas indicate these occurrences of *variational collapse* when FG-VI with a score-based divergence is used to approximate a three-dimensional Gaussian with a non-diagonal correlation matrix $\mathbf{C}$.

as a whole must be positive definite. Each panel in the figure visualizes a two-dimensional slice of the three-dimensional (convex) set of positive-definite correlation matrices for a fixed value of C_{23}. The blue regions of each slice indicate the areas where the NQPs in Propositions 9 and 10 yield finite and nonzero estimates of the variances; the red regions indicate areas of variational collapse. (It can be shown, in three dimensions, that the forward and reverse score-based divergences share the same areas of variational collapse.) Interestingly, the leftmost panel shows that variational collapse does not occur in two dimensions or arise purely from pairwise correlations. On the other hand, the remaining panels show that variational collapse occurs in three dimensions whenever the correlation matrix is dense and has at least one off-diagonal element of sufficiently large magnitude.

The above example shows that variational collapse occurs when some pairwise correlations are large and the assumption of factorization is strongly violated. This may not be a problem in settings where we expect the target covariance $\mathbf{\Sigma}$ to be sparse. However, in many settings, we do not have a priori knowledge on the structure of $\mathbf{\Sigma}$. In this case, it seems fraught to minimize a score-based divergence, particularly if the goal of FG-VI is to provide some quantification of uncertainty (whether it be with marginal variances, marginal precisions or generalized variance). Still, a score-based divergence may work well in conjunction with a richer family of approximations—for instance, a multivariate Gaussian with a dense covariance matrix, as demonstrated by Cai et al. (2024b).

5. Ordering of Divergences for FG-VI

We have seen that different divergences $D(q, p)$ yield different solutions to the problem of variational inference in eq. (1); we have also seen, in turn, that these different solutions yield different estimates of the marginal variances, precisions, and entropy. In the setting where both p and q are Gaussian, these estimates provide a natural way to compare and even order the divergences; see Definition 2. In this section, we show that the divergences in Table 1 can be ordered in this way; that is, for $0 < \alpha_1 < \alpha_2 < 1$, we prove that

$$S(q||p) \prec \mathrm{KL}(q||p) \prec \mathrm{D}_{\alpha_1}(p||q) \prec \mathrm{D}_{\alpha_2}(p||q) \prec \mathrm{KL}(p||q) \prec S(p||q),$$

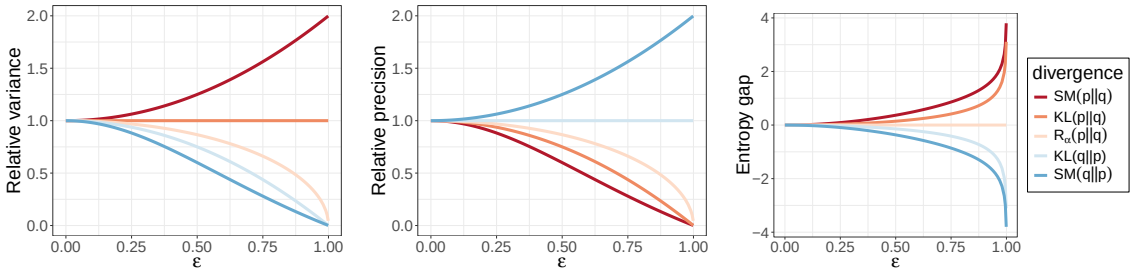

Figure 2: Variances, precisions, and entropy estimated by FG-VI with different divergences. In the left and center panels, the variance is normalized by the variance of p, and the precision by the precision of p, both along the first coordinate. In the right panel, we plot the difference between the estimated entropy and the entropy of p. Here FG-VI was used to approximate a 2-dimensional Gaussian with correlation ε. The α-divergence was computed for $\alpha = 0.5$. The ordering of the curves matches the predictions of Theorem 3.

as stated in Theorem 3. We also provide slightly weaker results on α-divergences for the case were $\alpha > 1$.

Figure 2 illustrates the consequences of this ordering when FG-VI is used to approximate a 2-dimensional Gaussian target with varying degrees of correlation. As the amount of correlation (denoted by ε) increases from zero to one—from no correlation to perfect correlation—the target Gaussian more starkly violates the assumption of factorization. From left to right, the panels plot the estimates of variance, precision, and entropy that are obtained from FG-VI with different divergences, and we see that for all values of ε, these estimates are ordered exactly as predicted by Theorem 3.

We have already proven one of these orderings. Recall from section 4.1 that when FG-VI is based on minimizing $\mathrm{KL}(q||p)$, it underestimates the marginal variances, at least one of them strictly. On the other hand, when FG-VI is based on minimizing $\mathrm{KL}(p||q)$, it correctly estimates the marginal variances. Thus we have already shown that $\mathrm{KL}(q||p) \prec \mathrm{KL}(q||p)$. The rest of this section is devoted to proving the other orderings in Theorem 3.

5.1 Ordering of Score-based Divergences

In this section we prove the two outermost orderings in Theorem 3; specifically, we show that $S(q||p) \prec \mathrm{KL}(q||p)$ and $\mathrm{KL}(p||q) \prec S(p||q)$. At a high level, these proofs are obtained by analyzing the Karush-Kuhn-Tucker (KKT) conditions of the quadratic programs in Propositions 9 and 10.

Proof of $S(q||p) \prec KL(q||p)$. Consider the diagonal covariance matrix $\mathbf{\Psi}$ that FG-VI estimates by minimizing $S(q||p)$ on the left side of this ordering. This matrix is obtained from the solution of the NQP in eq. (59) for the unknown nonnegative variables $s_i = \Psi_{ii}\Sigma^{-1}_{ii}$. The solution to this NQP must satisfy the KKT conditions: namely, for each component of $\mathbf{s}$, we have either that (i) $s_i = 0$ and $(\mathbf{Hs})_i > 1$ or (ii) $s_i > 0$ and $(\mathbf{Hs})_i = 1$. We examine each of these cases in turn:

$\boxed{s_i = 0}$ This is a case of variational collapse where by minimizing $S(q||p)$ we estimate that $\Psi_{ii} = 0$.

$\boxed{s_i > 0}$ In this case we observe that $(\mathbf{Hs})_i = 1$ from the KKT condition and also that $H_{ij} \geq 0$ and $H_{ii} = 1$ from eq. (60). Thus we see that

$$s_i = H_{ii} s_i = 1 - \sum_{j \neq i} H_{ij} s_j \leq 1, \tag{64}$$

or equivalently, that $\Psi_{ii} \Sigma_{ii}^{-1} \leq 1$.

We have thus shown that $\Psi_{ii} \Sigma_{ii}^{-1} \leq 1$ for all of the marginal variances estimated by minimizing the score-based divergence $S(q||p)$. Next we show that this inequality must be strict for at least one component. The claim is trivially true if any $s_i = 0$, since this implies $\Psi_{ii} \Sigma_{ii}^{-1} = 0$. Suppose to the contrary that every element of $\mathbf{s}$ is strictly positive. Since by assumption $\mathbf{\Sigma}$ is not diagonal, it must also be true that $\mathbf{\Sigma}^{-1}$ is not diagonal, and therefore from eq. (60) we can pick some i and k such that $H_{ik} > 0$. But then, continuing from eq. (64), we see that

$$s_i = 1 - \sum_{j \neq i} H_{ij} s_j \leq 1 - H_{ik} s_k < 1, \tag{65}$$

or equivalently that $\Psi_{ii} \Sigma_{ii}^{-1} < 1$, thus proving the claim. In sum we have shown that $\Psi_{ii} \leq 1/\Sigma_{ii}^{-1}$ for all of the marginal variances estimated by minimizing $S(q||p)$, and also that this inequality is strict for at least one component. Finally, we recall from eq. (25) that $1/\Sigma_{ii}^{-1}$ is the marginal variance estimated by minimizing $\mathrm{KL}(q||p)$. Per Definition 2 we conclude that $S(q||p) \prec \mathrm{KL}(q||p)$. ■

We use the same methods to prove an analogous ordering for the forward score-matching and KL divergences.

Proof of $S(p||q) \succ KL(p||q)$. To prove the order we consider the variances Ψ_{ii} that minimize $S(q||p)$; they are found by solving the NQP in eq. (62) for the nonnegative variables $t_i = \Psi_{ii}^{-1} \Sigma_{ii}$. The solution must satisfy the KKT conditions: either (i) $t_i = 0$ and $(\mathbf{Jt})_i > 1$ or (ii) $t_i > 0$ and $(\mathbf{Jt})_i = 1$. We examine each of these cases in turn:

$\boxed{t_i = 0}$ This is a case of variational collapse where by minimizing $S(p||q)$ we estimate that $\Psi_{ii}^{-1} = 0$; that is, $\Psi_{ii} = \infty$, and so $\Psi_{ii} > \Sigma_{ii}$.

$\boxed{t_i > 0}$ In this case, we observe that $(\mathbf{Jt})_i = 1$ from the KKT condition and also that $J_{ij} \geq 0$ and $J_{ii} = 1$ from eq. (63). Thus we see that

$$t_i = J_{ii} t_i = 1 - \sum_{j \neq i} J_{ij} t_j \leq 1, \tag{66}$$

or equivalently, that $\Psi_{ii}^{-1} \Sigma_{ii} \leq 1$ and $\Psi_{ii} \geq \Sigma_{ii}$. Moreover, since $\mathbf{\Sigma}$ is non-diagonal, $\mathbf{J}$ must also be non-diagonal, and hence there must exist at least one coordinate j such that $\Psi_{jj} > \Sigma_{jj}$.

We have shown that $\Psi_{ii} \geq \Sigma_{ii}$ for all i, and also that this inequality is strict for at least one component of the variance. Finally, we recall that the correct marginal variance (namely, Σ_{ii}) is estimated by VI when minimizing $\mathrm{KL}(p||q)$. Per Definition 2 we conclude that $S(p||q) \succ \mathrm{KL}(p||q)$. ■

5.2 Ordering of KL and α-divergences

Next we prove the two intermediate orderings in Theorem 3; specifically, for any $\alpha \in (0,1)$, we show that $\mathrm{KL}(q||p) \prec \mathrm{D}_\alpha(p||q) \prec \mathrm{KL}(p||q)$. These proofs are obtained by applying the inequalities in Lemma 5 to the fixed-point equations for the optimal marginal variances in eqs. (53–54), obtained by minimizing the α-divergence. The same approach allows us to show that $\mathrm{D}_\alpha(p||q) \succ \mathrm{KL}(p||q)$ for $\alpha > 1$.

Proof of $KL(q||p) \prec D_\alpha(p||q)$ for $\alpha \in (0,1)$. Let $\boldsymbol{\Psi}$ denote the diagonal covariance matrix estimated by minimizing $\mathrm{D}_\alpha(p||q)$, and consider the matrix $[\alpha\boldsymbol{\Sigma}^{-1} + (1-\alpha)\boldsymbol{\Psi}]^{-1}$ that appears on the right side of its fixed-point equation in eq. (54). For $\alpha \in (0,1)$, it is clear that this matrix is positive-definite and non-diagonal, and thus we can apply the inequality in eq. (15) of Lemma 5 to its diagonal elements. In this way, we find:

$$\begin{aligned}
\alpha\Psi_{ii}\Sigma_{ii}^{-1} &= \Psi_{ii}\left[\alpha\boldsymbol{\Sigma}^{-1} + (1-\alpha)\boldsymbol{\Psi}^{-1}\right]_{ii} - (1-\alpha) && (67)\\
&\geq \frac{\Psi_{ii}}{\left[\alpha\boldsymbol{\Sigma}^{-1} + (1-\alpha)\boldsymbol{\Psi}^{-1}\right]^{-1}_{ii}} - (1-\alpha), && (68)\\
&= (\Psi_{ii}/\Psi_{ii}) - (1-\alpha), && (69)\\
&= \alpha. && (70)
\end{aligned}$$

Dividing both sides by $\alpha\Sigma_{ii}^{-1}$, we see that $\Psi_{ii} \geq 1/\Sigma_{ii}^{-1}$ for all i. Next we appeal to the strict inequality of Lemma 5 and conclude that $\Psi_{jj} > 1/\Sigma_{jj}^{-1}$ for some j. On the left and right sides of these inequalities appear, respectively, the marginal variances estimated by minimizing $D_\alpha(p,q)$ and $\mathrm{KL}(q,p)$. Thus we have shown $\mathrm{D}_\alpha(p||q) \succ \mathrm{KL}(q,p)$. ■

Proof of $D_\alpha(p||q) \prec KL(p||q)$ for $\alpha \in (0,1)$. The proof follows the same steps as the previous one, but now we combine the inequalities of Lemma 5 with the fixed-point equation in eq. (53). In this way we find:

$$(1-\alpha)\Psi_{ii}^{-1}\Sigma_{ii} = \Psi_{ii}^{-1}\left[\alpha\boldsymbol{\Psi} + (1-\alpha)\boldsymbol{\Sigma}\right]_{ii} - \alpha \geq \frac{\Psi_{ii}^{-1}}{\left[\alpha\boldsymbol{\Psi} + (1-\alpha)\boldsymbol{\Sigma}\right]^{-1}_{ii}} - \alpha = 1-\alpha. \quad (71)$$

Dividing both sides by $(1-\alpha)\Psi_{ii}^{-1}$, we see that $\Sigma_{ii} \geq \Psi_{ii}$ for all i, and we know from the strict inequality in eq. (16) of Lemma 5 that $\Sigma_{jj} > \Psi_{jj}$ for some j. On the left and right sides of these inequalities appear, respectively, the marginal variances estimated by minimizing $\mathrm{KL}(p||q)$ and $\mathrm{D}_\alpha(p||q)$. Thus we have shown $\mathrm{KL}(p||q) \succ \mathrm{D}_\alpha(p||q)$ for $\alpha \in (0,1)$. ■

Proof of $D_\alpha(p||q) \succ KL(p||q)$ for $\alpha > 1$. To prove this ordering we need to justify the steps in eq. (71) when $\alpha > 1$, as opposed to when $\alpha \in (0,1)$. But when $\alpha > 1$, it is not obvious that the matrix $\boldsymbol{\Sigma}_\alpha = \alpha\boldsymbol{\Psi} + (1-\alpha)\boldsymbol{\Sigma}$ is positive-definite; this must be demonstrated before

we can apply the inequality in Lemma 5. Therefore we begin by showing that $\mathbf{\Sigma}_\alpha$ is positive-definite whenever the matrices $\mathbf{\Psi}$, $\mathbf{\Psi}^{-1}$, and $\mathbf{\Phi}_\alpha$ are positive-definite. (We know that $\mathbf{\Psi} \succ 0$ and $\mathbf{\Psi}^{-1} \succ 0$ from Proposition 7, and we know that $\mathbf{\Phi}_\alpha \succ 0$ because it is necessary for the α-divergence between two multivariate Gaussians to be well-defined.) Recall that $\mathbf{\Sigma}_\alpha = \mathbf{\Sigma}\mathbf{\Phi}_\alpha\mathbf{\Psi}$. It follows that

$$\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}_\alpha\mathbf{\Psi}^{-\frac{1}{2}} = \left(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}\right)\left(\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Phi}_\alpha\mathbf{\Sigma}^{\frac{1}{2}}\right)\left(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}}\right)^{-1}. \tag{72}$$

Thus the matrix $\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}_\alpha\mathbf{\Psi}^{-\frac{1}{2}}$ on the left side and the inner matrix $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Phi}_\alpha\mathbf{\Sigma}^{\frac{1}{2}}$ on the right side are related by a *similarity transformation* and therefore share the same eigenvalues.[4] Since $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Phi}_\alpha\mathbf{\Sigma}^{\frac{1}{2}}$ is positive-definite, the eigenvalues shared by these two matrices are positive. Next, we note that the matrix $\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}_\alpha\mathbf{\Psi}^{-\frac{1}{2}}$ is symmetric, and since all of its eigenvalues are positive, it is also positive-definite. Finally, we can write

$$\mathbf{\Sigma}_\alpha = \mathbf{\Psi}^{\frac{1}{2}}\left(\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}_\alpha\mathbf{\Psi}^{-\frac{1}{2}}\right)\mathbf{\Psi}^{\frac{1}{2}}, \tag{73}$$

and since the inner matrix $\mathbf{\Psi}^{-\frac{1}{2}}\mathbf{\Sigma}_\alpha\mathbf{\Psi}^{-\frac{1}{2}}$ on the right side is positive-definite, it follows that $\mathbf{\Sigma}_\alpha$ is also positive-definite. We can now apply Lemma 5 exactly as in eq. (71) to obtain

$$(1-\alpha)\Psi_{ii}^{-1}\Sigma_{ii} \geq 1-\alpha \tag{74}$$

with a strict inequality for some j. But now, because $\alpha > 1$, we obtain the *reverse* inequality when dividing both sides by $(1-\alpha)$, and we conclude that $\Psi_{ii} \geq \Sigma_{ii}$ with a strict inequality for some j. Thus we have shown $\mathrm{D}_\alpha(p||q) \succ \mathrm{KL}(p||q)$ for $\alpha > 1$. ■

5.3 Ordering of α-divergences for $\alpha \in (0,1)$

In this section we prove the ordering of the α-divergences in Theorem 3. This proof is more technical, relying on a detailed analysis of the fixed point equations in Proposition 8.

Proof of $D_{\alpha_1}(p||q) \prec D_{\alpha_2}(p||q)$ for $0 < \alpha_1 < \alpha_2 < 1$. Let $\mathbf{\Psi}(\alpha)$ denote the diagonal covariance matrix that FG-VI estimates by minimizing $\mathrm{D}_\alpha(p||q)$ and note by inspection of $\mathrm{D}_\alpha(p||q)$ that $\mathbf{\Psi}(\alpha)$ is smooth. For $\alpha \in (0,1)$, we know from the results of Section 5.2 that $\Psi_{ii}(\alpha)$ is sandwiched between its limiting values at zero and one:

$$\frac{1}{\Sigma_{ii}^{-1}} = \lim_{\alpha\to 0^+}\Psi_{ii}(\alpha) \leq \Psi_{ii}(\alpha) \leq \lim_{\alpha\to 1^-}\Psi_{ii}(\alpha) = \Sigma_{ii}. \tag{75}$$

We also know from Lemma 5 that $\Sigma_{ii}\Sigma_{ii}^{-1} > 1$ whenever $\sum_{j\neq i}|\Sigma_{ij}| > 0$, and that in this case, the above inequalities are strict. A useful picture of this situation is shown in the left panel of Figure 3. Consider any component $\Psi_{ii}(\alpha)$ of the estimated variances that is *not* constant on the unit interval. (As shown previously, there must be at least one such component.) In this case, there are only two possibilities: either $\Psi_{ii}(\alpha)$ is strictly increasing, or it is not. If the former is always true, then it follows that $\mathrm{D}_{\alpha_1}(p||q) \prec \mathrm{D}_{\alpha_2}(p||q)$ whenever $0 < \alpha_1 < \alpha_2 < 1$. We shall prove *by contradiction* that this is indeed the case.

4. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and an invertible matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, it is well known and straightforward to show that the matrices $\mathbf{A}$ and $\mathbf{BAB}^{-1}$ have the same eigenvalues.

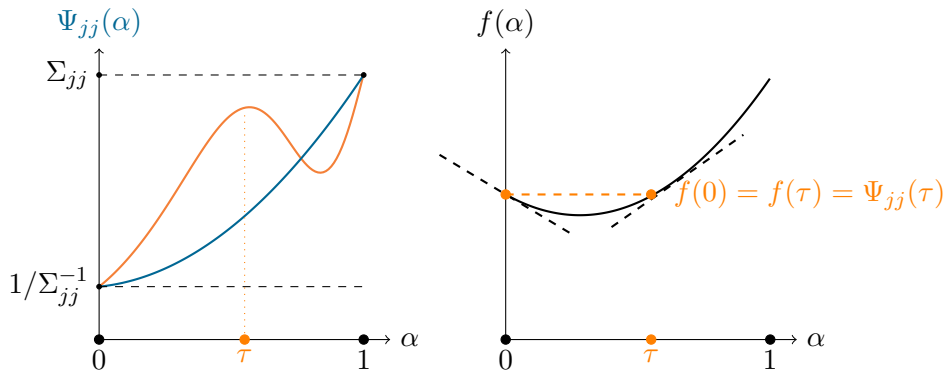

Figure 3: (Left) Either $\Psi_{jj}(\alpha)$ is strictly increasing over $\alpha \in (0,1)$, or it is not, with some minimal point τ of vanishing derivative. We prove the former by showing that no such point τ exists. (Right) The proof is based on properties of the function $f(\alpha)$ in eq. (83). The function is convex; it also satisfies $f(0) = f(\tau) = \Psi_{jj}(\tau)$ and $f'(0)f'(\tau) < 0$.

Assume that there are one or more diagonal components, $\Psi_{ii}(\alpha)$, that are neither constant nor strictly increasing over the unit interval; also, let $\mathcal{I}$ denote the set that contains the indices of these components. Since each such component must have at least one stationary point, we can also write:

$$\mathcal{I} = \Big\{ i \,\Big|\, \Sigma_{ii}\Sigma^{-1}_{ii} > 1 \text{ and } \Psi'_{ii}(\alpha) = 0 \text{ for some } \alpha \in (0,1) \Big\}. \tag{76}$$

For each such component, there must also exist some *minimal* point $\tau_i \in (0,1)$ where its derivative vanishes. We define $\tau = \min_{i\in\mathcal{I}} \tau_i$ and $j = \operatorname{argmin}_{i\in\mathcal{I}} \tau_i$, so that by definition

$$\Psi'_{ii}(\tau) \geq 0 \quad \text{for all } i, \tag{77}$$

$$\Psi'_{jj}(\tau) = 0. \tag{78}$$

To obtain the desired contradiction, we shall prove that there exists no point τ (as imagined in Figure 3) with these properties.

We start by rewriting the fixed-point equation in Proposition 8 and eq. (54) in a slightly different form,

$$\Psi_{ii}(\alpha) = [\mathbf{\Phi}(\alpha)^{-1}]_{ii} = \big[\alpha\mathbf{\Sigma}^{-1} + (1-\alpha)\Psi^{-1}(\alpha)\big]^{-1}_{ii}, \tag{79}$$

where on the right side we have explicitly indicated all sources of dependence on α. Next, we differentiate both sides of eq. (79) for $i=j$ and at $\alpha=\tau$:

$$\Psi'_{jj}(\tau) = \mathbf{e}_j^\top \left\{ \frac{d}{d\alpha}\Big[\tau\mathbf{\Sigma}^{-1} + (1-\tau)\mathbf{\Psi}^{-1}(\alpha)\Big]^{-1} + \frac{d}{d\alpha}\Big[\alpha\mathbf{\Sigma}^{-1} + (1-\alpha)\mathbf{\Psi}^{-1}(\tau)\Big]^{-1} \right\}\Bigg|_{\alpha=\tau} \mathbf{e}_j, \tag{80}$$

where on the right side we have separated out the different sources of dependence on α. Evaluating the first term on the right side, we find that

$$\mathbf{e}_j^\top \left\{ \frac{d}{d\alpha}\Big[\tau\mathbf{\Sigma}^{-1} + (1-\tau)\mathbf{\Psi}^{-1}(\alpha)\Big]^{-1} \right\}\Bigg|_{\alpha=\tau} \mathbf{e}_j = \mathbf{e}_j^\top \mathbf{\Phi}^{-1}(\tau)\mathbf{\Psi}^{-1}(\tau)\mathbf{\Psi}'(\tau)\mathbf{\Psi}^{-1}(\tau)\mathbf{\Phi}^{-1}(\tau)\,\mathbf{e}_j, \tag{81}$$

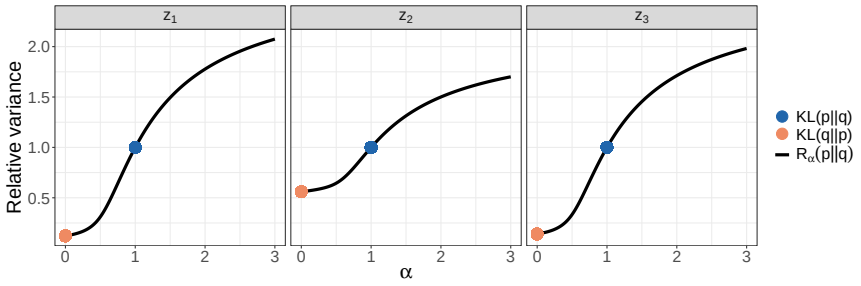

Figure 4: Marginal variances of q in eq. (4) that minimize $\mathrm{D}_\alpha(q||p)$ as a function of α. The target p is a three dimensional Gaussian. The plot shows the variances of q normalized by the variances of p. The variances of q are a strictly increasing function of α, indicating that the α-divergences are ordered. While we prove this to be the case for any Gaussian target when $\alpha \in (0,1)$, it remains an open problem to prove this when $\alpha > 1$.

and now we recognize that this term is nonnegative: of the matrices on the right side of eq. (81), note that $\mathbf{\Phi}(\alpha)$ and $\mathbf{\Psi}(\alpha)$ are positive definite and $\mathbf{\Psi}'(\tau)$ is positive semidefinite by virtue of the defining properties of τ in eqs. (77–78). Dropping this first term from the right side of eq. (80), we obtain the inequality

$$0 \geq \frac{d}{d\alpha} \left\{ \mathbf{e}_j^\top \left[\alpha \mathbf{\Sigma}^{-1} + (1-\alpha) \mathbf{\Psi}^{-1}(\tau) \right]^{-1} \mathbf{e}_j \right\} \Big|_{\alpha=\tau}, \tag{82}$$

where we have exploited that the left side of eq. (80) vanishes by the second defining property of τ in eq. (78).

Next we show that eq. (82) contains a contradiction. To do so, we examine the properties of the function defined on $[0,1]$ by

$$f(\alpha) = \mathbf{e}_j^\top \left[\alpha \mathbf{\Sigma}^{-1} + (1-\alpha) \mathbf{\Psi}^{-1}(\tau) \right]^{-1} \mathbf{e}_j. \tag{83}$$

This is, of course, exactly the function that is differentiated in eq. (82), whose right side is equal to $f'(\tau)$. This function possesses several key properties. First, we note that

$$f(0) = \mathbf{e}_j^\top \left[\mathbf{\Psi}^{-1}(\tau) \right]^{-1} \mathbf{e}_j = \Psi_{jj}(\tau) = \mathbf{e}_j^\top \left[\tau \mathbf{\Sigma}^{-1} + (1-\tau) \mathbf{\Psi}^{-1}(\tau) \right]^{-1} \mathbf{e}_j = f(\tau). \tag{84}$$

Second, we note that f is convex on the unit interval; here, we are observing the convexity of the matrix inverse (Nordström, 2011). Third, we note that

$$f'(0) = -\mathbf{e}_j^\top \mathbf{\Psi}(\tau) \left[\mathbf{\Sigma}^{-1} - \mathbf{\Psi}^{-1}(\tau) \right] \mathbf{\Psi}(\tau) \, \mathbf{e}_j = \Psi_{jj}(\tau) \left[\mathbf{\Sigma}^{-1} - \mathbf{\Psi}^{-1}(\tau) \right]_{jj} < 0. \tag{85}$$

Finally, we note that the sign of $f'(\tau)$ must *oppose* the sign of $f'(0)$; as shown in the right panel of Figure (3), this follows from the preceding properties that $f(0) = f(\tau)$, that the function f is convex on $[0,1]$, and that $f'(0) < 0$. Thus we have shown that $f'(\tau) > 0$. But this is directly in contradiction with eq. (82), which states that $0 \geq f'(\tau)$, and we are forced to conclude that no such point τ exists. This completes the proof. ■

Combining the results from Sections 5.1-5.3, we obtain a proof of Theorem 3.

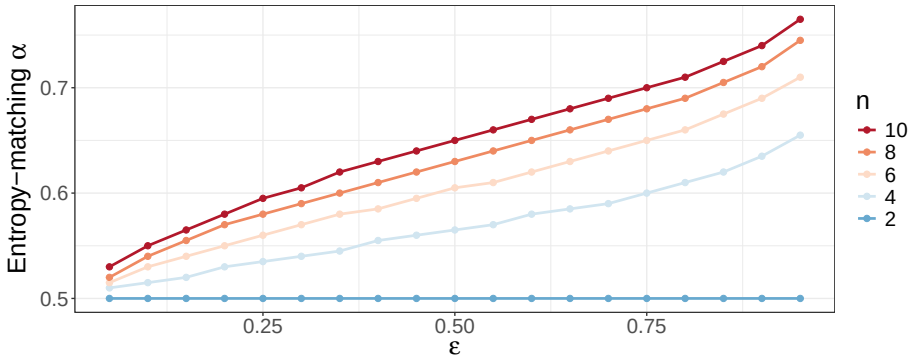

Figure 5: When p is Gaussian over $\mathbb{R}^n$, there always exists a unique $\alpha \in (0,1)$ such that the factorized approximation q minimizing $\mathrm{D}_\alpha(p||q)$ matches the entropy of p. The plots shows, however, that the entropy-matching value of α depends on the dimension and covariance structure of p. Above we vary the dimension n and constant correlation ε.

Remark 12 *The proof in this section does not generalize to α-divergences with $\alpha > 1$ because in this case the function $f(\alpha)$ in eq. (83) is not manifestly convex.*

We conjecture that the α-divergences are also ordered for $\alpha > 1$, and we provide empirical evidence for this ordering in Figure 4 where the target is a three-dimensional Gaussian. We leave a proof (or disproof) of this conjecture to future work.

5.4 Entropy-matching Solution for FG-VI

We observed in Section 4.1 that the reverse KL divergence yields a precision-matching solution for FG-VI and that the forward KL divergence yields a variance-matching solution. The ordering of divergences in Theorem 3 has another interesting consequence: it implies that some α-divergence yields a unique *entropy-matching* solution for FG-VI.

Corollary 13 *Let $q_\alpha = \mathrm{argmin}_{q \in \mathcal{Q}} D_\alpha(p||q)$. There exists a unique value $\alpha \in (0,1)$ such that q_α matches the entropy of p, or equivalently, that $|\mathbf{\Psi}(\alpha)| = |\mathbf{\Sigma}|$.*

Proof Let $\mathbf{\Psi}(\alpha)$ denote the covariance of q_α for $\alpha \in (0,1)$. Since $\Psi_{ii}(\alpha)$ is continuous with respect to α, so is $\log|\mathbf{\Psi}(\alpha)| = \sum_i \log \Psi_{ii}(\alpha)$. Moreover, from the ordering of α-divergences, we see that $\log|\Psi(\alpha)|$ is not only continuous, but strictly increasing over the unit interval $\alpha \in (0,1)$. In addition, for all components i along the diagonal, we have

$$1/\Sigma^{-1}_{ii} = \Psi_{ii}(0) \leq \Psi_{ii}(\alpha) \leq \Psi_{ii}(1) = \Sigma_{ii}, \tag{86}$$

with q_α interpolating smoothly between the precision-matching and the variance-matching approximations of FG-VI. From the impossibility result in Theorem 1, the precision-matching approximation (at $\alpha = 0$) underestimates the entropy of p, while the variance-matching approximation (at $\alpha = 1$) overestimates it. From continuity and strict monotonicity, it follows

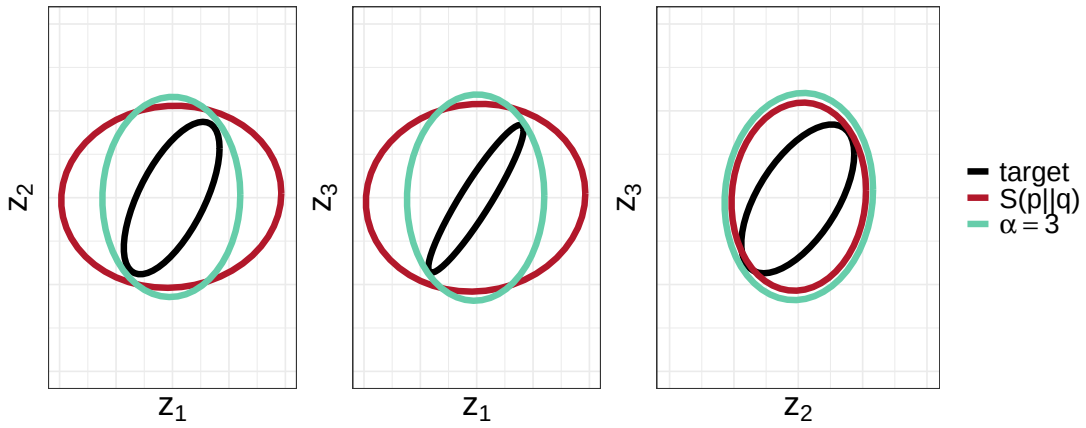

Figure 6: Example of a Gaussian target in dimension $d = 3$ where the variances obtained by minimizing $S(p||q)$ and $\mathrm{D}_\alpha(p||q)$ are not ordered. The non-ordering holds for any $\alpha \gtrapprox 2$ and is most clearly seen for $\alpha = 3$. The figure shows the two-dimensional projections of the variational fits.

that there exists a unique $\alpha \in (0, 1)$ such that $|\boldsymbol{\Psi}(\alpha)| = |\boldsymbol{\Sigma}|$, or equivalently, such that q_α matches the entropy of p. ■

The proof of Corollary 13 demonstrates the existence an entropy-matching solution, but it is not a constructive proof. Unfortunately, there does not exist a single value of α for which $|\boldsymbol{\Psi}(\alpha)| = |\boldsymbol{\Sigma}|$, even when p is Gaussian. We show this numerically for the case where the target covariance $\boldsymbol{\Sigma}$ has unit diagonal elements and constant off-diagonal elements $\varepsilon \in (0, 1)$. As seen in Figure 5, the entropy-matching value of α (obtained by a grid search) changes as we vary the amount of correlation and the dimension of the problem.

5.5 Non-ordering of the Score-based and α-divergences for $\alpha > 1$

Given that both $S(p||q)$ and $\mathrm{D}_\alpha(p||q)$ dominate $\mathrm{KL}(p||q)$, it is natural to ask whether an ordering exists between the score-based and α-divergences when $\alpha > 1$. Here we produce a counterexample to show that no such ordering exists. Searching through randomly generated covariance matrices, we find an example in dimension $d = 3$ where the variances that minimize $\mathrm{D}_\alpha(p||q)$ and $S(p||q)$ are not ordered for $\alpha \in [2, 3]$; Figure 6 illustrates this non-ordering when $\alpha = 3$. The target covariance in this example is (within two digits)

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.66 & 0.93 \\ 0.66 & 1 & 0.59 \\ 0.93 & 0.59 & 1 \end{pmatrix}. \tag{87}$$

Table 4 reports the variances of q obtained by minimizing $S(p||q)$ and $\mathrm{D}_\alpha(p||q)$ for this example when $\alpha = 2$ and $\alpha = 3$. Note that the score-based divergence yields the largest estimate (shown in bold) of the variance Ψ_{11}, but the α-divergence (with $\alpha = 3$) yields the largest estimate of the variances Ψ_{22} and Ψ_{33}. Thus these divergences are not ordered.

Divergence	Ψ_{11}	Ψ_{22}	Ψ_{33}
$S(p\|\|q)$	**6.18**	1.40	1.63
$\mathrm{D}_{\alpha=2}(p\|\|q)$	1.78	1.50	1.71
$\mathrm{D}_{\alpha=3}(p\|\|q)$	2.07	**1.70**	**1.98**

Table 4: Variances estimated by minimizing different divergences for the Gaussian target in eq. (87). In bold, the largest optimal variance obtained across all divergences. This example shows that $S(p||q)$ and $\mathrm{D}_\alpha(p||q)$ cannot be ordered for $\alpha > 1$.

6. Numerical Experiments

Does the ordering of divergences hold when FG-VI is applied to non-Gaussian targets? We study this question empirically on a range of target distributions. Several of these models are taken from Bayesian analysis problems, where the target p is only known up to a normalizing constant.

6.1 Optimization Procedure

When p is not Gaussian, some divergences in Table 1 are harder to minimize numerically than others. We use the following approaches.

Optimization of KL$(q||p)$. We minimize $\mathrm{KL}(q||p)$ by equivalently maximizing the evidence lower bound (ELBO), and we estimate the ELBO via Monte Carlo with draws from q, a procedure at the core of "black-box" VI (e.g. Kucukelbir et al., 2017).

Optimization of $D_\alpha(q||p)$. We use a similar procedure to optimize the α-divergence with $\alpha \in \{0.1, 0.5, 2\}$; however, this procedure is more fraught[5] as Monte Carlo estimators of the α-divergence and its gradients suffer from a large variance, especially as the dimension increases and as α approaches 1 (Geffner and Domke, 2021). For $\alpha = 2$, we find it helpful to construct Monte Carlo estimators with *oracle samples* from p, rather than q, when such samples are available (e.g., for simpler targets).

Optimization of KL$(p||q)$. We minimize $\mathrm{KL}(p||q)$ by choosing q in eq. (4) to match the mean and marginal variances of p. When possible, we calculate these statistics of p analytically; otherwise we estimate them by long runs of Markov chain Monte Carlo.

Optimization of $S(q||p)$. We minimize $\mathrm{S}(q||p)$ by adapting a recent *batch-and-match* (BaM) method inspired by proximal point algorithms (Cai et al., 2024b). BaM was originally developed for Gaussian variational families with dense covariance matrices, but here we implement BaM updates for FG-VI (derived in Appendix A). We do not empirically evaluate the forward score-based divergence as we do not have a reliable method to minimize $\mathrm{S}(p||q)$.

5. For more discussion on implementing VI with α-divergences, see e.g., Hernandez-Lobato et al. (2016); Li and Turner (2016); Dieng et al. (2017); Daudel et al. (2021, 2023).

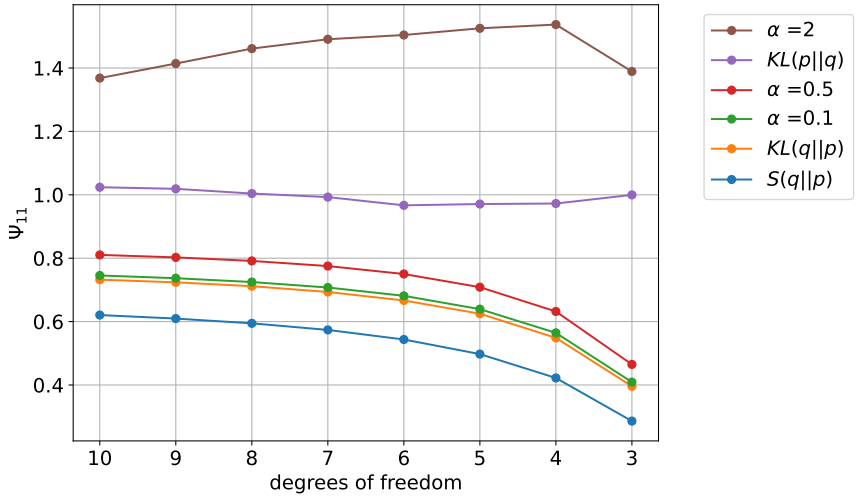

Figure 7: Variance estimated by FG-VI when targeting a two-dimensional Student-t with varying degrees of freedom d_f. The ordering of divergences for Gaussian targets is observed across a wide range of Student-t targets, from $d_f = 10$, where the target is nearly Gaussian, to $d_f = 3$, where the target is heavy tailed. (A similar plot is obtained for Ψ_{22}.)

Each method is implemented in `Python` and uses the `JAX` library to calculate derivatives (Bradbury et al., 2018). Optimization is performed with the `Adam` optimizer (Kingma and Ba, 2015). As a baseline, we run the optimizer for 500 iterations, and at each iteration, we use $B = 10,000$ draws from q (or p when using oracle samples) to estimate the relevant objective function and its gradient. This large number of draws is overkill for many problems, but it helps stabilize the optimization of α-divergences. In addition, we parallelize many calculations on GPU, keeping run times short.

For certain targets, we find it necessary to fine-tune the optimizer in order to avoid numerical instability. We do so by adjusting the learning rate of `Adam`, the number B of Monte Carlo draws, and the number of iterations. This proves particularly important when minimizing the α-divergences. Even so, we are only able to optimize the α-divergences for low-dimensional targets ($D \leq 10$). Code to reproduce the experiments can be found at `https://github.com/charlesm93/VI-ordering`.

6.2 Targets with Heavy Tails and Skew

First we experiment with non-Gaussian targets in two dimensions where we can systematically vary the degree of non-Gaussianity.

The Student-t distribution has heavier tails than the Gaussian, particularly for low degrees of freedom (d_f). We experiment with FG-VI on a wide range of Student-t targets, with d_f ranging from 10 (near Gaussian) to 3 (heavy-tailed), and we find that the ordering of variances in Theorem 3, derived for Gaussian targets, is preserved when the target is a Student-t distribution; see Figure 7.

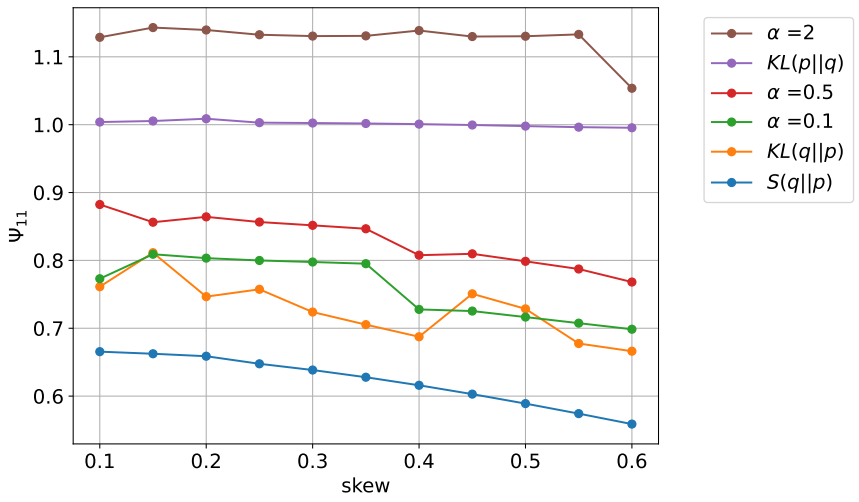

Figure 8: Variance of FG-VI when targeting a two-dimensional skewed Normal with skewness s. For $s = 0.1$, the target is near Gaussian and s increases, it becomes more asymmetric. (A similar plot is obtained for Ψ_{22}.)

The multivariate Skew-Normal distribution (Azzalini and Dalla Valle, 1996) has a skewness parameter s: for $s = 0$, the Skew-Normal exhibits no skew and reduces to a Gaussian, whereas for $s \geq 0.5$ it is heavily asymmetric. For skewed targets in this family, we find that the ordering of variances in Theorem 3 is generally preserved even as s increases; see Figure 8. However, the ordering between $\mathrm{KL}(q||p)$ and $\mathrm{D}_\alpha(q||p)$ for $\alpha = 0.1$ is sometimes violated. These violations are slight and may be due to numerical error in the optimization of these divergences.

6.3 Models from the `Inference Gym`

We next experiment with target distributions from the `inference gym`, a curated library of diverse models for the study of inference algorithms (Sountsov et al., 2020). We describe these target distributions briefly in Table 5 and provide more details in Appendix B.

Some of these target distributions are far from Gaussian, and so the ordering of variances in Theorem 3 is not guaranteed to hold. We may also observe slight violations of this ordering due to noisy solutions that arise from stochastic optimization.

Violations of the ordering in Theorem 3 are observed for the Eight Schools model, a hierarchical model that exhibits a funnel geometry; see Figure 9. In this example, most of the marginal variances are not ordered in the same way as predicted for Gaussian targets, although in many cases, the violations are slight. (The true marginal variances in this model were estimated by long runs of Markov chain Monte Carlo.)

When p is not Gaussian, an alternative ordering of divergences for FG-VI is suggested by the estimates they yield of the joint entropy. As shown in the final column of Table 5 and in Figure 10, the ordering in Theorem 3 *does* correspond—across all of the non-Gaussian targets in our experiments—to the ordering of entropies estimated by FG-VI. In these ex-

Model	n	Details	Variances	Entropy
Gaussian	10	Full-rank covariance matrix.	✓	✓
Student-t	2	Heavy tail.	✓	✓
Skew-Normal	2	Asymmetric.	?	?
Rosenbrock	2	High curvature.	✓	✓
Eight Schools	10	Small hierarchical model with funnel geometry.	✗	✓
German Credit	25	Logistic regression.	✓	✓
Radon Effect	91	Large hierarchical model with funnel geometry.	✗	✓
Stochastic Volatility	103	Stochastic volatility model.	✗	✓

Table 5: Target distributions p with dimension n. The two right-most columns indicate respectively whether the divergences are ordered according to the marginal variances (i.e., the ordering in Theorem 3 is preserved) or according to the entropy (equivalently, the generalized variance). Question marks indicate that the ordering in Theorem 3 is only slightly violated, possibly due to noise in the stochastic optimization.

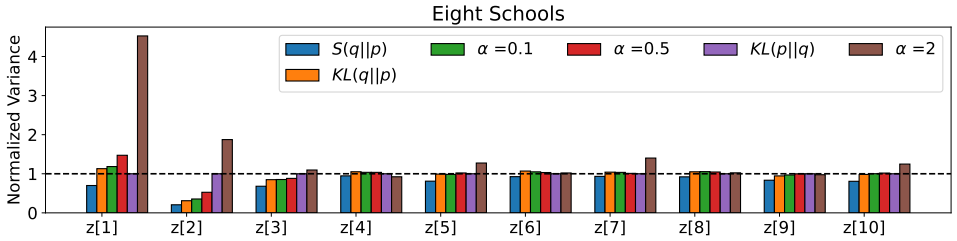

Figure 9: Variances from FG-VI when targeting the posterior distribution of the Eight Schools model. Some of the variances are ordered in the same way as predicted for Gaussian targets, but most are not. In this case, the target posterior is highly non-Gaussian, exhibiting a funnel shape, and thus strongly violating the assumptions of Theorem 3. The variances reported here, unlike those appearing in Theorem 3, are also noisy estimates from a stochastic optimization.

periments, however, we are not able to compare the entropies estimated by FG-VI to the actual entropies. To compute the entropy of non-Gaussian distributions, it is necessary to estimate their normalizing constants; this can be done by invoking a Gaussian-like approximation, as in bridge sampling (Meng and Schilling, 2002; Gronau et al., 2020). However, it seems inadequate to compare a theory developed for Gaussian targets to an empirical benchmark that relies on a Gaussian approximation.

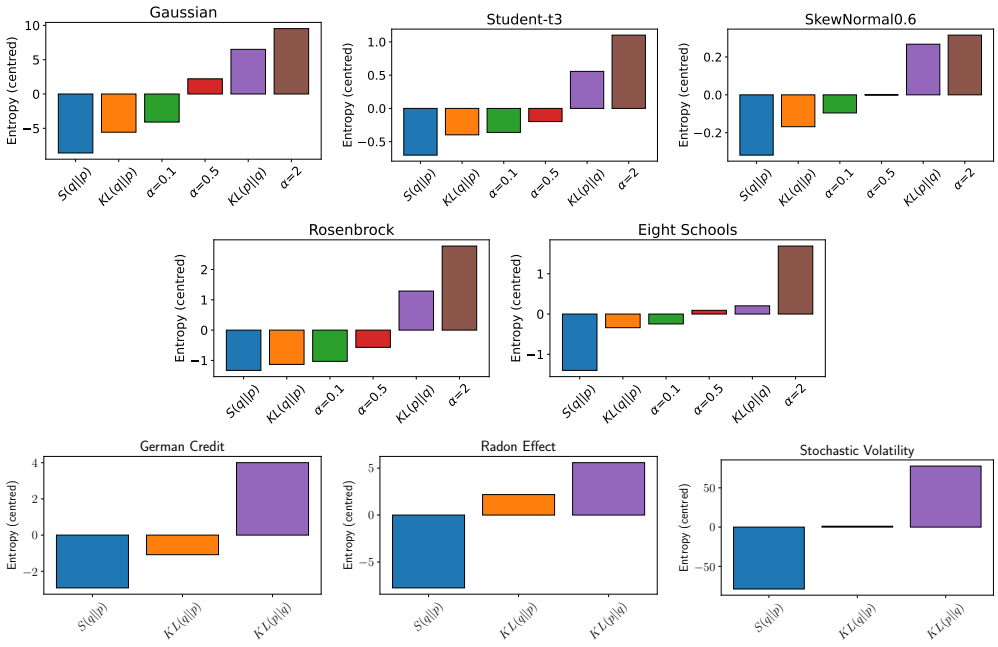

Figure 10: Ordering of entropy across targets. The zero on each vertical axis corresponds to the average of the entropies estimated by the approximations in each panel.

7. Discussion

Divergences over probability spaces are ubiquitous in the statistics and machine learning literature, but it remains challenging to reason about them. This challenge is salient in VI, where the divergence does not serve merely as an object for theoretical study but also bears directly on the algorithmic implementation of the method. Our work reveals how the choice of divergence can align (or misalign) with intelligible inferential goals, and it is in line with a rich and recent body of work on assessing the quality of VI (e.g, Yao et al., 2018; Wang and Blei, 2018; Huggins et al., 2020; Dhaka et al., 2021; Biswas and Mackey, 2024). A distinguishing feature of our paper is that it analyzes multiple and competing inferential goals—goals which cannot be simultaneously met and which are each best served by a different divergence.

Our results can inform the choice of divergence to minimize in eq. (1) in tandem with the use of a factorized approximation. This choice will in turn depend on the problem at hand. We discuss some examples:

- **Bayesian statistics.** Here, uncertainty is primarily quantified by the marginal variances Σ_{ii}, and these are matched by minimizing $\mathrm{KL}(p||q)$. In general, however, this divergence is not straightforward to minimize, though strategies such as expectation-propagation can in some limits attain an optimal solution (Vehtari et al., 2020). It is much more straightforward to minimize divergences such as $\mathrm{KL}(q||p)$ and $S(q||p)$, but for FG-VI we have seen how marginal variances can be significantly underestimated by factorized approximations that minimize these divergences. This provides further

motivation for: (i) corrective procedures (Giordano et al., 2018) to adjust the estimate of marginal variances from F-VI; (ii) VI methods with richer (non-factorized) families (e.g Opper and Archambeau, 2009; Dhaka et al., 2021; Zhang et al., 2022); (iii) altogether alternative methods such as Markov chain Monte Carlo (MCMC; Robert and Casella, 2004), when allowed by a computational budget.

- **Pre-conditioning of Markov chain Monte Carlo.** Several MCMC algorithms benefit from transforming the latent variables over which the Markov chain is constructed. This technique is known as pre-conditioning. In Hamiltonian Monte Carlo (HMC; Neal, 2012; Betancourt, 2018) the mass matrix serves as a pre-conditioner, and one typically restricts it to be diagonal for computational reasons. Neal (2012) recommends setting the diagonal elements of the (inverted) mass matrix to early estimates of the marginal variances, an idea implemented in popular adaptive HMC algorithms (e.g Hoffman and Gelman, 2014). Recently, it was proposed that this approach might not be optimal (Hird and Livingstone, 2023), and that a better choice for a diagonal mass matrix might be based on estimates of the marginal precisions (Tran and Kleppe, 2024). This suggests a novel application of F-VI, one where it is used to tune HMC's mass matrix by minimizing $\mathrm{KL}(q||p)$. Investigation of this method is left as future work.

- **Information theory.** Here uncertainty is primarily measured by the entropy. In certain limiting cases, the entropy may be accurately estimated by minimizing $\mathrm{KL}(q||p)$ (Parisi, 1988; Margossian and Saul, 2023); in general, however, this is not the case, as we have shown analytically when p and q are Gaussian. One possible strategy in this setting is to instead minimize the α-divergence, for $\alpha \in (0, 1)$. While there always exists an α to match the entropy, we have shown that this value depends on p, and so it remains an open problem to find an actionable divergence for entropy estimation.

VI is also deployed in applications where the primary goal is not to quantify uncertainty. A prominent example is latent variable models, where VI is used to approximate a marginal likelihood. In this context, the KL and α-divergences have been ordered based on the (lower or upper) bound they provide on the marginal likelihood (Li and Turner, 2016; Dieng et al., 2017). (Interestingly, this ordering agrees with the one provided in this paper, even though it is based on a different criterion.) This type of ordering cannot be extended to score-based divergences, however, as they do not provide bounds on the marginal likelihood.

We have seen that the results of VI can be improved, for certain inferential goals, by choosing a particular divergence. A complementary and more common approach is to use a richer family $\mathcal{Q}$ of variational approximations (e.g, Hoffman and Blei, 2015; Johnson et al., 2016; Dhaka et al., 2021). The deficits of FG-VI may be addressed, for instance, by Gaussian variational approximations with full covariance matrices (Opper and Archambeau, 2009; Cai et al., 2024b). In high dimensions, Dhaka et al. (2021) argue that the results of VI can be improved more easily by optimizing $\mathrm{KL}(q||p)$ over a richer variational family than by optimizing certain alternative divergences. Nevertheless, even as the variational family becomes richer, it is not likely to contain the target distribution p, and therefore the approximating distribution q must in some way be compromised. For example, if $\mathcal{Q}$ is a family of symmetric distributions and the target p is asymmetric, then an approximation

will not be able to match both the mode *and* the mean of the target. It stands to reason that further impossibility theorems, in the spirit of Theorem 1, can be derived to elucidate these trade-offs.

Acknowledgments

We thank David Blei, Diana Cai, Robert Gower, Chirag Modi, and Yuling Yao for helpful discussions. We are grateful to anonymous reviewers from the conference on Uncertainty in Artificial Intelligence for their comments on a previous paper (Margossian and Saul, 2023), which provided some of the motivation for this manuscript. Finally we thank the editor, Barbara Engelhardt, and two anonymous reviewers for their feedback on this manuscript.

Appendix A. Batch and Match Algorithm for FG-VI

Batch and Match (BaM) (Cai et al., 2024b, BaM) is an iterative algorithm for VI that attempts to estimate and minimize the reverse score-based divergence, $\mathrm{S}(q||p)$. At each iteration, BaM minimizes a regularized objective function

$$\mathcal{L}^{\mathrm{BaM}}(q) = \widehat{\mathcal{D}}_{q_t}(q||p) + \frac{2}{\lambda_t}\mathrm{KL}(q||q_t), \tag{88}$$

where

- $q_t \in \mathcal{Q}$ is a current approximation,
- $\hat{\mathcal{D}}_{q_t}(q||p)$ is a Monte Carlo estimator of the score-based divergence

$$\hat{\mathcal{D}}_{q_t}(q||p) = \frac{1}{B}\sum_{b=1}^{B} ||\nabla \log q\big(\mathbf{z}^{(b)}\big) - \nabla \log p\big(\mathbf{z}^{(b)}\big)||^2_{\boldsymbol{\Psi}}, \tag{89}$$

  using draws $\mathbf{z}^{(b)} \sim q_t$.

- $\lambda_t > 0$ is the learning rate, or step size, at the t^{th} iteration.

Cai et al. (2024b, Appendix C) derive the BaM updates which minimize eq. (88) when $\mathcal{Q}$ is the family of Gaussian distributions with *dense* covariance matrices. In this appendix, we derive the analogous updates to minimize eq. (88) when $\mathcal{Q}$ is the family of Gaussian distributions with *diagonal* covariance matrices. These were the updates used for the experiments in Section 6 of the paper.

We follow closely the derivation in Cai et al. (2024b), noting only the essential differences for the case of FG-VI. We use $\mathbf{g}^{(b)} = \nabla \log p\big(\mathbf{z}^{(b)}\big)$ as shorthand for the score at the b^{th} sample. As before, the following averages need to be computed at each iteration:

$$\bar{\mathbf{z}} = \frac{1}{B}\sum_{b=1}^{B}\mathbf{z}^{(b)}, \quad \bar{\mathbf{g}} = \frac{1}{B}\sum_{b=1}^{B}\mathbf{g}^{(b)}. \tag{90}$$

For FG-VI, we instead compute *diagonal* matrices $\mathbf{C}$ and $\boldsymbol{\Gamma}$ with nonnegative elements

$$C_{ii} = \frac{1}{B}\sum_{b=1}^{B}\Big(z_i^{(b)} - \bar{z}_i\Big)^2, \quad \Gamma_{ii} = \frac{1}{B}\sum_{b=1}^{B}\Big(g_i^{(b)} - \bar{g}_i\Big)^2. \tag{91}$$

Let q_{t+1} denote the minimizer of eq. (88), with mean $\boldsymbol{\nu}_{t+1}$ and diagonal covariance matrix $\boldsymbol{\Psi}^{t+1}$. The update for the mean takes the same form as in the original formulation of BaM:

$$\boldsymbol{\nu}^{t+1} = \frac{\lambda_1}{1+\lambda_t}\left(\bar{\mathbf{z}} + \boldsymbol{\Psi}^{t+1}\bar{\mathbf{g}}\right) + \frac{1}{1+\lambda_t}\boldsymbol{\nu}^t. \tag{92}$$

To obtain the covariance update, we substitute this result into eq. (88) and differentiate with respect to the *diagonal* elements of $\boldsymbol{\Psi}^{t+1}$. For FG-VI, we obtain a simple quadratic equation for each updated variance:

$$\left(\Gamma_{ii} + \frac{1}{1+\lambda_t}\bar{g}_i^2\right)\left(\Psi_{ii}^{t+1}\right)^2 + \frac{1}{\lambda_t}\Psi_{ii}^{t+1} - \left(C_{ii} + \frac{1}{\lambda_t}\Psi_{ii}^t + \frac{(\nu_i^t - \bar{z}_i)^2}{1+\lambda_t}\right) = 0. \tag{93}$$

Eq. (93) always admits a positive root, and for FG-VI, this positive root is the BaM update for the i^{th} diagonal element of $\boldsymbol{\Psi}^{t+1}$.

At a high level, BaM works the same way with diagonal covariance matrices as it does with full covariance matrices. Each iteration involves a *batch* step which draws the B samples from q_t and a *match* step that performs the calculations in eqs. (90-93). We find a constant learning rate $\lambda_t = 1$ to work well for the experiments in Section 6.

Appendix B. Models from the `Inference Gym`

In this appendix, we provide details about the models and data sets from the `inference gym` used in Section 6.

Rosenbrock Distribution. (n = 2) A transformation of a normal distribution, with non-trivial correlations (Rosenbrock, 1960). The contour plots of the density function have the shape of a crescent moon. The joint distribution is given by:

$$p(z_1) = \text{normal}(0, 10) \tag{94}$$

$$p(z_2|z_1) = \text{normal}(0.03(z_1^2 - 100), 1). \tag{95}$$

We use FG-VI to approximate $p(z_1, z_2)$.

Eight Schools. ($n = 10$) A Bayesian hierarchical model of the effects of a test preparation program across 8 schools (Rubin, 1981):

$$p(\mu) = \text{normal}(5, 3^2) \tag{96}$$

$$p(\tau) = \text{normal}^+(0, 10) \tag{97}$$

$$p(\theta_i|\mu, \sigma) = \text{normal}(\mu, \tau^2) \tag{98}$$

$$p(y_i|\theta_i) = \text{normal}(\theta_i, \sigma_i^2), \tag{99}$$

where normal$^+$ is a normal distribution truncated at 0. The observations are $\mathbf{x} = (y_{1:8}, \sigma_{1:8})$ and the latent variables are $\mathbf{z} = (\mu, \tau, \theta_{1:8})$. We use VI to approximate the posterior distribution $p(\mathbf{z}|\mathbf{x})$.

German Credit. ($n=25$) A logistic regression applied to a credit data set (Dua and Graff, 2017), with covariates $\mathbf{x} = x_{1:24}$, observations $\mathbf{y} = y_{1:24}$, coefficients $\mathbf{z} = z_{1:24}$ and an intercept z_0:

$$p(\mathbf{z}) = \text{normal}(0, \mathbf{I}_n), \tag{100}$$

$$p(y_i|\mathbf{z}, \mathbf{x}) = \text{Bernoulli}\left(\frac{1}{e^{-\mathbf{z}^T\mathbf{x}_i - z_0}}\right) \tag{101}$$

We use FG-VI to approximate $p(\mathbf{z}|y_i, \mathbf{x})$.

Radon Effect. ($n = 91$) A hierarchical model to measure the Radon level in households (Gelman and Hill, 2007). Here, we restrict ourselves to data from Minnesota. For each household, we have three covariates, $\mathbf{x}_i = x_{i,1:3}$, with corresponding coefficients $\mathbf{w} = w_{1:3}$, and a county level covariate, $\theta_{j[i]}$, where $j[i]$ denotes the county of the i^{th} household. The model is:

$$p(\mu) = \text{normal}(0, 1) \tag{102}$$

$$p(\tau) = \text{Uniform}(0, 100) \tag{103}$$

$$p(\theta_j|\mu, \tau) = \text{normal}(\mu, \tau^2) \tag{104}$$

$$p(w_k) = \text{normal}(0, 1) \tag{105}$$

$$p(\sigma) = \text{Uniform}(0, 100) \tag{106}$$

$$p(\log y_i) = \text{normal}(\mathbf{w}^T\mathbf{x}_i + \theta_{j[i]}, \sigma^2). \tag{107}$$

We use FG-VI to approximate $p(\mu, \tau, \theta, \mathbf{w}, \sigma|\mathbf{x}, y)$.

Stochastic volatility. (n=103) A time series model with 100 observations. The original model by Kim et al. (1998) is given by:

$$p(\sigma) = \text{Cauchy}^+(0, 2) \tag{108}$$

$$p(\mu) = \text{exponential}(1) \tag{109}$$

$$p((\phi + 1)/2) = \text{Beta}(20, 1.5) \tag{110}$$

$$p(z_i) = \text{normal}(0, 1) \tag{111}$$

$$h_1 = \mu + \sigma z_1/\sqrt{1 - \phi^2} \tag{112}$$

$$h_{i>1} = \mu + \sigma z_i + \phi(h_{i-1} - \mu) \tag{113}$$

$$p(y_i \mid h_i) = \text{normal}(0, \exp(h_i/2)). \tag{114}$$

The original model was developed for a time series of 3000 observations, but we only work with the first 100 observations. For these observations, we use FG-VI to model $p(\sigma, \mu, \phi, \mathbf{z}|\mathbf{y})$.

References

Adelchi Azzalini and Alessandra Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, 2000.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434v1*, 2018.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Christopher M. Bishop, David Spiegelhalter, and John Winn. VIBES: A variational inference engine for Bayesian networks. *Advances Neural Information Processing Systems 15*, pages 793–800, 2002.

Niloy Biswas and Lester Mackey. Bounding Wasserstein distance with couplings. *Journal of the American Statistical Association*, 2024. DOI: 10.1080/01621459.2023.2287773.

David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55:77–84, 2012.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL `https://arxiv.org/abs/1601.00670`.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Jacob Burbea. The convexity with respect to Gaussian distributions of divergences of order α. *Utilitas Mathematica*, 26:171–192, 1984.

Diana Cai, Chirag Modi, Charles C. Margossian, Robert M. Gower, David M. Blei, and Lawrence K. Saul. EigenVI: Score-based variational inference with orthogonal function expansions. In *Advances in Neural Information Processing Systems 37*, pages 132691–132721, 2024a.

Diana Cai, Chirag Modi, Loucas Pillaud-Vivien, Charles C. Margossian, Robert M. Gower, David M. Blei, and Lawrence K. Saul. Batch and match: Black-box variational inference with a score-based divergence. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5258–5297. PMLR, 2024b.

Aoxiang Chen, David J Nott, and Linda SL Tan. Weighted Fisher divergence for high-dimensional Gaussian variational inference. *arXiv:2503.04246*, 2025.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.

Thomas A. Courtade. Links between the logarithmic Sobolev inequality and the convolution inequalities for entropy and Fisher information. *arXiv:1608.05431*, 2016.

Kamélia Daudel, Randal Douc, and Francois Roueff. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *The Annals of Statistics*, 49:2250–2270, 2021.

Kamélia Daudel, Joe Benton, and Arnaud Doucet. Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics. *Journal of Machine Learning Research*, 24:1–83, 2023.

Akash Kumar Dhaka, Alejandro Catalina, Manushi Welandawe, Michael Riis Andersen, Jonathan H Huggins, and Aki Vehtari. Challenges and opportunities in high-dimensional variational inference. In *Advances in Neural Information Processing Systems 34*, pages 7787–7798, 2021.

Adji B. Dieng, Dustin Tran, R Ranganath, J Paisly, and David M. Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems 30*, pages 2732–2741, 2017.

Dheera Dua and C. Graff. UCL machine learning repository, 2017. URL `http://archive.ics.ucl.edu/ml`.

Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3650–3659, 2021.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel-Hierarchical Models*. Cambridge University Press, 2007.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Ark Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science, 2013.

Alison Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70, 2002.

Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

Rian Giordano, Tamara Broderick, and Michael I. Jordan. Covariances, robustness, and variational Bayes. *Journal of Machine Learning Research*, 19:1–49, 2018.

Ryan Giordano, Martin Ingram, and Tamara Broderick. Black box variational inference with a deterministic objective: Faster, more accurate, and even more black box. *Journal of Machine Learning Research*, 25:1–39, 2023.

Quantin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92, 2020.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernández-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1511–1520, 2016.

Max Hird and Samuel Livingstone. Quantifying the effectiveness of linear preconditionning in Markov chain Monte Carlo. *arXiv:2312.04898*, 2023.

Tomáš Hobza, Domingo Morales, and Leandro Pardo. Rényi statistics for testing equality of autocorrelation coefficients. *Statistical Methodology*, 6:424–436, 2009.

Matthew D. Hoffman and David M. Blei. Stochastic structured variational inference. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 361–369, 2015.

Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15: 1593–1623, 2014.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Jonathan H. Huggins, Mikołaj Kasprzak, Trevor Campbell, and Tamara Broderick. Validated variational inference via practical posterior error bounds. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pages 1792–1802, 2020.

Matthew J. Johnson, David K. Duvenaud, Alex Wiltschko, Ryan P. Adams, and Sandeep R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems 29*, pages 2946–2954, 2016.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Sangjoon Kim, Neil Shepard, and Siddhartha Chib. Stochastic volatility: likelihood inference and comparison with arch models. *The Review of Economic Studies*, 65:361–393, 1998.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. *Internation Conference on Learning Representations*, 2014.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18: 1–45, 2017.

Yingzhen Li and Richard E. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems 29*, pages 1073–1081, 2016.

Friedrich Liese and Igor Vajda. *Convex Statistical Distances*. Teubner, Leipzig, 1987.

David J.C. MacKay. *Information theory, inference, and learning algorithms*. 2003.

Charles C. Margossian and David M. Blei. Amortized variational inference: When and why? In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, pages 2434–2449, 2024.

Charles C. Margossian and Lawrence K. Saul. The shrinkage-delinkage 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, 2023.

Xiao-Li Meng and Stephen Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11:552–586, 2002. doi: doi:10.1198/106186002457.

Tom Minka. Divergence measures and message passing. *Technical Report MSR-TR-2005-173*, 2005.

Chirag Modi, Diana Cai, and Lawrence K. Saul. Batch, match, and patch: Low-rank approximations for score-based variational inference. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pages 4510–4518. PMLR, 2025.

Christian A. Naesseth, Fredrik Lindsten, and David M. Blei. Markovian score climbing: Variational inference with KL($p||q$). In *Advances in Neural Information Processing Systems 34*, pages 15499–15510, 2020.

Radford M. Neal. MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. CRC Press, 2012.

Kenneth Nordström. Convexity of the inverse and Moore-Penrose inverse. *Linear Algebra and Its Applications*, 434:1489–1512, 2011.

Manfred Opper and Cédric Archambeau. The variational Gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

Georgio Parisi. *Statistical Field Theory*. Addison-Wesley, 1988.

Christian P Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

Howard H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *Computer Journal*, 3:175–184, 1960.

Donald B. Rubin. Estimation in parallelized randomized experiments. *Journal of Educational Statistics*, 6:377–400, 1981.

Pavel Sountsov, Alexey Radul, and contributors. Inference gym, 2020. URL `https://pypi.org/project/inference_gym`.

Jakub M. Tomczak. *Deep Generative Modeling*. Springer, 2022.

Jimmy H. Tran and Tore S. Kleppe. Tuning diagonal scale matrices for HMC. *arxiv2403.07495*, 2024.

Richard E. Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. In David Barber, A. Taylan Cemgil, and Silvia Chiappa, editors, *Bayesian Time Series Models*, chapter 5, pages 109–130. Cambridge University Press, 2011.

Aki Vehtari, Andrew Gelman, Tuomas Sivula, Pasi Jylanki, Dustin Tran, Swupnil Sahai, Paul Blomstedt, John P. Cunningham, Schiminovich, and Christian P. Robert. Expectation propagation as a way of life: A framework for Bayesian inference on partitioned data. *Journal of Machine Learning*, 21:1–53, 2020.

Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 2024. To appear.

Martin J. Wainwright and Michael I. Jordan. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Bo Wang and Donald M. Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, R5:373 – 380, 2005.

Yixin Wang and David M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114:1147–1161, 2018.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5577–5586, 2018.

Lu Zhang, Bob Carpenter, Andrew Gelman, and Aki Vehtari. Pathfinder: Parallel quasi-Newton variational inference. *Journal of Machine Learning Research*, 23(306):1–49, 2022.