The Shrinkage-Delinkage Trade-off: An Analysis of Factorized Gaussian Approximations for Variational Inference

Charles C. Margossian[1] **Lawrence K. Saul**[1]

[1]Center for Computational Mathematics, Flatiron Institute, New York, NY, USA

Abstract

When factorized approximations are used for variational inference (VI), they tend to underestimate the uncertainty—as measured in various ways—of the distributions they are meant to approximate. We consider two popular ways to measure the uncertainty deficit of VI: (i) the degree to which it underestimates the componentwise variance, and (ii) the degree to which it underestimates the entropy. To better understand these effects, and the relationship between them, we examine an informative setting where they can be explicitly (and elegantly) analyzed: the approximation of a Gaussian, p, with a dense covariance matrix, by a Gaussian, q, with a diagonal covariance matrix. We prove that q always underestimates both the componentwise variance and the entropy of p, *though not necessarily to the same degree*. Moreover we demonstrate that the entropy of q is determined by the trade-off of two competing forces: it is decreased by the shrinkage of its componentwise variances (our first measure of uncertainty) but it is increased by the factorized approximation which delinks the nodes in the graphical model of p. We study various manifestations of this trade-off, notably one where, as the dimension of the problem grows, the per-component entropy gap between p and q becomes vanishingly small even though q underestimates every componentwise variance by a constant multiplicative factor. We also use the shrinkage-delinkage trade-off to bound the entropy gap in terms of the problem dimension and the condition number of the correlation matrix of p. Finally we present empirical results on both Gaussian and non-Gaussian targets, the former to validate our analysis and the latter to explore its limitations.

1 INTRODUCTION

Variational inference (VI) is a popular methodology for approximate Bayesian inference [Jordan et al., 1999, Wainwright and Jordan, 2008, Blei et al., 2017]. Given a target distribution, p, VI searches for a tractable distribution, $q \in \mathcal{Q}$, that minimizes the Kullback-Leibler (KL) divergence to p. A common choice for $\mathcal{Q}$ is to use a family of factorized distributions. The KL-divergence can then be optimized in a scalable manner for high-dimensional distributions [Wainwright and Jordan, 2008], which is crucial, for instance, to train models such as variational auto-encoders over large data sets [Kingma and Welling, 2013].

Factorized VI has its roots in the mean-field approximations to certain Gibbs distributions from statistical physics [Parisi, 1988, MacKay, 2003]. In this approach, the approximating distribution is modeled as

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

In most applications, the target distribution $p(\mathbf{z})$ does <u>not</u> factorize. By its very nature, factorized VI cannot estimate the correlations between different elements of $\mathbf{z}$. A more subtle shortcoming of factorized VI is that it also fails to correctly estimate the marginal distributions, $p(z_i)$. This failure typically manifests as an approximation q with an uncertainty deficit relative to p, a phenomenon which has been studied both empirically and theoretically [e.g MacKay, 2003, Wang and Titterington, 2005, Bishop, 2006, Turner and Sahani, 2011, Blei et al., 2017, Giordano et al., 2018]. There exists several measures of uncertainty, and we focus on two: (i) the componentwise variance and (ii) the entropy. The componentwise variance plays a crucial role in Bayesian modeling, especially when estimating the posterior distribution over interpretable variables. Meanwhile the entropy provides a multivariate notion of uncertainty and, in statistical physics, can be linked to the free energy, a quantity of interest for many problems.

Intuitively, we expect factorized VI to shrink the variance

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

of q to minimize its overlap with the tails of p. It is less clear how it should affect the entropy of q: on the one hand, this entropy is decreased by any shrinkage in the variance, but it is increased by the factorized approximation, which delinks the nodes in the full-covariance graphical model of p. Hence entropy is driven by a trade-off between two competing forces. We call this the *shrinkage-delinkage trade-off*. This trade-off hints that the adequacy of factorized VI may depend on the way we elect to measure its uncertainty deficit.

The goal of this paper is to understand the uncertainty deficit of factorized VI in the most informative setting where it can be rigorously analyzed. To this end, we study the special case where p is a Gaussian distribution over $\mathbb{R}^n$ with a full covariance matrix and q is a Gaussian distribution over $\mathbb{R}^n$ with a diagonal covariance matrix. This choice of q is natural when p is a multivariate distribution over $\mathbf{z} \in \mathbb{R}^n$, and leads to factorized Gaussian variational inference (FG-VI)—a popular method among practionners due notably to "black box" implementations such as automatic differentiation variational inference (ADVI) [Kucukelbir et al., 2017]. Our paper expands on previous analyses of FG-VI [Bishop, 2006, Turner and Sahani, 2011] in many ways, but perhaps most significantly by identifying—and elucidating—the shrinkage-delinkage trade-off of factorized VI, which in the considered setting can be written explicitly.

Our analysis is grounded in two fundamental inequalities. First we show that if p is multivariate Gaussian, and if q is the distribution (optimally) estimated by FG-VI, then

$$\mathrm{Var}_q(z_i) \leq \mathrm{Var}_p(z_i). \tag{2}$$

Second, under the same assumptions, we show that

$$\mathcal{H}(q) \leq \mathcal{H}(p), \tag{3}$$

where $\mathcal{H}(\cdot)$ denotes the entropy. This second inequality, relating the entropies of p and q, formalizes an observation [MacKay, 2003, Bishop, 2006] that q tends to be more "compact" than p. Our proofs of these inequalities hold generally for Gaussian distributions over $\mathbb{R}^n$; to the best of our knowledge, they are more direct and more general than previous demonstrations. While both inequalities reveal an uncertainty deficit, we will see that the two notions of uncertainty are not equivalent. Indeed, we provide one example where q underestimates each componentwise variance by a constant multiplicative factor, but the per-component entropy gap between p and q can be arbitrarily small. This discrepancy arises because the entropy gap in FG-VI is in fact *equal* to the KL divergence minimized by FG-VI when it targets a multivariate Gaussian. But, as we will see, this choice of objective function can harm the estimation of marginal variances.

The inequalities in eq. (2–3) anchor our subsequent analysis. As shown in Figure 1, the amount of shrinkage in FG-VI depends in general on the number of components of $\mathbf{z} \in \mathbb{R}^n$ as well as the degree of correlation between these components. With this motivation, we derive an upper bound on the entropy gap in eq. (3) in terms of the problem dimensionality, n, and the condition number of the true correlation matrix.

Finally we examine the relevance for some of our findings when FG-VI is applied to non-Gaussian target distributions. For these experiments, we draw on several examples from the Bayesian literature. We find that, while the variance shrinkage (2) does not hold systematically, it holds on average in the considered examples. We do not have a reliable method to empirically estimate the entropy, but make an argument that eq. (3) may hold in the studied examples.

Our results build on those of many previous studies. MacKay [2003], Bishop [2006], Turner and Sahani [2011], and Blei et al. [2017] all use a two-dimensional Gaussian to illustrate that the approximations from VI are more "compact" than the distributions they target. We formalize this observation in the general n-dimensional setting, while highlighting the difference between componentwise variance and entropy as measures of uncertainty—a difference that becomes more critical in high-dimensional settings. Experiments on non-Gaussian models also suggest a more nuanced picture, showing for instance that FG-VI does not always underestimate every componentwise variance, though in the studied examples variance shrinkage holds *on average*. Previous studies have also examined other measure of uncertainty, such as the frequentist intervals obtained by variational Bayes estimators [e.g. Wang and Titterington, 2005]. Finally, many have been motivated by the uncertainty deficit of factorized VI to develop new methods for inference. These include post-hoc corrections of variational approximations [Giordano et al., 2018] or, for certain models, careful decompositions of p using conditional distributions to justify the assumption of factorization [Agrawal and Domke, 2021].

The code for all results and figures is available on GitHub.

2 PRELIMINARIES

We analyze FG-VI in the setting where $p(\mathbf{z})$ is multivariate Gaussian with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. In this setting FG-VI has a particularly simple solution. (An earlier statement of this solution can be found in Turner and Sahani [2011].)

Proposition 2.1. *Let $q(\mathbf{z})$ be multivariate Gaussian with mean $\boldsymbol{\nu}$ and diagonal covariance $\boldsymbol{\Psi}$. Then the variational parameters minimizing KL$(q||p)$ are given by $\boldsymbol{\nu} = \boldsymbol{\mu}$ and*

$$\Psi_{ii} = \frac{1}{\Sigma^{-1}_{ii}}, \tag{4}$$

where the denominator Σ^{-1}_{ii} denotes a diagonal element of the matrix inverse $\boldsymbol{\Sigma}^{-1}$.

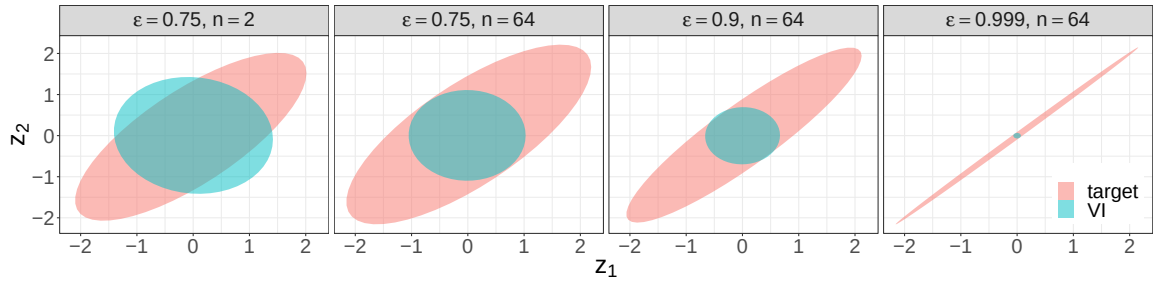

Figure 1: *FG-VI's approximation of a multivariate Gaussian whose correlation matrix has constant off-diagonal terms. FG-VI's variance shrinkage grows with both increasing dimensionality (n) and correlation (ε). For $n = 64$, the distributions are projected onto their first two coordinates. Despite what the picture suggests, the entropy gap between the approximation and the target is actually quite small (Section 3.2). In this sense, the lower-dimensional projection is misleading.*

Proof. The variational parameters $\boldsymbol{\nu}$ and $\boldsymbol{\Psi}$ are estimated by minimizing the KL-divergence

$$\mathrm{KL}(q||p) = \mathbb{E}_q[\log q(\mathbf{z})] - \mathbb{E}_q[\log p(\mathbf{z})], \tag{5}$$

where each expectation is taken with respect to the measure q. Note that only the second term in eq. (5) depends on the variational mean $\boldsymbol{\nu}$, and it is given by

$$-\mathbb{E}_q[\log p(\mathbf{z})] = \tfrac{1}{2}(\boldsymbol{\nu}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\nu}-\boldsymbol{\mu}) + \ldots \tag{6}$$

where the ellipses indicate terms that do not depend on $\boldsymbol{\nu}$. By minimizing this expression, it follows at once that $\boldsymbol{\nu} = \boldsymbol{\mu}$. With this substitution, eq. (5) simplifies to

$$\mathrm{KL}(q||p) = \tfrac{1}{2}\left[\mathrm{trace}\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right) - \log\left|\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}\right| - n\right], \tag{7}$$

and the result in eq. (4) follows by minimizing the above expression with respect to the diagonal elements of $\boldsymbol{\Psi}$. □

In sections 2, 3, and 4 of the paper, we assume that q is the factorized Gaussian distribution whose variances are given by eq. (4). *We emphasize in general that* $\Psi_{ii} \neq \Sigma_{ii}$. However, it is true that $\boldsymbol{\Psi} = \boldsymbol{\Sigma}$ when $\boldsymbol{\Sigma}$ is diagonal.

Many of our results will not be expressed directly in terms of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, but in terms of two related (but dimensionless) matrices. The first is the *correlation matrix* $\mathbf{C}$ with elements

$$C_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}. \tag{8}$$

Note that $C_{ii} = 1$, a simple fact that we will often exploit, and also that $\mathbf{C}$ reduces to the identity matrix when $\boldsymbol{\Sigma}$ is diagonal. At the other extreme, we may consider the case where all the off-diagonal elements of $\mathbf{C}$ are equal to some constant $\varepsilon > 0$. This is explored visually in Figure. (1). In appendix A we show that $\Psi_{ii} \to 0$ as $\varepsilon \to 1$ for fixed n, and that $\Psi_{ii} \to (1-\varepsilon)\Sigma_{ii}$ as $n \to \infty$ for fixed ε. Note that FG-VI underestimates the variance in both limits.

In addition to the correlation matrix, we also define the diagonal *shrinkage* matrix $\mathbf{S}$ with dimensionless entries

$$S_{ii} = \frac{\Sigma_{ii}}{\Psi_{ii}} = \Sigma_{ii}\,\Sigma_{ii}^{-1}. \tag{9}$$

We will use the matrices $\mathbf{C}$ and $\mathbf{S}$ to analyze how FG-VI underestimates the uncertainty of p. The uncertainty in axis-aligned directions is measured by the variances Σ_{ii}, but a multivariate measure of uncertainty is provided by the entropy

$$\mathcal{H}(p) = -\mathbb{E}_p \log p(\mathbf{z}). \tag{10}$$

To what extent does FG-VI underestimate this entropy? As shown next, the answer is very naturally expressed in terms of the correlation matrix $\mathbf{C}$ and the shrinkage matrix $\mathbf{S}$.

Proposition 2.2. *Let p and q be defined as above. Then their difference in entropy is given by*

$$\mathcal{H}(p) - \mathcal{H}(q) = \tfrac{1}{2}\log|\mathbf{S}| - \tfrac{1}{2}\log|\mathbf{C}|^{-1}. \tag{11}$$

Proof. A standard calculation for multivariate Gaussian distributions [Cover and Thomas, 2006] gives $\mathcal{H}(p) = \frac{1}{2}\log|\Sigma|(2\pi e)^n$, and an analogous result holds for $\mathcal{H}(q)$. Let $\Delta\mathcal{H} = \mathcal{H}(p) - \mathcal{H}(q)$. Then we see that

$$\Delta\mathcal{H} = \tfrac{1}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}|\log\boldsymbol{\Psi}| = \tfrac{1}{2}\log\left|\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-\frac{1}{2}}\right|. \tag{12}$$

Now from the definitions in eqs. (8–9), it can be verified by direct substitution that $\boldsymbol{\Psi}^{-\frac{1}{2}}\boldsymbol{\Sigma}\,\boldsymbol{\Psi}^{-\frac{1}{2}} = \mathbf{S}^{\frac{1}{2}}\mathbf{C}\,\mathbf{S}^{\frac{1}{2}}$. It follows from the basic properties of determinants that

$$\Delta\mathcal{H} = \tfrac{1}{2}\log\left|\mathbf{S}^{\frac{1}{2}}\mathbf{C}\mathbf{S}^{\frac{1}{2}}\right| = \tfrac{1}{2}\log|\mathbf{S}| - \tfrac{1}{2}\log|\mathbf{C}|^{-1}.$$

□

3 SHRINKAGE-DELINKAGE TRADE-OFF

In this section we prove that FG-VI systematically underestimates the variance and entropy of a multivariate Gaussian distribution. We will see, however, that a large shrinkage in *all* componentwise variances does not imply a correspondingly large shrinkage in the entropy.

3.1 FUNDAMENTAL INEQUALITIES

Theorem 3.1 (Variance shrinkage)**.** *The solution for FG-VI in eq. (4) underestimates the variance; that is,*

$$\Psi_{ii} \leq \Sigma_{ii}, \tag{13}$$

and the inequality is strict for some component of the variance (i.e., $\Psi_{ii} < \Sigma_{ii}$) if $\mathbf{\Sigma}$ is not purely diagonal.

Proof. Let $\mathbf{C}$ denote the correlation matrix in eq. (8). It can be verified by direct calculation that

$$C_{ij}^{-1} = \Sigma_{ij}^{-1}\sqrt{\Sigma_{ii}\Sigma_{jj}}. \tag{14}$$

As further notation, let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $\mathbf{C}$, and let $\mathbf{e}_i$ denote the unit vector along the ith axis. Then from the solution in eq. (4), it follows that

$$\begin{aligned}
\frac{\Sigma_{ii}}{\Psi_{ii}} &= C_{ii}^{-1}, && (15)\\
&= C_{ii}^{-1} + C_{ii} - 1 && (16)\\
&= \mathbf{e}_i^\top(\mathbf{C}^{-1} + \mathbf{C})\,\mathbf{e}_i - 1, && (17)\\
&\geq \min_{\|\mathbf{e}\|=1}[\mathbf{e}^\top(\mathbf{C}^{-1} + \mathbf{C})\,\mathbf{e} - 1], && (18)\\
&= \min_i\,(\lambda_i^{-1} + \lambda_i - 1), && (19)\\
&\geq \min_{\lambda>0}\,(\lambda^{-1} + \lambda - 1) = 1, && (20)
\end{aligned}$$

where in the last step we have used the fact that the correlation matrix $\mathbf{C}$ has strictly positive eigenvalues. This proves eq. (13). Now suppose that $\mathbf{\Sigma}$ is not purely diagonal. Then $\mathbf{C}$ is also not diagonal; hence there must be some unit vector $\mathbf{e}_i$ that is not an eigenvector of $\mathbf{C}$. In this case the inequality in eq. (18) is strict, showing that $\Sigma_{ii} > \Psi_{ii}$. □

Next we examine the difference in entropy given by eq. (11). First we show that this difference is determined by the trade-off of competing entropic forces.

Theorem 3.2 (The Shrinkage-Delinkage Tradeoff)**.** *Consider the entropy difference in eq. (11) from FG-VI:*

$$\mathcal{H}(p) - \mathcal{H}(q) = \tfrac{1}{2}\log|\mathbf{S}| - \tfrac{1}{2}\log|\mathbf{C}|^{-1}.$$

Both terms in this difference are nonnegative: that is,

$$\log|\mathbf{S}| \;\geq\; 0, \tag{21}$$

$$\log\frac{1}{|\mathbf{C}|} \;\geq\; 0. \tag{22}$$

Before proving the theorem we consider the meaning of these inequalities. Conceptually, the first inequality shows that any shrinkage of variances (from Theorem 3.1) reduces the entropy of q and thus contributes to a larger difference in eq. (11). The second inequality shows that the factorization of q acts as a counterbalance to this effect: the entropy of p is necessarily reduced by the presence of correlations, but such correlations cannot be modeled by q. Thus the factorization of q must (to some extent) oppose the entropy difference in eq. (11), and the net difference is determined by the trade-off of these forces. Visually the factorization of q is represented by the delinkage of nodes in the full-covariance graphical model for p. This is the essence of the *shrinkage-delinkage* tradeoff for FG-VI.

Proof. The bound on $\log|\mathbf{S}|$ in eq. (21) follows at once from Theorem 3.1:

$$\log|\mathbf{S}| = \sum_{i=1}^{n}\log\frac{\Sigma_{ii}}{\Psi_{ii}} \geq 0. \tag{23}$$

As before, let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of $\mathbf{C}$ so that $\log|\mathbf{C}| = \sum_i \log\lambda_i$. From Jensen's inequality, we have:

$$\sum_{i=1}^{n}\log\lambda_i \leq n\log\left[\tfrac{1}{n}\sum_{i=1}^{n}\lambda_i\right] = n\log\tfrac{1}{n}\,\text{trace}(\mathbf{C}) = 0, \tag{24}$$

which proves eq. (22). □

To prove that q underestimates the entropy of p, we need the following result which is important in its own right.

Proposition 3.3. *The entropy gap between p and q is equal to the KL divergence minimized by FG-VI:*

$$\mathcal{H}(p) - \mathcal{H}(q) = \text{KL}(q\|p). \tag{25}$$

Proof. The identity follows by substituting the solution from eq. (4) into the KL divergence in eq. (5). This yields the entropy gap, namely $\text{KL}(q,p) = \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\log|\mathbf{\Psi}|$, computed in eq. (12). □

It follows that FG-VI is minimizing the entropy gap between p and q when it targets a multivariate Gaussian. As suggested by the trade-off in Theorem 3.2, however, the entropy gap can be minimized despite a large shrinkage in componentwise variances.

The nonnegativity of the KL divergence in eq. (25) also leads to the other fundamental inequality of this section.

Theorem 3.4 (Entropy gap)**.** *The solution for FG-VI in eq. (4) underestimates the entropy; that is,*

$$\mathcal{H}(q) \leq \mathcal{H}(p), \tag{26}$$

and this inequality is strict if $\mathbf{\Sigma}$ is not purely diagonal.

An immediate implication of this theorem is that the shrinkage term in eq. (21) dominates the shrinkage-delinkage trade-off in Theorem 3.2.

Remark 3.5. *We see also from Theorem 3.4 that $|\mathbf{\Psi}| \leq |\mathbf{\Sigma}|$. This inequality can be viewed as a multivariate analog of the result, in Theorem 3.1, that $\Psi_{ii} \leq \Sigma_{ii}$.*

3.2 DEMONSTRATION OF THE TRADE-OFF

Figure 2 illustrates the shrinkage-delinkage trade-off in FG-VI, and how it is resolved, for multivariate Gaussian distributions with two types of covariance matrices:

- *Squared exponential kernel:* This type of covariance matrix arises in models involving Gaussian processes [e.g. Rasmussen and Williams, 2006, Chapter 2]. For this example we sampled a random input $\mathbf{x} \sim \text{uniform}(0, 200)^n$ and set the covariance matrix via the kernel function $\Sigma_{ij} = \exp(-(x_i - x_j)^2/\rho^2)$. We use the hyperparameter $\rho > 0$ to vary the degree of correlation.
- *Constant off-diagonal:* The posterior distributions of models with exchangeable data [Gelman et al., 2013, chapter 5] can generate such covariance matrices, or at least covariance matrices whose subblocks have the described structure. For this example we set $\Sigma_{ii} = 1$ along the diagonal and $\Sigma_{ij} = \varepsilon$ for all $i \neq j$, and we used the hyperparameter $\varepsilon > 0$ to vary the degree of correlation.

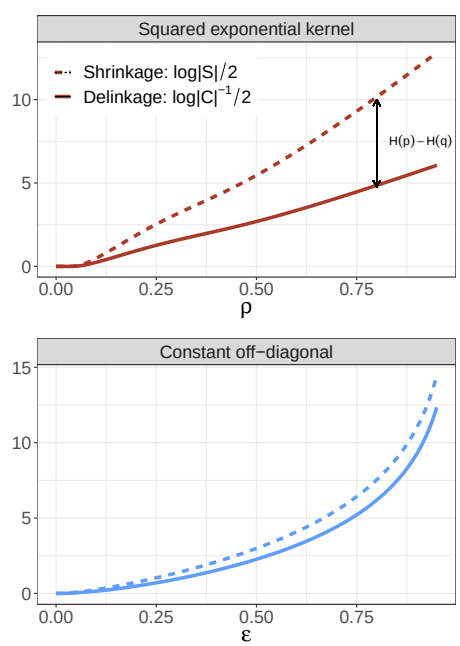

Figure 2: *Shrinkage-delinkage trade-off in FG-VI when the Gaussian target over $\mathbb{R}^n$ has a squared-exponential-kernel covariance matrix (top) or a covariance matrix with constant off-diagonal terms (bottom). Here $n = 10$.*

Figure 2 plots the opposing contributions from the shrinkage and delinkage terms in eq. (11) using solid and dashed lines. In each panel, the difference between these curves reveals the degree to which FG-VI underestimates the entropy of the multivariate Gaussian distribution it is being used to approximate. It can also be seen that FG-VI manages the shrinkage-delinkage trade-off differently for different types of covariance matrices. While in the squared exponential kernel case the entropy gap is large, it is smaller when the covariance matrix has constant off-diagonal terms: there, the shrinkage and delinkage terms in eq. (11) are almost perfectly balanced.

This last finding may come as a surprise in light of earlier results, shown in Figure 1, where the variational approximation is clearly too "compact." But the two-dimensional projections in Figure 1 are misleading. In higher dimensions, the approximating sphere of FG-VI gains more in volume than its target ellipse; this discrepancy arises because each added component is independent for q but strongly correlated for p. The overall effect is that the opposing terms in eq. (11) are nearly balanced. Hence even when FG-VI hardly underestimates the (per-component) entropy, it may still grossly underestimate the componentwise variance. This contrast becomes more acute in the asymptotic limit of n.

Theorem 3.6. *Suppose $\mathbf{\Sigma}$ has constant off-diagonal terms, $\varepsilon > 0$. Then the per-component entropy gap vanishes in the limit $n \to \infty$, whereas every componentwise variance shrinks by a constant factor:*

$$\lim_{n\to\infty} \tfrac{1}{n}\left(\mathcal{H}(p) - \mathcal{H}(q)\right) = 0. \tag{27}$$

$$\lim_{n\to\infty} \left(\Psi_{ii}/\Sigma_{ii}\right) = 1 - \varepsilon. \tag{28}$$

The proof is given in Appendix A. The theorem also shows that the average of the diagonal elements in the shrinkage matrix also converges to a constant factor:

$$\lim_{n\to\infty} \tfrac{1}{n}\text{trace}(\mathbf{S}) = (1-\varepsilon)^{-1}. \tag{29}$$

This example highlights the roots of FG-VI in mean-field approximations from statistical physics [Parisi, 1988]. As is well known, the mean-field approximation for the free energy becomes exact in the limit $n \to \infty$ for certain spin systems with infinite-range interactions. The infinite-range interactions in these systems are analogous, for the Gaussian models we study here, to the assumption of constant off-diagonal terms in the covariance matrix [e.g Mukherjee et al., 2018, Margossian and Mukherjee, 2021]. An important takeaway is that factorized approximations can work well to estimate the entropy—to wit, minimizing the KL-divergence with FG-VI on a Gaussian target is equivalent to minimizing the entropy gap—but still fail to accurately compute the componentwise variances. This can become an important limitation as we apply VI beyond problems in statistical physics and more broadly to Bayesian modeling, where estimation of the variances is critical.

FG-VI's limited ability to estimate the marginal variance is a product of both the choice of the approximating family (factorized Gaussians) and the choice the objective function, $\text{KL}(q||p)$. In Appendix A we show that when minimizing

the *reverse* KL-divergence, $\mathrm{KL}(p||q)$, for a target p, we obtain, in the above example, the opposite result: exact estimations of the marginal variances but an arbitrarily large entropy gap.

4 BOUNDS ON $\log|\mathbf{S}|$ AND $\log|\mathbf{C}|$

In the last section, we saw that the shrinkage-delinkage trade-off played out differently for different types of covariance matrices; we also proved certain asymptotic results that depended on the detailed structure of the covariance matrix (e.g., constant off-diagonal). In this section, we derive more general bounds on the terms in this trade-off that depend only the problem dimensionality, n, and the condition number, R, of the correlation matrix, $\mathbf{C}$.

4.1 OPTIMIZATIONS FOR UPPER BOUNDS

Consider the space of all correlation matrices with condition number R. We denote this space by the set

$$\mathcal{C}_R = \{\mathbf{C} \in \mathcal{S}_+^n \mid C_{ii}=1\ \forall i,\ \lambda_{\max}(\mathbf{C})=R\lambda_{\min}(\mathbf{C})\}. \tag{30}$$

The set contains the intersection of those $n\times n$ matrices that are positive semidefinite (i.e., lying in the cone $\mathcal{S}_+^n$), whose diagonal elements are equal to unity, and whose largest eigenvalue is R times larger than its smallest one.

If the condition number of the correlation matrix $\mathbf{C}$ is known to be R, then we can (in principle) compute the following upper bounds on the terms in eq. (11):

$$\log|\mathbf{S}| \leq \max_{\mathbf{C}\in\mathcal{C}_R}\left[\sum_{i=1}^n \log C_{ii}^{-1}\right], \tag{31}$$

$$\log|\mathbf{C}| \leq \max_{\mathbf{C}\in\mathcal{C}_R}\left[\sum_{i=1}^n \log \lambda_i(\mathbf{C})\right]. \tag{32}$$

In eq. (31), we have used the fact from eq. (15) that $S_{ii}=C_{ii}^{-1}$, while in eq. (32), we have written the determinant of a matrix as the product of its eigenvalues.

In practice, however, it is difficult to perform the optimizations over the set $\mathcal{C}_R$ in eq. (31-32). Instead we consider a more tractable relaxation; the essential idea is to optimize over a larger set of matrices, one that is characterized only in terms of constraints on its eigenvalues. We denote this constrained set of eigenvalues by

$$\Lambda_R = \left\{\boldsymbol{\lambda}\in\mathbb{R}_+^n \,\middle|\, \lambda_1\geq\ldots\geq\lambda_n = R\lambda_1,\ \sum_{i=1}^n \lambda_i = n\right\}. \tag{33}$$

Note that the set $\mathcal{C}_R$ of correlation matrices is contained strictly within the set of matrices with eigenvalues in Λ_R. In particular, a matrix in $\mathcal{C}_R$ is constrained to have ones along its diagonal, while a matrix with eigenvalues in Λ_R is only constrained to have a trace equal to n. With the above relaxation, we obtain the following upper bounds on the terms $\log|\mathbf{S}|$ and $\log|\mathbf{C}|$ in eq. (11).

Proposition 4.1. *Suppose that the correlation matrix* $\mathbf{C}$ *has condition number* R*. Then*

$$\log|\mathbf{S}| \leq n\log\frac{1}{n}\left[\max_{\boldsymbol{\lambda}\in\Lambda_R}\sum_{i=1}^n \lambda_i^{-1}\right] \tag{34}$$

$$\log|\mathbf{C}| \leq \max_{\boldsymbol{\lambda}\in\Lambda_R}\left[\sum_{i=1}^n \log\lambda_i\right]. \tag{35}$$

Proof. The second bound is immediate from eq. (32) and the relaxation in eq. (33). For the first bound, recall that $S_{ii}=C_{ii}^{-1}$, and note from Jensen's equality that

$$\tfrac{1}{n}\textstyle\sum_i \log C_{ii}^{-1} \leq \log\tfrac{1}{n}\sum_i C_{ii}^{-1} = \log\left[\tfrac{1}{n}\sum_i \lambda_i^{-1}(\mathbf{C})\right].$$

The bound in eq. (34) follows from the above in concert with the relaxion in eq. (33). □

4.2 SOLUTIONS FROM SYMMETRY

The optimizations over Λ_R in eqs. (34–35) have a great deal of structure that we can exploit to compute their solutions. We analyze each of these optimizations in turn.

Lemma 4.2. *Let* $\boldsymbol{\lambda}\in\Lambda_R$ *be the solution that maximizes the right side of eq. (34). Then at most one* λ_i *is not equal to either* λ_1 *or* λ_n*.*

Proof. We prove the lemma by contradiction. Suppose there exists a solution with intermediate elements λ_i and λ_j that satisfy $\lambda_1 > \lambda_i > \lambda_j > \lambda_n$. Consider the effect on this solution of a perturbation that adds some small amount $\delta>0$ to λ_i and subtracts the same amount from λ_j. Note that for sufficiently small δ, this perturbation will not leave the set Λ_R; however, it will *expand* the separation of λ_i from λ_j. As a result the objective $\sum_i \lambda_i^{-1}$ has a gain

$$f(\delta) = \frac{1}{\lambda_i+\delta} - \frac{1}{\lambda_i} + \frac{1}{\lambda_j-\delta} - \frac{1}{\lambda_j}. \tag{36}$$

Next we evaluate the derivative $f'(\delta)$ at $\delta=0$; doing so we find $f'(0)=\lambda_j^{-2}-\lambda_i^{-2}>0$. But this yields a contradiction, because any solution must be maximal, and hence stationary (i.e., $f'(0)=0$), with respect to small perturbations. □

The above lemma greatly restricts the form of the solutions that we must consider for the optimization in eq. (34). The next lemma does the same for the optimization in eq. (35).

Lemma 4.3. *Let* $\boldsymbol{\lambda}\in\Lambda_R$ *be the solution that maximizes the right side of eq. (35). Then* $\lambda_i=\lambda_j$ *whenever* $1<i<j<n$*.*

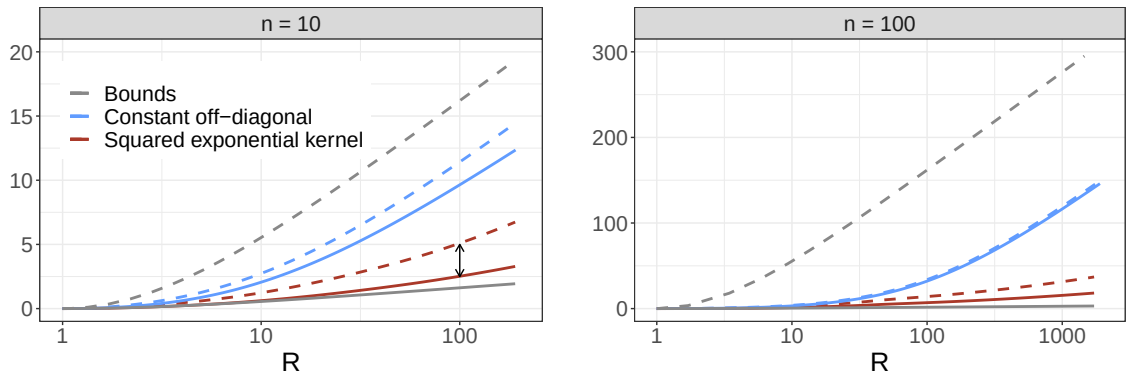

Figure 3: *Bound on the entropy gap. The blue and red curves are replotted from Figure 2. The dotted gray line is an upper bound on the shrinkage term,* $\log|S|/2$*, and the solid gray line is a lower bound on the delinkage term,* $\log|C|^{-1}/2$*. Hence the difference between the gray lines provides an upper bound on the entropy gap,* $\mathcal{H}(p)-\mathcal{H}(q)=log|S|/2-\log|C|^{-1}/2$ *(Theorem 3.2).*

Proof. We prove this lemma in similar fashion. Suppose there exists a solution with intermediate elements λ_i and λ_j that satisfy $\lambda_1 \geq \lambda_i > \lambda_j \geq \lambda_n$. Consider the effect on this solution of a perturbation that adds some small amount $\delta > 0$ to λ_j and subtracts the same amount from λ_i. Again, for sufficiently small δ, this perturbation will not leave the set Λ_R; however, it will *diminish* the separation of λ_i from λ_j. As a result the objective $\sum_i \log \lambda_i$ has a gain

$$g(\delta) = \log(\lambda_i - \delta) + \log(\lambda_j + \delta) - \log(\lambda_i) - \log(\lambda_j). \quad (37)$$

Evaluating the derivative, we find $g'(0) = \lambda_j^{-1} - \lambda_i^{-} > 0$. As before this yields a contradiction, because any solution must be maximal, and hence stationary (i.e., $g'(0)=0$), with respect to small perturbations. □

Now let us consider, at a high level, how these lemmas simplify the optimizations in eqs. (34–35). The lemmas show that for each optimization, there exist *three* elements—the maximum element λ_1, the minimum element λ_n, and some intermediate element λ_k for $1 < k < n$—from which the remaining $n-3$ elements of the solution can be deduced by symmetry. Note also that any solution in Λ_R must satisfy the *two additional* constraints that $\sum_i \lambda_i = n$ and $\lambda_1 = R\lambda_n$. In Appendix B, we show that by exploiting these symmetries and constraints in concert, we can reduce the optimizations in eqs. (34–35) to a sequence of one-dimensional problems for which we have closed-form solutions.

Figure 3 plots the bounds on the shrinkage and delinkage terms as a function of the condition number, R, for problems with dimensionalities $n=10$ (*left*) and $n=100$ (*right*). These bounds provide envelopes between which the actual values of the competing terms in eq. (11) must lie. To illustrate this, the figure also shows the corresponding values of these terms for the squared-exponential-kernel and constant-off-diagonal covariance matrices introduced in the previous section. (Notice that in this figure, unlike Figure 2, these values are plotted against the condition number of the correlation matrix rather than the hyperparameters ρ or ε.)

Using similar methods, it also possible to upper-bound the entropy gap and the trace of the shrinkage matrix (which reflects the average shrinkage in componentwise variance). The derivations of these additional bounds are relegated to Appendices C and D.

5 NON-GAUSSIAN MODELS

Do our results extend in any way when FG-VI is applied to non-Gaussian models? In this section we suppose that p is a *non-Gaussian* target with covariance $\mathbf{\Sigma}$. Our previous analysis of FG-VI targets was based on the variance estimator,

$$(\Psi_G)_{ii} := 1/(\Sigma^{-1})_{ii},$$

and the corresponding shrinkage matrix with diagonal elements $(S_G)_{ii} = \Sigma_{ii}/(\Psi_G)_{ii}$. But neither $\mathbf{\Psi}_G$ nor $\mathbf{S}_G$ will be returned by FG-VI when it is applied to a non-Gaussian target with covariance $\mathbf{\Sigma}$.

To explore these issues, we applied ADVI [Kucukelbir et al., 2017] with a factorized Gaussian approximation to study the posterior distributions in several Bayesian models as well as one "adversarial" target (Table 1). These test targets represent a diversity of applications. The GLM and 8-schools models are taken from the model data base PosteriorDB [Magnusson et al., 2022], while the disease map Gaussian process model and sparse kernel interaction model [Agrawal et al., 2019] are studied by [Margossian et al., 2020]. We also included a mixture of well-separated spherical Gaussians; for this target, the approximation by FG-VI collapses to one of the modes, so that FG-VI can underestimate the componentwise variances by an arbitrarily large amount (e.g., if the modes are widely separated). Note that in all cases, before applying ADVI, we transformed any

constrained (e.g., nonnegative) variables of the target distribution to be unconstrained variables over $\mathbb{R}$.

We estimated the posterior covariance using long runs of Markov chain Monte Carlo, specifically 16,000 draws using the software Stan [Carpenter et al., 2017], except in the mixture example, where the covariance was calculated analytically. We then estimated (i) the shrinkage matrix, $\mathbf{S}$, when targeting the posterior and (ii) the shrinkage matrix, $\mathbf{S}_G$, when targeting a Gaussian with the same covariance as the posterior. For non-Gaussian posteriors, we observe that FG-VI does *not* always underestimate the componentwise variance; see Figure 4. On the other hand, for all models in Table 1, we see that $\frac{1}{n}\text{trace}(\mathbf{S}) > 1$, meaning that the componentwise variances are underestimated *on average* (Figure 5). In addition, for the Bayesian models, we observe that $\text{trace}(\mathbf{S}) \approx \text{trace}(\mathbf{S}_G)$. The mixture target, however, provides a counter-example where $\text{trace}(\mathbf{S}) \gg \text{trace}(\mathbf{S}_G)$.

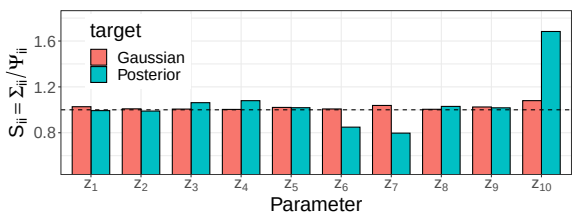

Figure 4: *Shrinkage matrix for FG-VI when targeting the posterior distribution of* `8schools_nc` *versus targeting a Gaussian with the same covariance matrix. For the non-Gaussian target, we may have* $S_{ii} < 1$.

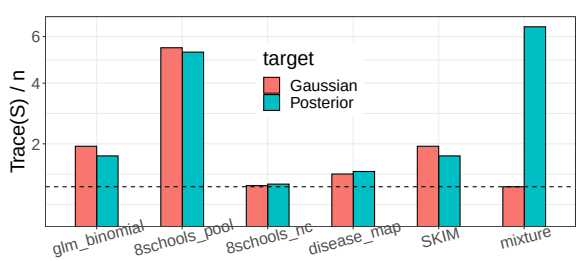

Figure 5: *Trace of shrinkage matrix for various models when targeting the true posterior versus targeting a Gaussian with the same covariance matrix.*

In addition to the shrinkage in componentwise variances, we would have also liked to evaluate the entropy gap in these models. It is easy to obtain an upper bound on this gap by observing that the Gaussian maximizes the entropy among all continuous distributions with a given covariance $\mathbf{\Sigma}$ [Cover and Thomas, 2006]. Thus we have

$$\mathcal{H}(p) - \mathcal{H}(q) \leq \tfrac{1}{2}(\log|\mathbf{\Sigma}| - \log|\mathbf{\Psi}|). \tag{38}$$

We observed this upper bound to be positive for all the models in Table 1. But we also know that FG-VI can overestimate the entropy in non-Gaussian models: Turner and Sahani [2011] demonstrated this for a mixture of largely overlapping 1-dimensional Gaussians. It is an open question to understand the general conditions under which FG-VI underestimates the entropy. We next note that our upper-bound on the entropy gap does not immediately apply to (38), because it is unclear how $\log|\mathbf{\Psi}|$ and $\log|\mathbf{\Psi}_G|$ compare. To evaluate the entropy gap empirically in a non-Gaussian model—as would be required to investigate this question further—it is necessary to estimate the normalizing constant of the posterior. Candidate methods for this, such as bridge sampling [e.g Gronau et al., 2020, Meng and Schilling, 2002], rely on a proposal distribution which (roughly) approximates the target. Typically, a Gaussian-like approximation is used for these proposals, but this is precisely the assumption we want to relax. In other words, we do not wish to compare a theory for Gaussian targets to an empirical benchmark which relies on a Gaussian approximation. We leave this issue to future work.

6 DISCUSSION

In this paper we have shown that FG-VI underestimates the componentwise variance and joint entropy of a multivariate Gaussian distribution. Furthermore we expressed the entropy gap as a trade-off between two competing terms and observed that it was equal to the KL divergence minimized by FG-VI. Our analysis helps to understand why FG-VI can greatly underestimate the componentwise variances even when it effectively minimizes the entropy gap and KL divergence. Our results also suggest that better estimates of variance may be obtained by changing the objective function or using a different family of approximations.

This research has practical implications on when to use FG-VI. When the target distribution exhibits strong correlations, FG-VI can return poor estimates of the marginal variance; this is a limitation in Bayesian modeling where we often care about the posterior variance of interpretable variables. On the other hand, FG-VI can still produce good estimates of the entropy, notably in the limit where n is large. This is one reason that factorized approximations have been widely used for many problems in statistical physics.

An open question is whether a shrinkage-delinkage tradeoff operates when FG-VI is applied to non-Gaussian targets. For such targets, we have produced a counter-example showing that FG-VI can overestimate a particular componentwise variance. On the other hand, we have observed that these variances are underestimated on average, and moreover that the shrinkage term, $\log|\mathbf{S}|$, remains positive. It requires further investigation to make more general statements about the entropy gap. Finally, it would also be interesting to extend our analysis beyond FG-VI—for instance to approximations based on a diagonal plus low-rank covariance matrix, rather than a strictly diagonal one [e.g Zhang et al., 2022].

Call	n	Description	$\frac{1}{n}\log\lvert\mathbf{\Sigma}/\mathbf{\Psi}\rvert$
`glm_binomial`	3	General linear model with a binomial likelihood.	0.291
`8schools_nc`	10	Hierarchical model with a non-centered parameterization.	0.011
`8schools_pool`	9	Same model but with a small, fixed population variance value to enforce strong partial pooling and create a high posterior correlation.	0.339
`disease_map`	102	Gaussian process model with Poisson likelihood. Applied to disease map of Finland using 100 randomly sampled counties (out of 911).	0.066
`SKIM`	305	Sparse kernel interaction model, applied to a Prostate cancer microarray data set on a subset of 200 SNPS.	0.033
`Mixture`	2	Mixture of well-separated Gaussians with spherical covariance matrices.	3.051

Table 1: *Non-Gaussian targets for numerical experiments.*

7 ACKNOWLEDGMENTS

We thank David Blei, Bob Carpenter, Justin Domke and Chirag Modi for feedback on this manuscript.

References

Abhinav Agrawal and Justin Domke. Amortized variational inference in simple hierarchical models. *Neural Information Processing Systems*, 2021.

Raj Agrawal, Jonathan H Huggins, Brians Trippe, and Tamara Broderick. The Kernel interaction trick: Fast Bayesian discovery of pairwise interactions in high dimensions. *International Conference on Machine Learning*, 2019.

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

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

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddel. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1 –32, 2017. doi: 10.18637/jss.v076.i01.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc, 2006.

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.

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

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

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.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013.

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

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

Måns Magnusson, Paul-Christian Bürkner, and Aki Vehtari. posteriordb: a set of posteriors for bayesian inference and probabilistic programming, 2022.

Charles C Margossian and Sumit Mukherjee. Simulating ising and potts models at critical and cold temperatures using auxiliary gaussian variables. *arXiv:2110.10801*, 2021.

Charles C. Margossian, Aki Vehtari, Daniel Simpson, and Raj Agrawal. Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent gaussian models and beyond. *Neural Information Processing Systems*, 2020.

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

Rajarshi Mukherjee, Sumit Mukherjee, and Ming Yuan. Global testing against sparse alternatives under Ising models. *Annals of Statistics*, 46, 2018. doi: doi:10.1214/17-AOS1612.

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

Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

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.

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

Bo Wang and Donald M. Titterington. Inadequacy of interval estimates corresponding to variational bayesian approximations. *Artificial Intelligence and Statistics*, 2005.

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