Interpolating Between Sampling and Variational Inference with Infinite Stochastic Mixtures

Abstract

Sampling and Variational Inference (VI) are two large families of methods for approximate inference that have complementary strengths. Sampling methods excel at approximating arbitrary probability distributions, but can be inefficient. VI methods are efficient, but may misrepresent the true distribution. Here, we develop a general framework where approximations are stochastic mixtures of simple component distributions. Both sampling and VI can be seen as special cases: in sampling, each mixture component is a delta-function and is chosen stochastically, while in standard VI a single component is chosen to minimize divergence. We derive a practical method that interpolates between sampling and VI by solving an optimization problem over a mixing distribution. Intermediate inference methods then arise by varying a single parameter. Our method provably improves on sampling (reducing variance) and on VI (reducing bias+variance despite increasing variance). We demonstrate our method's bias/variance trade-off in practice on reference problems, and we compare outcomes to commonly used sampling and VI methods. This work takes a step towards a highly flexible yet simple family of inference methods that combines the complementary strengths of sampling and VI.

1 INTRODUCTION

We are concerned with the familiar and general case of approximating a probability distribution, such as occurs in Bayesian inference when both the prior over latent variables and the likelihood function connecting them to data are known, but computing the posterior exactly is intractable. There are two largely separate families of techniques for approximating such intractable inference problems: Markov Chain Monte Carlo (MCMC) sampling, and Variational Inference (VI) [Bishop, 2006, Murphy, 2012].

Sampling-based methods, including MCMC, approximate a distribution with a finite set of representative points. MCMC methods are stochastic and sequential, generating a sequence of sample points that, given enough time, become representative of the underlying distribution increasingly well. MCMC sampling is (typically) asymptotically unbiased, at the expense of high variance, leading to long run times in practice. Similar to the approach we take here, sampling methods are studied at different scales: both in terms of their asymptotic limit (i.e. their bias at infinitely many samples) and their practical behavior for finite samples or other resource limits [Korattikara et al., 2014, Angelino et al., 2016].

Variational Inference (VI) refers to methods that produce an approximate distribution by minimizing some quantification of divergence between the approximation and the desired posterior distribution [Blei et al., 2017, Zhang et al., 2019]. For the purposes of this paper, we will use VI to refer to the most common flavor of variational methods, namely minimizing the Kullback-Leibler (KL) divergence between an approximate distribution from a fixed family and the desired distribution [Bishop, 2006, Wainwright and Jordan, 2008, Murphy, 2012, Blei et al., 2017]. The best-fitting approximate distribution is often used directly as a proxy for the true posterior in subsequent calculations, which can greatly simplify those downstream calculations if the approximate distribution is itself easy to integrate. In contrast to MCMC, VI is often used in cases where speed is more important than asymptotic bias [Angelino et al., 2016, Blei et al., 2017, Zhang et al., 2019].

In this work, our goal is to develop an intermediate family of methods that "interpolate" between MCMC and VI, inspired by a simple and intuitive picture (Figure 1): we propose applying sampling methods *in the space of variational parameters* such that the resulting approximation is a stochastic

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

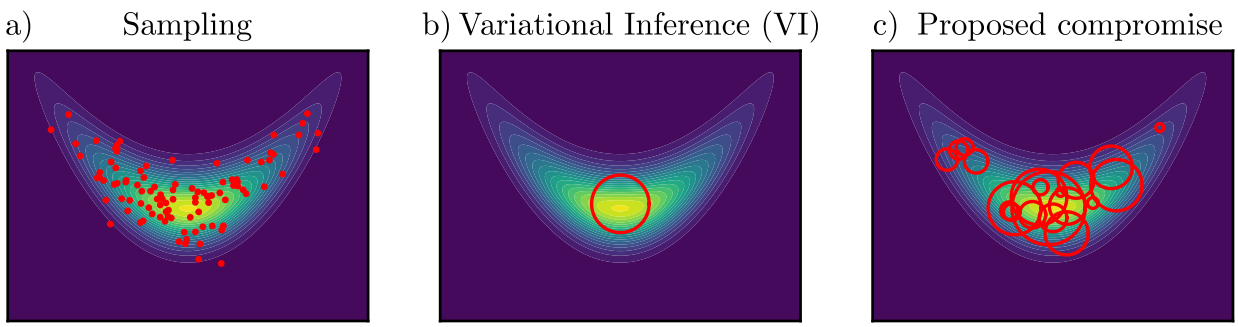

Figure 1: Conceptual introduction on a toy 2D example. **a)** Sampling methods approximate the underlying $\mathrm{p}(\mathbf{x})$ with a stochastic set of representative points. **b)** Variational Inference (VI) methods begin by selecting an approximating distribution family, $\mathrm{q}(\mathbf{x};\theta)$, here an isotropic Gaussian plotted as an ellipse at its 1σ contour. The optimal parameters θ^* are chosen to minimize $\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))$. **c)** We propose using a stochastic mixture of component distributions, where *parameters* θ are sampled rather than the variable of interest $\mathbf{x}$.

mixture of variational "component" distributions [Yin and Zhou, 2018]. This extends sampling by replacing the sampled points with extended components, and it extends VI by replacing the single best-fitting variational distribution with a stochastic mixture of more localized components. This is qualitatively distinct from previous variational methods that use *stochastic optimization*: rather than stochastically optimizing a single variational approximation [Hoffman et al., 2013, Salimans et al., 2015], we use stochasticity to construct a *random mixture* of variational components that achieves lower asymptotic bias than any one component could. As we will show below, this framework generalizes both sampling and VI, where sampling and VI emerge as special cases of a single optimization problem.

This paper is organized as follows. In section 2, we set up the problem and our notation, and describe how both classic sampling and classic VI can be understood as special cases of stochastic mixtures. In section 3, we introduce an intuitive framework for reasoning about infinite stochastic mixtures and define an optimization problem that captures the trade-off between sampling and VI. Section 4 introduces an approximate objective and closed-form solution and describes a simple practical algorithm. Section 5 gives empirical and theoretical results that show how our method interpolates the bias and variance of sampling and VI. Finally, section 6 concludes with a summary, related work, limitations, and future directions.

2 SETUP AND NOTATION

Let $\mathrm{p}^*(\mathbf{x}) = \mathrm{Z}\mathrm{p}(\mathbf{x})$ denote the unnormalized probability distribution of interest, with unknown normalizing constant Z. For instance, in the common case of a probabilistic model with latent variables $\mathbf{x}$, observed data $\mathcal{D}$, and joint distribution $\mathrm{p}(\mathbf{x}, \mathcal{D})$, we are interested in approximations to the posterior distribution $\mathrm{p}(\mathbf{x}|\mathcal{D})$. This is intractable in general, but we assume that we have access to the un-normalized posterior $\mathrm{p}^*(\mathbf{x}|\mathcal{D}) = \frac{1}{\mathrm{Z}}\mathrm{p}(\mathcal{D}|\mathbf{x})\mathrm{p}(\mathbf{x})$.[1] Let $\mathrm{q}(\mathbf{x};\theta)$ be any "simple" distribution that may be used used in a classic VI context (such as mean-field or Gaussian), and let $\mathrm{m}_T(\mathbf{x})$ be a mixture containing T of these simple distributions as components, defined by a set of T parameters $\{\theta^{(1)}, \ldots, \theta^{(T)}\}$:

$$\mathrm{m}_T(\mathbf{x}) \equiv \frac{1}{T}\sum_{t=1}^{T} \mathrm{q}(\mathbf{x};\theta^{(t)})\,. \tag{1}$$

For example, if q is a multivariate normal with mean μ and covariance Σ, then $\theta^{(t)} = \{\mu^{(t)}, \Sigma^{(t)}\}$ and $\mathrm{m}_T(\mathbf{x})$ would be a mixture of T component normal distributions [Gershman et al., 2012].

We will study properties of distributions over component parameters, which we denote $\psi(\theta)$ [Ranganath et al., 2016]. If the set of $\theta^{(t)}$ is drawn randomly from $\psi(\theta)$, then as $T \to \infty$, $\mathrm{m}_T(\mathbf{x})$ approaches the idealized infinite mixture,

$$\mathrm{m}(\mathbf{x}) \equiv \int_\theta \mathrm{q}(\mathbf{x};\theta)\psi(\theta)\mathrm{d}\theta\,. \tag{2}$$

Sampling and VI as special cases of the mixing distribution. Let $\theta^* = \arg\min_\theta \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))$ be the parameters corresponding to the classic single-component variational solution. VI corresponds to the special case where the mixing distribution $\psi(\theta)$ is a Dirac delta around θ^*, or $\psi(\theta) = \delta(\theta - \theta^*)$, in which case the mixture $\mathrm{m}_T(\mathbf{x})$ is equivalent to $\mathrm{q}(\mathbf{x};\theta^*)$ regardless of the number of components T. Sampling can also be seen as a special case of $\psi(\theta)$ in which each component narrows to a Dirac delta ($\psi(\theta)$ places negligible mass on regions of θ-space where components have appreciable width), and the means of the

[1] To reduce clutter, $\mathcal{D}$ will be dropped in the remainder of the paper, and we will use only $\mathrm{p}(\mathbf{x})$ and $\mathrm{p}^*(\mathbf{x})$.

components are distributed according to $\mathrm{p}(\mathbf{x})$. This requires that the component family $\mathrm{q}(\mathbf{x};\theta)$ is capable of expressing a Dirac-delta at any point $\mathbf{x}$, such as a location-scale family. Thus, both sampling and VI can be seen as limiting cases of stochastic mixture distributions, $\mathrm{m}_T(\mathbf{x})$, defined by a distribution over component parameters, $\psi(\theta)$. In what follows, we will show how designing the mixing distribution $\psi(\theta)$ allows us to create mixtures that trade-off the complementary strengths of sampling and VI.

3 CONCEPTUAL FRAMEWORK

3.1 DECOMPOSING $\mathrm{KL}(\mathrm{m}||\mathrm{p})$ INTO MUTUAL INFORMATION AND EXPECTED KL

The idealized infinite mixture $\mathrm{m}(\mathbf{x})$ is fully defined by the chosen component family $\mathrm{q}(\mathbf{x};\theta)$ and the mixing distribution $\psi(\theta)$. Consider the variational objective with respect to the entire mixture, $\mathrm{KL}(\mathrm{m}||\mathrm{p})$:

$$\mathrm{KL}(\mathrm{m}||\mathrm{p}) = \int_{\mathbf{x}} \mathrm{m}(\mathbf{x}) \log \frac{\mathrm{m}(\mathbf{x})}{\mathrm{p}^*(\mathbf{x})} \mathrm{d}\mathbf{x} + \log \mathrm{Z}\,, \tag{3}$$

where Z is the normalizing constant of $\mathrm{p}^*(\mathbf{x})$ and is irrelevant for constructing $\mathrm{m}(\mathbf{x})$. Instead of (3), one can use the equivalent objective of maximizing the *Evidence Lower BOund* or ELBO [Bishop, 2006, Murphy, 2012, Blei et al., 2017]. Regardless, minimizing (3) or maximizing the ELBO for mixtures is intractable in general. However, as first shown by Jaakkola and Jordan [1998] for finite mixtures, it admits the following useful decomposition:

$$\begin{aligned}\mathrm{KL}(\mathrm{m}||\mathrm{p}) = &\underbrace{\int_{\theta} \psi(\theta) \int_{\mathbf{x}} \mathrm{q}(\mathbf{x};\theta) \log \frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{p}^*(\mathbf{x})} \mathrm{d}\mathbf{x}\, \mathrm{d}\theta}_{\text{(i) Expected KL}} \\ &- \underbrace{\int_{\theta} \psi(\theta) \int_{\mathbf{x}} \mathrm{q}(\mathbf{x};\theta) \log \frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{m}(\mathbf{x})} \mathrm{d}\mathbf{x}\, \mathrm{d}\theta}_{\text{(ii) Mutual Information } \mathcal{I}[\mathbf{x};\theta]}\end{aligned} \tag{4}$$

(dropping $\log \mathrm{Z}$). The first term, (i), is the **Expected KL Divergence** for each component when the parameters are drawn from $\psi(\theta)$. This term quantifies, on average, how well the mixture components match the target distribution. In isolation, Expected KL is minimized when all components individually minimize $\mathrm{KL}(\mathrm{q}||\mathrm{p})$, i.e. when $\psi(\theta) \to \delta(\theta - \theta^*)$. This tendency to concentrate $\psi(\theta)$ to the single best variational solution is balanced by the second term, (ii), which is the **Mutual Information** between $\mathbf{x}$ and θ, which we will write $\mathcal{I}[\mathbf{x};\theta]$, under the joint distribution $\mathrm{q}(\mathbf{x};\theta)\psi(\theta)$. This term should be *maximized*, and, importantly, it does not depend on $\mathrm{p}^*(\mathbf{x})$. Mutual Information is maximized when the components are as diverse as possible, which encourages the components to become narrow and to spread out over diverse regions of $\mathbf{x}$ *regardless* of how well they agree with $\mathrm{p}(\mathbf{x})$. This decomposition of $\mathrm{KL}(\mathrm{m}||\mathrm{p})$ into Mutual Information (between $\mathbf{x}$ and θ) and Expected KL (between q and p) is convenient because approximations to Mutual Information are well-studied, and minimizing Expected KL can leverage standard tools from VI.

3.2 TRADING OFF BETWEEN MUTUTAL INFORMATION AND EXPECTED KL

We will refer back to this decomposition of the $\mathrm{KL}(\mathrm{m}||\mathrm{p})$ objective into Expected KL (between q and p) and Mutual Information (between each $\mathbf{x}$ and θ) throughout. Figure 2 depicts a two-dimensional space with Expected KL on the x-axis and Mutual Information on the y-axis. Any given mixing distribution ψ can be placed as a point in this space, but in general many ψ's may map to the same point.

Sampling and VI live at extreme points in this space. Classic VI, where $\psi(\theta) = \delta(\theta - \theta^*)$, corresponds to the blue point (c), because by definition θ^* achieves the minimum possible KL, and $\mathcal{I}[\mathbf{x};\theta]$ is zero. Classic sampling corresponds to the green point (d), with $\psi(\theta)$ placing mass only on Dirac-delta-like components, and selecting each component with probability $\mathrm{p}(\mu)$, where μ is the mean of q determined by θ.

Towards the goal of constructing mixtures that trade-off properties of sampling and VI, we propose to view the two terms in (4) as separate objectives that may be differently weighted, and maximizing the objective

$$\mathcal{L}(\psi, \lambda) = \mathcal{I}[\mathbf{x};\theta] - \lambda \mathbb{E}_{\psi}\left[\mathrm{KL}(\mathrm{q}||\mathrm{p})\right] \tag{5}$$

for a given hyperparameter λ with respect to the mixing distribution ψ. This objective may alternatively be viewed as the Lagrangian of a constrained optimization problem over the mixing density ψ, where Mutual Information is maximized subject to a constraint on Expected KL. This is a concave maximization problem with linear constraints, defining a Pareto front of solutions that each achieve a different balance between Expected KL and Mutual Information. In practice, maximizing Mutual Information necessitates approximations [Poole et al., 2019], so there may be good mixture approximations that are not found in practice, such as the yellow point (e) in Figure 2. In section 4 below, we use an approximation to Mutual Information that has the property, illustrated by the orange curve (f) in Figure 2, of connecting VI (c) to sampling (d), controlled by varying λ. As shown on the right of Figure 2, our method produces mixtures that behave like classic samples when $\lambda = 1$, that behave like classic VI when $\lambda \to \infty$, and that exhibit intermediate behavior at intermediate values of λ.

We emphasize that this frame is quite general: any stochastic mixture can be reasoned about in terms of its Expected KL and Mutual Information, and this is a natural space in which to think about interpolating sampling and VI. A similar decomposition of $\mathrm{KL}(\mathrm{m}||\mathrm{p})$ (or the ELBO) has been used by previous methods that optimize mixtures [Zobay,

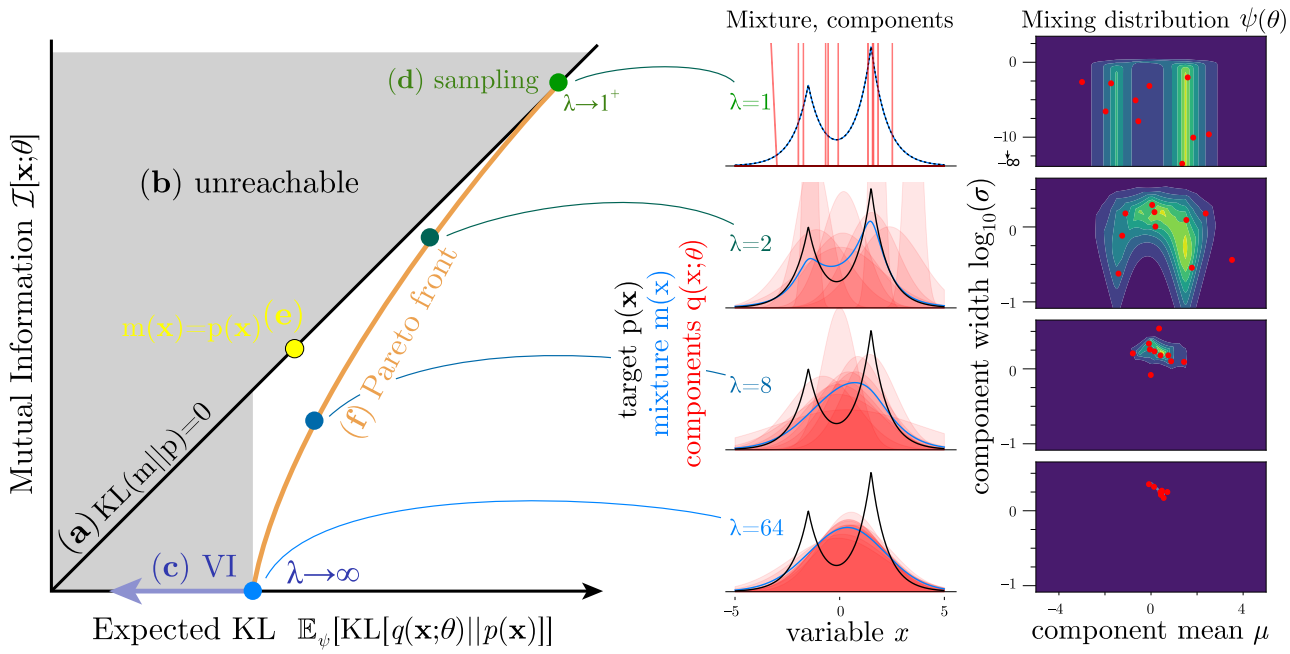

Figure 2: *Left*: Understanding mixtures in terms of Mutual Information and Expected KL. **a)** The quality of any infinite mixture (in terms of $\mathrm{KL}(\mathrm{m}||\mathrm{p})$) is given by its distance from the y=x line (black diagonal line). **b)** Two unreachable regions are shaded in gray: above the y=x line (because $\mathrm{KL}(\mathrm{m}||\mathrm{p}) \geq 0$), and to the left of the single-component variational solution, since VI achieves the minimum $\mathrm{KL}(\mathrm{q}||\mathrm{p})$. **c)** When $\psi(\theta) = \delta(\theta - \theta^*)$ as in classic VI, Expected KL is at its minimum and Mutual Information is zero. Increasing the expressiveness of q corresponds to moving left along the x-axis (blue arrow). **d)** Because sampling is unbiased, it is a mixture that lives on the $\mathrm{KL}(\mathrm{m}||\mathrm{p}) = 0$ or $y = x$ line. If $\mathbf{x}$ is discrete, the coordinates of the point marked (d) are $(\mathcal{H}[\mathbf{x}], \mathcal{H}[\mathbf{x}])$, i.e. the entropy of $\mathrm{p}(\mathbf{x})$. When $\mathbf{x}$ is continuous, both Mutual Information and Expected KL grow unboundedly together as the individual components narrow. **e)** Any point on the y=x line implies $\mathrm{m}(\mathbf{x}) = \mathrm{p}(\mathbf{x})$, and this may be possible without resorting to sampling for certain combinations of p and q. However, such mixtures are not guaranteed to exist for all problems, and are difficult to find due to the intractability of Mutual Information. **f)** We propose a family of mixture approximations, parameterized by λ, that connects VI to sampling in a natural and principled way. Points on this curve correspond to solutions to the (approximate version of the) objective in (5). *Middle*: Examples in a 1D toy problem, where $\mathrm{p}(\mathbf{x})$ is an unequal mixture of two heavy-tailed distributions (black lines), and $\mathrm{q}(\mathbf{x};\theta)$ is a single Gaussian component with parameters $\theta = \{\mu, \log\sigma\}$ (transluscent red components). *Right*: Varying λ controls the mixing distribution over θ (contours). Red points correspond to the Gaussian components in the middle.

2014, Jaakkola and Jordan, 1998, Gershman et al., 2012, Yin and Zhou, 2018]. The primary difference between these previous methods is how they approximate (or lower-bound) Mutual Information. In the next section, we introduce a new approximation that is particularly efficient, and is the first to our knowledge that can produce sampling-like behavior with finitely many components.

4 APPROXIMATE OBJECTIVE

Maximizing Mutual Information, as is required by (5), is a notoriously difficult problem that arises in many domains, and there is a large collection of approximations and bounds in the literature [Jaakkola and Jordan, 1998, Brunel and Nadal, 1998, Gershman et al., 2012, Wei and Stocker, 2016, Kolchinsky and Tracey, 2017, Poole et al., 2019]. Previous work has optimized *finite* mixtures by considering how each of T components interacts with the other $T-1$ components, resulting in quadratic scaling with T [Gershman et al., 2012, Guo et al., 2016, Miller et al., 2017, Kolchinsky and Tracey, 2017, Yin and Zhou, 2018, Poole et al., 2019]. Beginning instead with *infinite* mixtures, we find that the local geometry of θ-space is sufficient to provide an approximation to Mutual Information *that can be evaluated independently for each value of* θ.

4.1 STAM'S INEQUALITY

Mutual Information between $\mathbf{x}$ and θ can be written as

$$\begin{aligned}\mathcal{I}[\mathbf{x};\theta] &= \mathcal{H}[\theta] - \mathbb{E}_{\mathrm{m}(\mathbf{x})}\left[\mathcal{H}[\hat{\theta}|\mathbf{x}]\right] \\ &= \mathcal{H}[\theta] - \mathbb{E}_{\psi(\theta)}\big[\underbrace{\mathbb{E}_{\mathrm{q}(\mathbf{x}|\theta)}[\mathcal{H}[\hat{\theta}|\mathbf{x}]]}_{\mathcal{H}[\hat{\theta}|\theta]}\big]\end{aligned} \tag{6}$$

where $\mathcal{H}[\theta]$ is the entropy of $\psi(\theta)$ and $\mathcal{H}[\hat{\theta}|\mathbf{x}]$ is the entropy of $\mathrm{q}(\hat{\theta}|\mathbf{x}) = \frac{\mathrm{q}(\mathbf{x};\hat{\theta})\psi(\hat{\theta})}{\mathrm{m}(\mathbf{x})}$, i.e. the distribution of *inferred* θ values for a given $\mathbf{x}$. The second line follows simply from expanding the definition of $\mathrm{m}(\mathbf{x})$ in the outer expectation. The term $\mathcal{H}[\hat{\theta}|\theta]$ can be thought of in terms of a statistical estimation problem: $\hat{\theta}$ is the "recovered" value of θ after passing through the "channel" $\mathbf{x}$. Bounding the error of such estimators is a well-studied problem in statistics.

From (6), a lower-bound on Mutual Information can be derived from an *upper bound* on $\mathcal{H}[\hat{\theta}|\theta]$ for each θ. For this, we draw inspiration from Stam's inequality [Stam, 1959, Dembo et al., 1991, Wei and Stocker, 2016], which states

$$\mathcal{H}[\hat{\theta}|\theta] \leq \frac{1}{2}\log\left|2\pi e\mathcal{F}(\theta)^{-1}\right| , \tag{7}$$

where $|\cdot|$ is a determinant, and $\mathcal{F}(\theta)$ is the Fisher Information Matrix, defined as

$$\mathcal{F}(\theta)_{ij} = -\mathbb{E}_{\mathrm{q}(\mathbf{x};\theta)}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log \mathrm{q}(\mathbf{x};\theta)\right] .$$

The Fisher Information Matrix is also the local metric on the *statistical manifold* with coordinates θ [Amari, 2016]; it is used to quantify how "distinguishable" θ is from $\theta + d\theta$. Note that (7) can be viewed as the entropy of a Gaussian approximation to $\mathrm{q}(\hat{\theta}|\mathbf{x})$ with precision matrix $\mathcal{F}(\theta)$; this approximation is most accurate when $\mathrm{q}(\mathbf{x};\theta)$ itself is narrow and approximately Gaussian [Wei and Stocker, 2016].

Combining (6) and (7), we propose to use

$$\mathcal{I}_{\mathcal{F}}[\mathbf{x};\theta] \equiv \mathcal{H}[\theta] - \frac{1}{2}\mathbb{E}_{\psi(\theta)}\left[\log\left|2\pi e\mathcal{F}(\theta)^{-1}\right|\right] \tag{8}$$

as a proxy for the intractable $\mathcal{I}[\mathbf{x};\theta]$ in (5).

Note that $\mathcal{I}_{\mathcal{F}}[\mathbf{x};\theta]$ is not strictly a *bound* on $\mathcal{I}[\mathbf{x};\theta]$, but may be seen as an *approximation* to it [Wei and Stocker, 2016]. Briefly, this is because the original Stam's inequality, as stated in (7), assumes θ is a scalar location parameter, and assumes the high-precision limit where $\mathrm{q}(\hat{\theta}|\mathbf{x})$ is well-approximated by a Gaussian. Despite this, $\mathcal{I}_{\mathcal{F}}[\mathbf{x};\theta]$ is well-suited for our purposes, since (i) it leads to a remarkably simple and easy to implement expression for $\psi(\theta)$ below; (ii) we can prove that it leads to sampling when $\lambda = 1$ and VI when $\lambda \to \infty$; and (iii) the inequality in (7) is nonetheless likely to be strict, since we neglect the prior information contained in $\psi(\theta)$ when estimating $\hat{\theta}$ and therefore over-estimate the entropy.[2]

[2] By analogy to the Bayesian Cramér-Rao bound [Gill and Levit, 1995, Fauß et al., 2021], a tighter variant of (7) could be derived that takes into account the prior, though possibly at the expense of added complexity; we leave this to future work.

4.2 CLOSED-FORM MIXING DISTRIBUTION

Substituting $\mathcal{I}_{\mathcal{F}}[\mathbf{x};\theta]$ for $\mathcal{I}[\mathbf{x};\theta]$ in (5) gives the following approximate objective,

$$\mathcal{L}_{\mathcal{F}}(\psi,\lambda) = \mathcal{H}[\theta] + \mathbb{E}_{\psi}\left[\frac{1}{2}\log|\mathcal{F}| - \lambda\,\mathrm{KL}(\mathrm{q}||\mathrm{p}^*)\right] \tag{9}$$

having dropped additive constants and using $\log|\mathcal{F}^{-1}| = -\log|\mathcal{F}|$. This now resembles a maximum-entropy problem with an expected-value constraint, which has the following simple closed-form solution:

$$\log\psi(\theta) = \frac{1}{2}\log|\mathcal{F}(\theta)| - \lambda\,\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}^*(\mathbf{x})) \tag{10}$$

again dropping additive constants. Equation (10) is strikingly simple, and amenable to many existing MCMC sampling methods for drawing samples of θ from ψ.

Despite being derived from an approximation to our original objective, (10) nonetheless contains both sampling and VI as special cases. As $\lambda \to \infty$, the KL term dominates and $\psi(\theta)$ concentrates to $\delta(\theta - \theta^*)$, reproducing VI. When $\lambda = 1$, this mixing distribution also corresponds to "sampling" in the following sense:

Definition 1 (Sampling) *A stochastic mixture, defined by the component family* $\mathrm{q}(\mathbf{x};\theta)$ *and mixing distribution* $\psi(\theta)$*, is considered to be "sampling" if (i) it is* ***unbiased*** *in the limit of infinitely many components, i.e.* $\mathrm{m}(\mathbf{x}) \to \mathrm{p}(\mathbf{x})$*; and, (ii) it consists of* ***non-overlapping components****. That is, for small values of of* $0 < \epsilon \ll 1$*, wherever* $\mathrm{q}(\mathbf{x};\theta_i) > \epsilon$*, with high probability,* $\mathrm{q}(\mathbf{x};\theta_j) < \epsilon$*, for all pairs* θ_i, θ_j *drawn independently from* $\psi(\theta)$*.*

Lemma 4 in Appendix A.2 establishes that $\psi(\theta)$ with $\lambda = 1$ leads to sampling as defined here, assuming mixture components q are Gaussian. However, we conjecture that sampling arises from a broader class of q components as well.

4.3 IMPLEMENTATION

We implemented (10) in Stan [Carpenter et al., 2017], an open-source framework for probabilistic models and approximate inference algorithms. We sampled θ from $\psi(\theta)$ using Stan's default implementation of the No U-Turn Sampler (NUTS) [Hoffman and Gelman, 2014], but we emphasize that samples can be drawn from (10) using any existing sampling method. All comparisons to existing methods were with Stan's built-in NUTS sampler (over $\mathbf{x}$) and its built-in mean-field VI [Kucukelbir et al., 2017].

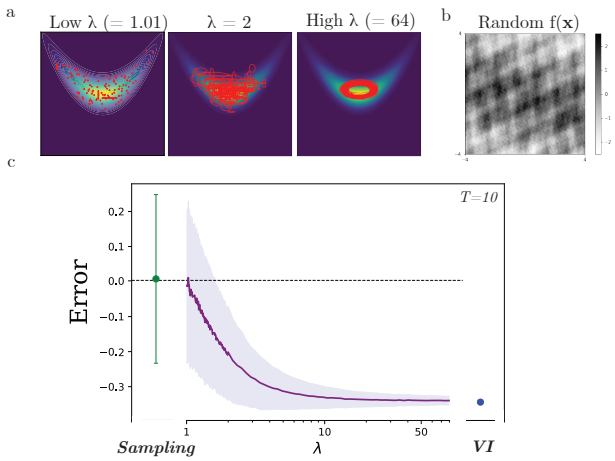

Figure 3: λ controls a bias/variance tradeoff, interpolating between sampling and VI. **a)** For an example 2D distribution (the banana distribution), we set q to Gaussian with diagonal covariance and sampled $\theta \sim \psi(\theta)$ using NUTS (see Appendix B.2 for sampling details). **b)** We selected $f(\mathbf{x})$ as a random mixture of sinusoids at different frequencies. We then calculated the bias and variance of computing $\mathbb{E}_{\mathrm{m}_T}[f(\mathbf{x})]$. **c)** The green point and error bars ("Sampling") indicate the estimated value of $\mathbb{E}_{\mathrm{p}(\mathbf{x})}[f(\mathbf{x})]$ and its variance using NUTS to draw samples of $\mathbf{x}$. The blue point ("VI") shows the value of $\mathbb{E}_{\mathrm{q}(\mathbf{x};\theta^*)}[f(\mathbf{x})]$ using Stan's built-in VI. Our method is shown in the middle across a range of λ values. Low λ provides unbiased but high variance estimators, while high λ provides a bias near that of standard VI and a vanishing variance. In panel (c), we used $T = 10$ independent samples for both classic NUTS and our method.

5 NAVIGATING BIAS/VARIANCE TRADE-OFFS FOR FINITE T

5.1 REDUCING MEAN SQUARED ERROR (MSE)

In this section, we expound the sense in which our method "interpolates" sampling and VI in terms of bias and variance, both analytically and empirically. In our experiments, we quantify bias and variance in terms of the Mean Squared Error (MSE) of the expectation of an arbitrary $f(\mathbf{x})$ using a random mixture of T components, $\mathrm{m}_T(\mathbf{x})$. In Figure 3, we show empirically that by increasing λ one can interpolate between the zero bias but high variance solution, equivalent to sampling, and the zero variance but high bias solution, equivalent to VI. Between these extremes, our method smoothly interpolates both bias and variance.

To show this empirically requires choosing a class of functions $f(\mathbf{x})$. We construct a random smooth function by discrete Fourier synthesis. Specifically, we select a series of sinusoid plane waves in the space of $\mathbf{x}$ with increasing frequency ω, random directions $\mathbf{t}$ and phase ϕ, such that $f(\mathbf{x}) = \sum_{\omega=1}^{N} a_\omega \sin(\omega \mathbf{t}^\top \mathbf{x} + \phi_\omega)$. The amplitudes a_ω are set according to a power law: $a_\omega = \omega^{-\alpha}$. An example of $f(\mathbf{x})$ is shown in Fig. 3b for $\alpha = -1$, and α is varied in Fig 4. Adjusting α allows flexibly setting the "wiggliness" of the synthesized function [Stein and Shakarchi, 2011].

We also tested our algorithm on three reference problems from posteriordb [Magnusson et al., 2021], now evaluating a large set of random fs, defined on the space of each model's unconstrained parameters, with $\alpha = -1$ (Figure B.1). The conclusion is similar: across many random fs, our algorithm performs on average as well as or better than both sampling (by reducing variance) and VI (by reducing bias).

5.2 CONSIDERATIONS FOR SELECTING λ

A first practical consideration for the choice of λ is the particular function $f(\mathbf{x})$ to be integrated. Since MSE can be decomposed into the sum of squared bias and variance, the value of λ that minimizes MSE occurs when $\frac{\partial \text{Bias}^2}{\partial \lambda} = -\frac{\partial \text{Var}}{\partial \lambda}$. Any factor that increases the variance but not the bias of an estimate for a fixed number of components T will push the optimal λ towards higher values.

One such factor is the smoothness of $f(\mathbf{x})$. Classic sampling can have problematically high variance when $f(\mathbf{x})$ is very jagged, as single points are not very representative of the surrounding function. Intuitively, then, we should expect that higher λ (more VI-like mixtures) is preferred when $f(\mathbf{x})$ is more "wiggly." To show this, we generated a random function with varying smoothness, integrated it over random mixtures approximating the 2D banana distribution, and plotted the resulting MSE, bias, and variance (Fig. 4). We adjusted smoothness by varying the power law decay, α, for a fixed set of phases and wave directions. At any value of λ, variance can be seen to increase as f is made more wiggly. With all else held equal, it is better to trade some variance for bias when the integrand changes quickly with $\mathbf{x}$.

Another factor that affects the optimal λ is the computational budget. If time allows a large number of T to be sampled, the optimal λ will approach 1 with a speed that depends on the particular problem (specifically, on $\frac{\partial \text{Bias}^2}{\partial \lambda}$). In our experiments we set a fixed T to demonstrate our algorithm's properties. However, if the number of components is not known in advance, a practitioner may also decrease λ adaptively over time as sampling continues.

5.3 ANALYTICAL RESULTS

While the MSE of the expected value of some $f(\mathbf{x})$ is a useful way to compare approximate inference methods, it depends on the somewhat arbitrary choice of f, and in practice, the f's of interest are often not known at the time of inference. This motivates using the following alternative definition of error that is independent of f and closely related

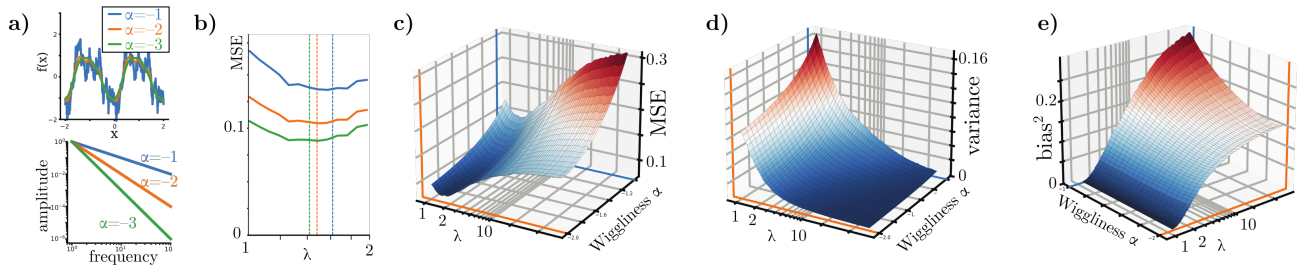

Figure 4: Flexibly trading bias for variance matters for integrating functions. **a)** We generated random functions of a specified smoothness by varying the decay of its power spectrum while randomizing phase. **b)** The λ with the smallest MSE for a fixed number samples $T = 100$ depends on the integrand's smoothness. **c)** Surface plots of MSE for $T = 100$ samples, varying λ and α. **d)** Variance is higher for smaller λ and more wiggly integrands. **e)** Bias vanishes near $\lambda = 1$.

to the variational objective of minimizing KL divergence:

$$\text{KL error} = \mathbb{E}[\mathrm{KL}(\mathrm{m}_T(\mathbf{x})||\mathrm{p}(\mathbf{x}))] = \underbrace{\mathrm{KL}(\mathrm{m}(\mathbf{x})||\mathrm{p}(\mathbf{x}))}_{\text{KL bias}} + \underbrace{\mathbb{E}\left[\mathrm{KL}(\mathrm{m}_T(\mathbf{x})||\mathrm{m}(\mathbf{x}))\right]}_{\text{KL variance}} . \tag{11}$$

That is, **KL bias** is the KL divergence from the infinite mixture $\mathrm{m}(\mathbf{x})$ to the true distribution, and **KL variance** is the average KL, over realizations of T independent mixture components, from $\mathrm{m}_T(\mathbf{x})$ to the infinite mixture $\mathrm{m}(\mathbf{x})$. Note that KL bias is identical to the infinite-mixture objective we started with in (4).

The following theorem establishes that for all finite T, we can always reduce the KL error, relative to sampling, using some $\lambda > 1$.

Theorem 1 (Improve on sampling) *If a mixture is sampling as in Definition 1, then* $\frac{\mathrm{d}}{\mathrm{d}\lambda}$*KL bias* $= 0$ *and* $\frac{\mathrm{d}}{\mathrm{d}\lambda}$*KL variance* < 0*. Thus,* $\frac{\mathrm{d}}{\mathrm{d}\lambda}$*KL error* < 0*.*

This theorem establishes the intuitive result that the variance of sampling can be reduced, minimally impacting its bias, by replacing samples with narrow mixture components. Importantly, Theorem 1 is based on how $\psi(\theta)$ changes with λ when using the closed-form expression for ψ we derived based on the approximate $\mathcal{L}_{\mathcal{F}}$ objective. Because the theorem is phrased in conditional terms ("if the mixture is sampling, then..."), we must further show that both conditions of "sampling" (Definition 1) are met when $\lambda = 1$. This is proved in Lemma 4 in Appendix A.2 for Gaussian components, though we suspect it holds for other component families as well.

We can also improve on VI using stochastic mixtures. However, this result is slightly more subtle, as there are three cases where one should expect VI to be optimal. First, if q is in the same family as p, then $\mathrm{q}(\mathbf{x}; \theta^*) = \mathrm{p}(\mathbf{x})$, then is no benefit to increasing T, and reducing λ only adds variance. Second, if T is small – in the most extreme case, if $T = 1$ – then reducing λ will again only add variance without reducing bias. Third, if p is lighter-tailed than $\mathrm{q}(\mathbf{x}; \theta^*)$, then a mixture of nearby qs will add variance to m [Lindsay, 1983], making the match to p worse. With these three exceptions in mind, the following theorem establishes conditions where we expect to reduce KL error relative to VI by using a large but finite $\lambda < \infty$.

Theorem 2 (Improve on VI) *Assume that* $\mathrm{p}(\mathbf{x})$ *is heavier-tailed than* $\mathrm{q}(\mathbf{x}; \theta^*)$ *and that* λ *is large. Then, there exists some finite* $T_0 > 1$ *such that for all* $T \geq T_0$*,* $\frac{\mathrm{d}}{\mathrm{d}\lambda}$*KL error* > 0*. Proof: see Appendix A.3.*

Note that this result depends on an additional conjecture that relates the curvature in parameter space of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ to the curvature of $\mathrm{KL}(\mathrm{q}||\mathrm{q}^*)$ that we believe holds as long as p is heavier-tailed than q. For details, see Appendix A.3.

6 DISCUSSION

Summary: Our work provides a new perspective on the relationship between the two dominant frameworks for approximate inference – sampling and VI – by viewing both as special cases of inference using a broader class of stochastic mixtures. Our main theoretical contribution is the framework shown in Figure 2, where mixtures that "interpolate" sampling and VI are analyzed in terms of how they trade off Mutual Information and Expected KL. We then derived an easy-to-use method based on an approximation to Mutual Information that uses the local geometry of the space of variational parameters. To demonstrate the ease and effectiveness of our method, we implemented it in the popular Stan language and demonstrated using a small set of reference problems how we "interpolate" sampling and VI by varying a single parameter, λ. Finally, we showed why such an intermediate inference scheme is useful in practice. On one hand, we proved that it is always possible to improve on classic sampling ($\lambda = 1$) by increasing λ: our method provably reduces the variance of sampling while minimally impacting its bias. On the other hand, our method provably reduces the bias of VI under certain conditions (and improves overall error if the number of mixture components is

sufficiently large).

Time and space complexity: By approximating Mutual Information using only *local* geometric information in (8), in our method each component can be selected independently of the others. This means we can select and evaluate T components in $\mathcal{O}(T)$ time and either $\mathcal{O}(T)$ space (if all are stored) or $\mathcal{O}(1)$ space (if components are evaluated online) – identical to traditional MCMC sampling algorithms. Further, we can run independent chains sampling $\theta \sim \psi(\theta)$ for a constant factor speedup. This improves on past work using mixture approximations, which incurred $\mathcal{O}(T^2)$ time and $\mathcal{O}(T)$ space complexity, since the optimization problem for the Tth component depends on the location of the other $T-1$ components, all of which must be in memory at once [Jaakkola and Jordan, 1998, Gershman et al., 2012, Salimans et al., 2015, Guo et al., 2016, Miller et al., 2017, Acerbi, 2018, Yin and Zhou, 2018] (but the $\mathcal{O}(T^2)$ complexity may be hardware-accelerated).

Related Work: The trade-offs between sampling and VI are well-studied, and many methods have been proposed to "close the gap" between them (see [Angelino et al., 2016, Zhang et al., 2019] for general reviews). Like these other methods, we aim to provide good approximations with high computational efficiency and low variance.

There are many methods that use mixture models to reduce the bias of variational inference. Theorem 2 shows that our method only "beats" classic VI when $T > T_0$ for some finite but potentially large T_0. This is the price we pay for drawing mixture components stochastically [Yin and Zhou, 2018]. When a mixture of T components is *optimized* rather than *sampled*, bias is reduced and variance remains near zero, as in previous work [Jaakkola and Jordan, 1998, Gershman et al., 2012, Zobay, 2014, Guo et al., 2016, Miller et al., 2017], but in previous work this optimization has incurred a $\mathcal{O}(T^2)$ cost while our method is $\mathcal{O}(T)$ and can be further parallelized. Further, with some notable exceptions [Anaya-Izquierdo and Marriott, 2007, Salimans et al., 2015], most mixture VI methods make strong assumptions about the family of components [Jaakkola and Jordan, 1998, Gershman et al., 2012, Acerbi, 2018, Miller et al., 2017]. Our framework and method is somewhat agnostic to the family of q, though we have only rigorously proved that is asymptotically unbiased when using Gaussian components.

Many methods use sampling in the service of variational inference, or vice versa, but do not provide a unifying approach to both. These typically use the samples to compute expectations used to update a variational approximation [Acerbi, 2018, Miller et al., 2017, Kucukelbir et al., 2017], rather than to generate the mixture components themselves.

There is also a large number of sampling approaches that aim to improve the efficiency of sampling by reducing its variance at the cost of some bias. Some of these use variational approaches as proposal distributions, but ultimately the posterior is approximated by a set of (possibly weighted) samples of the latent variables [de Freitas et al., 2001, Korattikara et al., 2014, Ma et al., 2015, Zhang et al., 2021]. By expanding each sample to a distribution, our approach allows each sample to cover more space with less variance and greater efficiency [Nalisnick and Smyth, 2017].

Despite some high-level similarities to other approaches, our framework is unusual in approximating the posterior by a sampled mixture of variational approximations. The Mixture Kalman filter [Chen and Liu, 2000] is a special case of this, which uses a sampled mixture of Gaussians, each constructed as a Kalman filter. A related approach is to *optimize* a parameterized function that generates mixture components [Salimans et al., 2015, Wolf et al., 2016, Yin and Zhou, 2018], and generative diffusion models can also be seen as a case of this approach [Sohl-Dickstein et al., 2015, Ho et al., 2020]. Our work differs in that we derived a closed-form mixing distribution that requires no additional learning or optimization and that is readily implemented in existing inference software (Stan, [Carpenter et al., 2017]).

Limitations and future work: Using $\mathcal{I}_{\mathcal{F}}[\mathbf{x};\theta]$ to approximate $\mathcal{I}[\mathbf{x};\theta]$ reduces the generality of our method, since the former is most appropriate for narrow and Gaussian-like components [Wei and Stocker, 2016]. Incorporating prior information from $\psi(\theta)$ into this bound, generalizing to other kinds of components, or even starting with alternative bounds on $\mathcal{I}[\mathbf{x};\theta]$ are all interesting avenues for future work. Another limitation of our theory is that our proof of Theorem 2 depends on a conjecture.

We currently only study mixtures with T *independent* mixture components without taking into account the cost of producing independent samples of θ. In reality, this cost depends on the quality of the sampler, warm-up and burn-in time, and a potentially large number of calls to $\log \mathrm{p}(\mathbf{x})$ [Zhang et al., 2021]. Further, λ dramatically changes the shape of $\log \psi(\theta)$, which may affect the efficiency of the sampler – we mitigated this slightly by scaling the mass parameter of NUTS with λ.

We have so far considered λ to be constant for a run of our algorithm, and this can lead to asymptotic bias even when T is large. A simple adjustment to make our method effective at both small and large T would be to decay λ as T grows, but note that this may require adapting the sampler parameters on the fly. Our method also requires evaluating $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ many times per sample of θ. This could be made more efficient by adapting the number of Monte Carlo evaluations (fewer samples from q are sufficient when λ is low and components are narrow), by accounting for stochastic likelihood evaluations [Ma et al., 2015], or by extending our method to mean-field message-passing [Jaakkola and Jordan, 1998], where $\nabla_\theta \mathrm{KL}(\mathrm{q}||\mathrm{p})$ can be computed in closed form [Hoffman et al., 2013].

References

Luigi Acerbi. Variational Bayesian Monte Carlo. *Advances in Neural Information Processing Systems*, 2018.

S Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016. ISBN 9784431559788. URL https://books.google.com/books?id=UkSFCwAAQBAJ.

Karim Anaya-Izquierdo and Paul Marriott. Local mixture models of exponential families. *Bernoulli*, 13 (3):623–640, 2007. ISSN 1350-7265. doi: 10.3150/07-BEJ6170. URL http://projecteuclid.org/euclid.bj/1186503479.

Elaine Angelino, Matthew James Johnson, and Ryan P. Adams. Patterns of Scalable Bayesian Inference. *Foundations and Trends in Machine Learning*, 9(2-3):119–247, 2016. doi: 10.1561/2200000052.

Christopher M Bishop. Pattern Recognition and Machine Learning. *Pattern Recognition*, page 738, 2006. ISSN 10179909. doi: 10.1117/1.2819119. URL http://www.library.wisc.edu/selectedtocs/bg0137.pdf.

David M. Blei, Alp Kucukelbir, and Jon D Mcauliffe. Variational Inference: A Review for Statisticians. *arXiv*, pages 1–41, 2017.

Nicolas Brunel and Jean Pierre Nadal. Mutual Information, Fisher Information, and Population Coding. *Neural Computation*, 10(7):1731–1757, 1998. ISSN 08997667. doi: 10.1162/089976698300017115.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017. ISSN 15487660. doi: 10.18637/jss.v076.i01.

Rong Chen and Jun S Liu. Mixture kalman filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):493–508, 2000.

Nando de Freitas, Pedro Højen-Sørensen, Michael I. Jordan, and Stuart Russel. Variational MCMC. *Uncertainty in Artificial Intelligence*, 2001.

Amir Dembo, Thomas M. Cover, and Joy A. Thomas. Information Theoretic Inequalities. *IEEE Transactions on Information Theory*, 37(6):1501–1518, 1991. ISSN 15579654. doi: 10.1109/18.104312.

Michael Fauß, Alex Dytso, and H. Vincent Poor. A variational interpretation of the Cramér–Rao bound. *Signal Processing*, 182, 2021. ISSN 01651684. doi: 10.1016/j.sigpro.2020.107917.

Samuel J. Gershman, Matthew D. Hoffman, and David M. Blei. Nonparametric Variational Inference. *Proceedings of the 29th International Conference on Machine Learning*, pages 235–242, 2012. ISSN 0899-7667. doi: 10.1162/089976699300016331. URL https://icml.cc/Conferences/2012/papers/360.pdf.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: A Bayesian Cramér-Rao bound. *Bernoulli*, 1(1):59–79, 1995. URL https://www.jstor.org/stable/3318681.

Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B. Dunson. Boosting Variational Inference. (Vi):1–17, 2016. URL http://arxiv.org/abs/1611.05559.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

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: 1351–1381, 2014.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013. ISSN 1532-4435. doi: citeulike-article-id:10852147. URL http://arxiv.org/abs/1206.7051.

Tommi S. Jaakkola and Michael I. Jordan. Improving the Mean Field Approximation via the Use of Mixture Distributions. In Michael I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.

Artemy Kolchinsky and Brendan D. Tracey. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7): 1–17, 2017. ISSN 10994300. doi: 10.3390/e19070361.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget. *International Conference on Machine Learning*, 32 (1):181–189, 2014. URL http://arxiv.org/abs/1304.5299.

Alp Kucukelbir, David M. Blei, Andrew Gelman, Rajesh Ranganath, and Dustin Tran. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18:1–45, 2017. ISSN 15337928.

Bruce G. Lindsay. The Geometry of Mixture Likelihoods: A General Theory. *The Annals of Statistics*, 11(1):86–94, 1983.

Yi-An Ma, Tianqi Chen, and Emily B. Fox. A Complete Recipe for Stochastic Gradient MCMC. *Advances in Neural Information Processing Systems*, pages 1–16, 2015. ISSN 10495258. URL http://arxiv.org/abs/1506.04696.

M. Magnusson, Paul-Christian Bürkner, and Aki Vehtari. posteriordb: A database of Bayesian posterior inference, 2021. URL https://github.com/stan-dev/posteriordb.

Andrew C. Miller, Nicholas J. Foti, and Ryan P. Adams. Variational Boosting: Iteratively Refining Posterior Approximations. *arXiv*, 2017. URL http://arxiv.org/abs/1611.06585.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, MA, 2012.

Eric Nalisnick and Padhraic Smyth. Variational Inference with Stein Mixtures. *NIPS2017 (Workshop)*, 2017. ISSN 00368075. doi: 10.1126/science.1070850. URL https://www.ics.uci.edu/$\sim$enalisni/AABI_paper30-Stein_Mixtures.pdf.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. *arXiv*, 2019. ISSN 23318422.

Rajesh Ranganath, Dustin Tran, and David M. Blei. Hierarchical Variational Models. *ICML*, 33:1–9, 2016.

Tim Salimans, Diederik P. Kingma, and Max Welling. Markov Chain Monte Carlo and Variational Inference: Bridging the Gap. *Proceedings of the 32nd International Conference on Machine Learning*, pages 1218–1226, 2015. URL http://arxiv.org/abs/1410.6460.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

A. J. Stam. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959. ISSN 00199958. doi: 10.1016/S0019-9958(59)90348-1.

Elias M Stein and Rami Shakarchi. *Fourier analysis: an introduction*, volume 1. Princeton University Press, 2011.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. ISSN 1935-8237. doi: 10.1561/2200000001.

Xue-Xin Wei and Alan A. Stocker. Mutual Information, Fisher Information, and Efficient Coding. *Neural computation*, 28(2), 2016. doi: 10.1162/NECO_a_0084.

Christopher Wolf, Maximilian Karl, and Patrick van der Smagt. Variational inference with hamiltonian monte carlo. *arXiv preprint arXiv:1609.08203*, 2016.

Mingzhang Yin and Mingyuan Zhou. Semi-Implicit Variational Inference. *International Conference on Machine Learning*, 35, 2018.

Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in Variational Inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026, 2019. ISSN 19393539. doi: 10.1109/TPAMI.2018.2889774.

Lu Zhang, Bob Carpenter, Andrew Gelman, and Aki Vehtari. Pathfinder: Parallel quasi-Newton variational inference. *arXiv*, 2021. URL http://arxiv.org/abs/2108.03782.

O. Zobay. Variational Bayesian inference with Gaussian-mixture approximations. *Electronic Journal of Statistics*, 8(1):355–389, 2014. ISSN 19357524. doi: 10.1214/14-EJS887.

Supplemental Information: Interpolating Between Sampling and Variational Inference with Infinite Stochastic Mixtures

A PROOFS AND DERIVATIONS

Throughout, we assume that θ forms a minimal statistical manifold [Amari, 2016], so that the degrees of freedom of q match the dimensionality of θ, and whenever $\mathrm{q}(\mathbf{x};\theta_i) = \mathrm{q}(\mathbf{x};\theta_j)$ for all $\mathbf{x}$, it must be that $\theta_i = \theta_j$.

Recall that in the main text, we defined the following objective:

$$\mathcal{L}(\psi,\lambda) \equiv \mathcal{I}[\mathbf{x};\theta] - \lambda \mathbb{E}_{\psi(\theta)}\left[\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}^*(\mathbf{x}))\right] , \qquad \text{((5) restated)}$$

where $\lambda \in [1,\infty)$ is a hyper-parameter, and ψ is a probability density on θ. We also introduced an **approximate objective** in which $\mathcal{I}[\mathbf{x};\theta]$ is replaced with

$$\mathcal{I}_{\mathcal{F}}[\mathbf{x};\theta] \equiv \mathcal{H}[\theta] - \frac{1}{2}\mathbb{E}_{\psi(\theta)}\left[\log\left|2\pi e \mathcal{F}(\theta)^{-1}\right|\right] . \qquad \text{((8) restated)}$$

This approximate objective is

$$\mathcal{L}_{\mathcal{F}}(\psi,\lambda) = \mathcal{H}[\theta] + \mathbb{E}_{\psi(\theta)}\left[\frac{1}{2}\log|\mathcal{F}(\theta)| - \lambda \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}^*(\mathbf{x}))\right] , \qquad \text{((9) restated)}$$

and it is maximized for a given λ by

$$\psi(\theta) = \frac{1}{Z}\exp\left(\frac{1}{2}\log|\mathcal{F}(\theta)| - \lambda \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))\right) \qquad \text{((10) restated)}$$

$$\text{where} \qquad Z = \int_\theta \exp\left(\frac{1}{2}\log|\mathcal{F}(\theta)| - \lambda \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))\right) \mathrm{d}\theta .$$

A.1 CHARACTERIZING THE PARETO FRONT

Let us begin with a set of results regarding the shape of the Pareto front that connects VI to Sampling in Figure 2.

Lemma 1 *$\mathcal{L}(\psi,\lambda)$ is concave in ψ, i.e. $\mathcal{L}(\omega\psi_1 + (1-\omega)\psi_2,\lambda) \geq \omega\mathcal{L}(\psi_1,\lambda) + (1-\omega)\mathcal{L}(\psi_2,\lambda)$ for $0 \leq \omega \leq 1$. Further, $\mathcal{L}_{\mathcal{F}}(\psi,\lambda)$ is* strictly *concave in ψ.*

Proof: The proof for $\mathcal{L}$ follows from the fact that $\mathbb{E}_{\psi(\theta)}\left[\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}^*(\mathbf{x}))\right]$ is *linear* in ψ, and $\mathcal{I}[\mathbf{x};\theta]$ is known to be concave in the marginal distribution of either variable [Braverman and Bhowmick, 2011]. The proof for $\mathcal{L}_{\mathcal{F}}$ is similar: the $\mathbb{E}_{\psi(\theta)}\left[\frac{1}{2}\log|\mathcal{F}(\theta)|\right]$ term is linear in ψ, and $\mathcal{H}[\theta]$ is strictly concave in ψ. This can be seen, for instance, by taking the

second variational derivative of $\mathcal{H}[\theta]$ with respect to ψ:

$$\begin{aligned}
\nabla^2_\psi \mathcal{H}[\theta]\big|_{\theta_i\theta_j} &= \nabla_\psi \left(\nabla_\psi \mathcal{H}[\theta]\big|_{\theta_i} \right)\big|_{\theta_j} \\
&= \nabla_\psi \left(-\nabla_\psi \int_\theta \psi(\theta) \log \psi(\theta) \mathrm{d}\theta\big|_{\theta_i} \right)\big|_{\theta_j} \\
&= \nabla_\psi \left(-1 - \log \psi(\theta_i) \right)\big|_{\theta_j} \\
&= \begin{cases} -\frac{1}{\psi(\theta_i)} & \text{if } \theta_i = \theta_j \\ 0 & \text{otherwise}\,. \end{cases}
\end{aligned}$$

Since $\psi(\theta) \geq 0$ everywhere, this implies that the curvature of $\mathcal{H}[\theta]$ is strictly negative at all values of θ. ■

Lemma 2 *Let $\mathcal{I}^*(\lambda)$ and $\mathbb{E}[\mathrm{KL}]^*(\lambda)$ denote the values of Mutual Information and Expected KL achieved by optima of $\mathcal{L}$ for a given λ. Then, λ defines the slope of the Pareto front:*

$$\lambda = \frac{\mathrm{d}\mathcal{I}^*/\mathrm{d}\lambda}{\mathrm{d}\mathbb{E}[\mathrm{KL}]^*/\mathrm{d}\lambda}\,.$$

Or, in the case of $\mathcal{L}_\mathcal{F}$, λ similarly defines the slope of

$$\lambda = \frac{\mathrm{d}\mathcal{I}^*_\mathcal{F}/\mathrm{d}\lambda}{\mathrm{d}\mathbb{E}[\mathrm{KL}]^*/\mathrm{d}\lambda}\,,$$

with $\mathcal{I}_\mathcal{F}$ in place of $\mathcal{I}$.

Proof: This follows from viewing $\mathcal{L}$ as the Lagrangian of a constrained optimization problem, with λ as a Lagrange multiplier. The same argument applies to both $\mathcal{L}$ and $\mathcal{I}$ as to $\mathcal{L}_\mathcal{F}$ and $\mathcal{I}_\mathcal{F}$, so we will just give the proof for one. Consider the constrained optimization problem of maximizing $\mathcal{I}$ (or $\mathcal{I}_\mathcal{F}$) subject to the constraint that $\mathbb{E}[\mathrm{KL}(\mathrm{q}||\mathrm{p})] = C$. The Lagrangian for this problem is identical to (5), but with C added:

$$\mathcal{L}(\psi, \lambda) \equiv \mathcal{I}[\mathbf{x}; \theta] - \lambda \left(\mathbb{E}_{\psi(\theta)} \left[\mathrm{KL}(\mathrm{q}(\mathbf{x}; \theta)||\mathrm{p}^*(\mathbf{x})) \right] - C \right)$$

Optimizing with respect to ψ, this is a concave maximization problem with a linear constraint. A well-known property of such problems is that, at the solution, the Lagrange multiplier (λ) is equal to the change in the objective ($\mathcal{I}^*$) per change in the constraint (C), or $\lambda = \frac{\mathrm{d}\mathcal{I}^*}{\mathrm{d}C}$. Since C is the constrained value of $\mathbb{E}[\mathrm{KL}(\mathrm{q}||\mathrm{p})]$, we also immediately have $\frac{\mathrm{d}\mathbb{E}[\mathrm{KL}]^*}{\mathrm{d}C} = 1$. This implies that

$$\lambda = \frac{\mathrm{d}\mathcal{I}^*_\mathcal{F}/\mathrm{d}C}{\mathrm{d}\mathbb{E}[\mathrm{KL}]^*/\mathrm{d}C}\,.$$

So far, we have treated λ as a function of C, but for all values of λ that correspond to a unique C, we can invert this relationship and treat C as a function of λ. Then, assuming $\frac{\mathrm{d}C}{\mathrm{d}\lambda} \neq 0$ for all $1 \leq \lambda < \infty$ that we are interested in, we have

$$\lambda = \frac{\mathrm{d}\mathcal{I}^*_\mathcal{F}/\mathrm{d}C \times \mathrm{d}C/\mathrm{d}\lambda}{\mathrm{d}\mathbb{E}[\mathrm{KL}]^*/\mathrm{d}C \times \mathrm{d}C/\mathrm{d}\lambda} = \frac{\mathrm{d}\mathcal{I}^*_\mathcal{F}/\mathrm{d}\lambda}{\mathrm{d}\mathbb{E}[\mathrm{KL}]^*/\mathrm{d}\lambda}\,.$$

Again using the fact that $C = \mathbb{E}[\mathrm{KL}]^*$ by construction, the condition that $\frac{\mathrm{d}C}{\mathrm{d}\lambda} \neq 0$ is equivalent to $\frac{\mathrm{d}\mathbb{E}[\mathrm{KL}]^*}{\mathrm{d}\lambda} \neq 0$. In other words, as long as changing λ has some effect on $\mathbb{E}[\mathrm{KL}]^*$, the combined effect on $\mathcal{I}^*$ and $\mathbb{E}[\mathrm{KL}]^*$ will be such that $\lambda = \frac{\mathrm{d}\mathcal{I}^*}{\mathrm{d}\mathbb{E}[\mathrm{KL}]^*}$. ■

A.2 SAMPLING-LIKE BEHAVIOR OF OUR METHOD

Recall our definition of sampling:

Definition 1 (Sampling) *A stochastic mixture, defined by the component family $\mathrm{q}(\mathbf{x}; \theta)$ and mixing distribution $\psi(\theta)$, is considered to be "sampling" if (i) it is **unbiased** in the limit of infinitely many components, i.e. $\mathrm{m}(\mathbf{x}) \to \mathrm{p}(\mathbf{x})$; and, (ii) it consists of **non-overlapping components**. That is, for small values of of $0 < \epsilon \ll 1$, wherever $\mathrm{q}(\mathbf{x}; \theta_i) > \epsilon$, with high probability $\mathrm{q}(\mathbf{x}; \theta_j) < \epsilon$, for all pairs θ_i, θ_j drawn independently from $\psi(\theta)$.*

We will assume throughout this section that q is a location-scale family, and in particular Gaussian for Lemma 4, but it may be fruitful for future work to consider other families of mixture components.

Lemma 3 *Sampling is an optimum of the original objective, $\mathcal{L}$, when $\lambda = 1$.*

Proof: When $\lambda = 1$, $\mathcal{L}$ simplifies back to $\mathrm{KL}(\mathrm{m}||\mathrm{p})$. Any **unbiased** mixture is a minimum of $\mathrm{KL}(\mathrm{m}||\mathrm{p})$. ■

Note, however, that this does not imply sampling is the unique optimum. In general there may be other unbiased mixing distributions $\psi(\theta)$ such that $\mathrm{m}(\mathbf{x}) = \mathrm{p}(\mathbf{x})$. For instance, if q is Gaussian and $\mathrm{p}(\mathbf{x})$ is itself a finite mixture of Gaussians, then $\psi(\theta)$ could concentrate on exactly those modes in p. In any case where there two such unbiased ψs, there are in fact infinitely many unbiased, since any mixture of them, $\alpha\psi_1(\theta) + (1-\alpha)\psi_2(\theta)$, will also be unbiased. Among all unbiased mixtures, sampling may in some sense be the worst choice – we conjecture that it has the highest variance of all unbiased mixtures.

Lemma 4 *When q is Gaussian and $\lambda = 1$, the optimal ψ that maximizes the approximate objective $\mathcal{L}_{\mathcal{F}}$ is both **unbiased** and has **non-overlapping components**.*

In other words, Lemma 4 states that the solution to the approximate objective $\mathcal{L}_{\mathcal{F}}$ "looks like" sampling when $\lambda = 1$, in the sense of Definition 1.

Proof: Without loss of generality, let us assume that θ is already parameterized in terms of its location and scale, $[\boldsymbol{\mu}, \boldsymbol{\sigma}]$, where $\boldsymbol{\mu}$ determines the mean of q and $\boldsymbol{\sigma}$ determines its covariance. Then, the Fisher Information Matrix is a block-diagonal matrix:[1]

$$\mathcal{F}(\theta) = \begin{bmatrix} \mathcal{F}(\boldsymbol{\mu}) & 0 \\ 0 & \mathcal{F}(\boldsymbol{\sigma}) \end{bmatrix}$$

where

$$\begin{aligned} \mathcal{F}(\boldsymbol{\mu}) &= \Lambda \\ \mathcal{F}(\boldsymbol{\sigma})_{ij} &= \frac{1}{2}\mathrm{Tr}\left(\Lambda \frac{\partial \Sigma}{\partial \boldsymbol{\sigma}_i} \Lambda \frac{\partial \Sigma}{\partial \boldsymbol{\sigma}_j}\right) . \end{aligned}$$

Λ and Σ are the precision matrix and covariance matrix of q, respectively. Both Λ and Σ are functions of the parameters $\boldsymbol{\sigma}$ but not of $\boldsymbol{\mu}$. To simplify further, let us assume that the covariance of q is diagonal, and that $\boldsymbol{\sigma}_i$ is the log standard deviation of the ith dimension of $\mathbf{x}$:

$$\Sigma(\boldsymbol{\sigma})_{ij} = \begin{cases} e^{2\boldsymbol{\sigma}_i} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

We emphasize that this simplification is for notational convenience only, and other parameterizations of $\Sigma(\boldsymbol{\sigma})$ are permissible. With this assumption, $\mathcal{F}(\boldsymbol{\sigma})$ becomes the identity matrix, and the log determinant of $\mathcal{F}(\theta)$ becomes simply

$$\log|\mathcal{F}(\theta)| = \log|\Lambda| .$$

So, for Gaussian q, the expression for ψ becomes

$$\log\psi(\theta) = \log\psi(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{2}\log|\Lambda(\boldsymbol{\sigma})| - \lambda\mathrm{KL}(\mathrm{q}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\sigma})||\mathrm{p}(\mathbf{x})) .$$

Next, we will split $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ into separate entropy and cross-entropy terms:

$$\begin{aligned} \mathrm{KL}(\mathrm{q}||\mathrm{p}) &= \mathbb{E}_{\mathrm{q}(\mathbf{x};\theta)}\left[\log \mathrm{q}(\mathbf{x};\theta)\right] - \mathbb{E}_{\mathrm{q}(\mathbf{x};\theta)}\left[\log \mathrm{p}(\mathbf{x})\right] \\ &= -\mathcal{H}[\mathrm{q}] + \mathcal{CE}[\mathrm{q}||\mathrm{p}] . \end{aligned}$$

[1] https://en.wikipedia.org/wiki/Fisher_information#Multivariate_normal_distribution

And note that when q is Gaussian, its entropy is given by

$$\mathcal{H}[\mathrm{q}] = \frac{1}{2}\log|2\pi e\Sigma| = \frac{1}{2}\log|\Sigma| + \text{constants}\,.$$

Taking $\lambda = 1$ and using the fact that $\log|\Sigma| = -\log|\Sigma^{-1}| = -\log|\Lambda|$ and combining the above three equations, the $\mathcal{H}[\mathrm{q}]$ and $\log|\mathcal{F}(\boldsymbol{\mu})|$ terms cancel in ψ and we are left – up to additive constants – with

$$\log\psi(\theta) = -\mathcal{CE}[\mathrm{q}||\mathrm{p}] = \mathbb{E}_{\mathrm{q}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\sigma})}\left[\log\mathrm{p}(\mathbf{x})\right]\,. \tag{A.1}$$

To summarize, equation (A.1) says that, using Gaussian components and letting $\lambda \to 1$, our method, derived from the $\mathcal{I}_{\mathcal{F}}$ approximation to $\mathcal{I}$, selects components simply according to the *cross entropy* between $\mathrm{q}(\mathbf{x};\theta)$ and $\mathrm{p}(\mathbf{x})$.

Note that (A.1) is not a proper distribution over θ. To see this, consider any sufficiently narrow component such that q behaves like a Dirac delta, or $\mathbb{E}_{\mathrm{q}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\sigma})}[\log\mathrm{p}(\mathbf{x})] \approx \log\mathrm{p}(\boldsymbol{\mu})$. Wherever this holds for some $\boldsymbol{\sigma}$, it will additionally hold for all *narrower* components at the same $\boldsymbol{\mu}$.[2] Therefore, below a particular scale where q behaves like a Dirac delta, (A.1) places uniform mass on the infinitely many qs that are at least as narrow. This effect is visible in the top-right panel of Figure 2. Also note that ψ is only improper for $\lambda = 1$; for all other $\lambda > 1$, a $(\lambda - 1)\mathcal{H}[\mathrm{q}]$ term remains, and ψ cannot place arbitrarily much mass on arbitrarily narrow components.

Despite its impropriety, we are free to draw samples of θ from this improper ψ when $\lambda = 1$ [Besag et al., 1995, Hobert and Casella, 1996]. We will then find that with probability approaching 1 we only ever see components that "look like" Dirac-deltas. This phenomenon is seen empirically in all of our experiments where we set $\lambda = 1$ and run HMC dynamics drawing $\theta \sim \psi(\theta)$. Since components will become arbitrarily narrow, we have the **non-overlapping components** property required by our definition of sampling.

Consider decomposing $\psi(\theta)$ into $\psi(\boldsymbol{\sigma})\psi(\boldsymbol{\mu}|\boldsymbol{\sigma})$. The previous paragraph establishes that the marginal distribution $\psi(\boldsymbol{\sigma})$ will allocate effectively all samples to parts of θ-space where components behave like Dirac deltas. This implies

$$\begin{aligned}\log\psi(\boldsymbol{\mu}|\boldsymbol{\sigma} = \text{narrow}) &= \mathbb{E}_{\mathrm{q}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\sigma})}\left[\log\mathrm{p}(\mathbf{x})\right] \\ &= \log\mathrm{p}(\boldsymbol{\mu})\,.\end{aligned}$$

Hence, $\mathrm{m}(\mathbf{x})$ will be a mixture of Dirac-delta-like components, each of which is chosen in proportion to the true probability of its mean, $\mathrm{p}(\boldsymbol{\mu})$. This means that $\mathrm{m}(\mathbf{x})$ will be **unbiased**. ■

Theorem 1 (Improve on sampling) *If a mixture is sampling as in Definition 1, then* $\frac{\mathrm{d}}{\mathrm{d}\lambda}$*KL bias* $= 0$ *and* $\frac{\mathrm{d}}{\mathrm{d}\lambda}$*KL variance* < 0*. Thus,* $\frac{\mathrm{d}}{\mathrm{d}\lambda}$*KL error* < 0*.*

Proof: Our approach will be to calculate the variational derivatives of KL bias and KL error with respect to ψ, then take the inner product (directional derivative) with the change in ψ per change in λ.

First, we need the sensitivy of ψ to changes in λ. Recall that the closed-form solution for ψ we get from solving $\mathcal{L}_{\mathcal{F}}$ is

$$\log\psi(\theta) = \frac{1}{2}\log|\mathcal{F}(\theta)| - \lambda\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x})) - \log Z(\lambda)\,.$$

The sensitivity of $\log\psi$ to λ is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}\log\psi(\theta) &= -\mathrm{KL}(\mathrm{q}||\mathrm{p}) + \frac{1}{Z}\int_{\theta'} e^{\frac{1}{2}\log|\mathcal{F}(\theta)| - \lambda\mathrm{KL}(\mathrm{q}||\mathrm{p})}\mathrm{KL}(\mathrm{q}||\mathrm{p})\mathrm{d}\theta' \\ &= \mathbb{E}_{\psi}[\mathrm{KL}(\mathrm{q}||\mathrm{p})] - \mathrm{KL}(\mathrm{q}||\mathrm{p})\,.\end{aligned}$$

Converting from $\log\psi$ to ψ, we get

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\psi(\theta) = \psi(\theta)\left(\mathbb{E}_{\psi}[\mathrm{KL}(\mathrm{q}||\mathrm{p})] - \mathrm{KL}(\mathrm{q}||\mathrm{p})\right) \tag{A.2}$$

[2] There is an implicit assumption here that $\log\mathrm{p}(\mathbf{x})$ is almost everywhere smooth, so that there is some small enough scale at which $\mathrm{p}(\mathbf{x})$ appears locally linear under q.

Recall that we defined KL bias $= \mathrm{KL}(\mathrm{m}||\mathrm{p})$ and KL variance $= \mathbb{E}[\mathrm{KL}(\mathrm{m}_T||\mathrm{m})]$. The variational derivative of Bias with respect to ψ, evaluated at θ is

$$\begin{aligned}\nabla_\psi \mathrm{KL}(\mathrm{m}||\mathrm{p}) =& \nabla_\psi \int_{\mathbf{x}} (\mathbb{E}_\psi[\mathrm{q}(\mathbf{x};\theta)]) \log \frac{(\mathbb{E}_\psi[\mathrm{q}(\mathbf{x};\theta)])}{\mathrm{p}(\mathbf{x})} \mathrm{d}\mathbf{x} \\ =& \int_{\mathbf{x}} \left(\mathrm{m}(\mathbf{x}) \frac{\mathrm{p}(\mathbf{x})}{\mathrm{m}(\mathbf{x})} \frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{p}(\mathbf{x})} + \mathrm{q}(\mathbf{x};\theta) \log \frac{\mathrm{m}(\mathbf{x})}{\mathrm{p}(\mathbf{x})} \right) \mathrm{d}\mathbf{x} \\ =& 1 + \mathbb{E}_{\mathrm{q}(\mathbf{x};\theta)} \left[\log \frac{\mathrm{m}(\mathbf{x})}{\mathrm{p}(\mathbf{x})} \right] . \end{aligned} \tag{A.3}$$

To get the sensitivity of Bias to λ we will take the inner-product of (A.2) with (A.3). This is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda} \text{KL bias} &= \left\langle \frac{\mathrm{d}\text{KL bias}}{\mathrm{d}\psi}, \frac{\mathrm{d}\psi}{\mathrm{d}\lambda} \right\rangle \\ &= \int_\theta \left(1 + \mathbb{E}_{\mathrm{q}(\mathbf{x};\theta)} \left[\log \frac{\mathrm{m}(\mathbf{x})}{\mathrm{p}(\mathbf{x})} \right] \right) \psi(\theta) \left(\mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{p})] - \mathrm{KL}(\mathrm{q}||\mathrm{p}) \right) \mathrm{d}\theta \\ &= \int_\theta (1+0) \psi(\theta) \left(\mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{p})] - \mathrm{KL}(\mathrm{q}||\mathrm{p}) \right) \mathrm{d}\theta \qquad \textbf{(unbiased)} \\ &= \mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{p})] - \mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{p})] \\ &= 0 . \end{aligned}$$

So, we can conclude that in the sampling limit, small changes in λ have no effect on Bias. Geometrically, this tells us the Pareto Front is tangent to the y=x line in that limit, as illustrated in Figure 2.

Next we will consider the variational derivative of the Variance component of KL error with respect to ψ, where

$$\begin{aligned}\text{KL variance} &\equiv \mathbb{E}_{1..T}[\mathrm{KL}(\mathrm{m}_T||\mathrm{m})] \\ &= \mathbb{E}_{1..T} \left[\int_{\mathbf{x}} \left(\frac{1}{T} \sum_{t=1}^{T} \mathrm{q}(\mathbf{x};\theta_t) \right) \log \frac{\left(\frac{1}{T} \sum_{j=1}^{T} \mathrm{q}(\mathbf{x};\theta_j) \right)}{\mathrm{m}(\mathbf{x})} \mathrm{d}\mathbf{x} \right] \end{aligned}$$

using the shorthand $\mathbb{E}_{1..T}[\ldots]$ to denote an expectation over independent draws of $\{\theta_t\} \sim \psi(\theta)$. We will apply the assumption of **non-overlapping components** to simplify $\mathrm{KL}(\mathrm{m}_T||\mathrm{m})$. Let $\int_{\mathbf{x}\in\mathrm{q}_t} \ldots \mathrm{d}\mathbf{x}$ denote an integral over just the region of $\mathbf{x}$-space where $\mathrm{q}(\mathbf{x};\theta_t) > \epsilon$ for some small ϵ. By assumption, these regions are disjoint for all pairs of θs, with high probability. Splitting the integral into T separate regions and rearranging terms inside the $\log$, we have

$$\begin{aligned}\text{KL variance} &\approx \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{1..T} \left[\int_{\mathbf{x}\in\mathrm{q}_t} \mathrm{q}(\mathbf{x};\theta_t) \log \left(\frac{\mathrm{q}(\mathbf{x};\theta_t)}{\mathrm{m}(\mathbf{x})} \left(\frac{1}{T} + \frac{1}{T} \sum_{j\neq t} \frac{\mathrm{q}(\mathbf{x};\theta_j)}{\mathrm{q}(\mathbf{x};\theta_t)} \right) \right) \mathrm{d}\mathbf{x} \right] + \mathcal{O}(\epsilon) \\ &= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{1..T} \left[\int_{\mathbf{x}\in\mathrm{q}_t} \mathrm{q}(\mathbf{x};\theta_t) \left(\log \frac{\mathrm{q}(\mathbf{x};\theta_t)}{\mathrm{m}(\mathbf{x})} + \log \left(\frac{1}{T} + \frac{1}{T} \sum_{j\neq t} \frac{\mathrm{q}(\mathbf{x};\theta_j)}{\mathrm{q}(\mathbf{x};\theta_t)} \right) \right) \mathrm{d}\mathbf{x} \right] + \mathcal{O}(\epsilon) \\ &= \mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{m})] + \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{1..T} \left[\int_{\mathbf{x}\in\mathrm{q}_t} \mathrm{q}(\mathbf{x};\theta_t) \log \left(\frac{1}{T} + \frac{1}{T} \sum_{j\neq t} \frac{\mathrm{q}(\mathbf{x};\theta_j)}{\mathrm{q}(\mathbf{x};\theta_t)} \right) \mathrm{d}\mathbf{x} \right] + \mathcal{O}(\epsilon) \end{aligned}$$

From here, we can get an upper-bound on Variance by noting that $\frac{\mathrm{q}(\mathbf{x};\theta_j)}{\mathrm{q}(\mathbf{x};\theta_t)} \leq 1$ by the non-overlapping assumption. Since there are $T-1$ of these in the sum, $\log \left(\frac{1}{T} + \frac{1}{T} \sum_{j\neq t} \frac{\mathrm{q}(\mathbf{x};\theta_j)}{\mathrm{q}(\mathbf{x};\theta_t)} \right) \leq \left(\frac{1}{T} + \frac{T-1}{T} \right) = 0$, and so

$$\text{KL variance} \leq \mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{m})] + \mathcal{O}(\epsilon) .$$

Note that this bound used the assumption of non-overlapping components, and is therefore only applicable for small λ and moderate values of T. Since we are interested in showing that $\frac{\mathrm{d}}{\mathrm{d}\lambda}$KL variance < 0 in the sampling limit, showing that the

upper bound on variance decreases with λ will suffice. Using this upper-bound, we get the following variational derivative of Variance with respect to ψ at each value of θ:

$$\begin{aligned}
\nabla_\psi \text{KL variance}\big|_\theta &\approx \nabla_\psi \int_\theta \psi(\theta) \int_{\mathbf{x}} \mathrm{q}(\mathbf{x};\theta) \log \frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{m}(\mathbf{x})} \mathrm{d}\mathbf{x}\mathrm{d}\theta \Big|_\theta \\
&= -\int_\theta \psi(\theta) \int_{\mathbf{x}} \mathrm{q}(\mathbf{x};\theta) \frac{\mathrm{m}(\mathbf{x})}{\mathrm{q}(\mathbf{x};\theta)} \frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{m}(\mathbf{x})^2} \mathrm{q}(\mathbf{x};\theta) \mathrm{d}\mathbf{x}\mathrm{d}\theta + \int_{\mathbf{x}} \mathrm{q}(\mathbf{x};\theta) \log \frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{m}(\mathbf{x})} \mathrm{d}\mathbf{x} \\
&= -1 + \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{m}(\mathbf{x}))\,.
\end{aligned}$$

Taking the inner product with $\frac{\mathrm{d}}{\mathrm{d}\lambda}\psi$, and applying the **unbiased** assumption,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\text{KL variance} &= \left\langle \frac{\mathrm{d}\text{KL variance}}{\mathrm{d}\psi}, \frac{\mathrm{d}\psi}{\mathrm{d}\lambda} \right\rangle \\
&= \int_\theta \left(-1 + \mathrm{KL}(\mathrm{q}||\mathrm{m})\right) \psi(\theta) \left(\mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{p})] - \mathrm{KL}(\mathrm{q}||\mathrm{p})\right) \mathrm{d}\theta \\
&= \int_\theta \left(-1 + \mathrm{KL}(\mathrm{q}||\mathrm{p})\right) \psi(\theta) \left(\mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{p})] - \mathrm{KL}(\mathrm{q}||\mathrm{p})\right) \mathrm{d}\theta \qquad \textbf{(unbiased)} \\
&= -\mathbb{E}_{\psi(\theta)}\left[\left(\mathrm{KL}(\mathrm{q}||\mathrm{p}) - \mathbb{E}_\psi[\mathrm{KL}(\mathrm{q}||\mathrm{p})]\right) \mathrm{KL}(\mathrm{q}||\mathrm{p})\right] \\
&= -\mathrm{var}\left(\mathrm{KL}(\mathrm{q}||\mathrm{p})\right)\,.
\end{aligned}$$

In other words, this says that the change in the (upper bound on) "Variance," defined as $\mathbb{E}_{1..T}\mathrm{KL}(\mathrm{m}_T||\mathrm{m})]$, is *negative*, with magnitude given by the variance of the values taken by $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ across all θ.

To summarize, we have shown that, in the sampling limit, where $\lambda = 1$, we have $\frac{\mathrm{d}}{\mathrm{d}\lambda}\text{KL bias} = 0$ and $\frac{\mathrm{d}}{\mathrm{d}\lambda}\text{KL variance} \leq 0$, which proves the lemma. ■

A.3 VI-LIKE BEHAVIOR OF OUR METHOD

Definition 2 (VI limit) *We model the large λ limit of our method using a Laplace approximation around the optimal $\theta^* = \arg\min_\theta \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))$:*

$$\begin{aligned}
&\psi(\theta) \approx \mathcal{N}(\theta; \theta^*, \Sigma^*) \\
\textit{where} \qquad &\Sigma^{*-1} = \lambda \nabla^2_\theta \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))\big|_{\theta^*}\,.
\end{aligned} \tag{A.4}$$

In other words, we approximate ψ by a normal distribution whose mean is θ^* and whose precision is set by the curvature of $\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))$ and scales with λ. We will assume, for the purposes of proofs related to the VI limit, that there is a single optimal θ^*.

Theorem 2 (Improve on VI) *Assume that $\mathrm{p}(\mathbf{x})$ is heaver-tailed than $\mathrm{q}(\mathbf{x};\theta^*)$. Then, there exists some $T_0 > 1$ such that for all $T \geq T_0$, $\frac{\mathrm{d}}{\mathrm{d}\lambda}$KL error > 0, when λ is sufficiently large. Proof: see Appendix A.3.*

Proof: As λ grows, the Laplace approximation in (A.4) becomes increasingly narrow. This allows us to approximate expectations under ψ using a second order Taylor approximation to the integrand. The general rule for multivariate Gaussians is

$$\mathbb{E}_{\mathcal{N}(\mathbf{y};\mu,\Sigma)}[f(\mathbf{y})] \approx f(\mu) + \frac{1}{2}\mathrm{Tr}\left(\Sigma\, \nabla^2_{\mathbf{y}} f\right)\big|_\mu$$

Recall that we defined KL error as $\mathbb{E}_{1..T}[\mathrm{KL}(\mathrm{m}_T(\mathbf{x})||\mathrm{p}(\mathbf{x}))]$. Approximating each $\psi(\theta_t)$ as a multivariate Gaussian, their product is also a multivariate Gaussian whose collective covariance is block-diagonal[3] containing T copies of Σ^* from (A.4), and whose collective mean is θ^* for each component. At this mean value where all T components' parameters are

[3]This assumes the T components are statistically independent draws from $\psi(\theta)$. The approach outlined here could be generalized to include correlation between θs in the off-block-diagonals to model variance of an autocorrelated chain of θ values.

equal to θ^*, $\mathrm{m}_T(\mathbf{x})$ becomes $\mathrm{q}(\mathbf{x};\theta^*)$. Hence, applying the Taylor series approximation to KL error, the $f(\mu)$ term is just $\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta^*)||\mathrm{p}(\mathbf{x}))$. The second term is

$$\frac{1}{2}\mathrm{Tr}\left(\begin{bmatrix}\Sigma^* & & & 0\\ & \Sigma^* & & \\ & & \ddots & \\ 0 & & & \Sigma^*\end{bmatrix}\nabla^2_{\theta_1,\ldots,\theta_T}\mathrm{KL}(\mathrm{m}_T||\mathrm{p})\right).$$

First, note that the zeros in the off-block-diagonal terms on the left mean that we can ignore interactions between θs across different mixture components in the Hessian term on the right. Second, there is $T-$fold symmetry between all components. So, this simplifies to

$$\frac{T}{2}\mathrm{Tr}\left(\Sigma^*\ \nabla^2_{\theta_1}\mathrm{KL}(\mathrm{m}_T||\mathrm{p})\right) = \frac{T}{2\lambda}\mathrm{Tr}\left((\nabla^2_\theta\mathrm{KL}(\mathrm{q}||\mathrm{p}))^{-1}\ \nabla^2_{\theta_1}\mathrm{KL}(\mathrm{m}_T||\mathrm{p})\right).$$

Next, since this Hessian is being evaluated around the point θ^*, all of $\theta_2,\ldots,\theta_T$ are equal to θ^*, and we can write the mixture as a function only of the component parameters we are varying in the Hessian. Call this mixture with $T-1$ components set to the variational solution m^*_T, defined as

$$\mathrm{m}^*_T(\mathbf{x};\theta) = \frac{T-1}{T}\mathrm{q}(\mathbf{x};\theta^*) + \frac{1}{T}\mathrm{q}(\mathbf{x};\theta).$$

We will now calculate each of these Hessians. Note: in what follows we will use θ_i and θ_j to indicate the i and jth indices of the vector θ, whereas we had used θ_t to indicate one of T vectors. For Σ^*, we need the second derivative (Hessian) of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_j\partial\theta_i}\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x})) &= \frac{\partial^2}{\partial\theta_j\partial\theta_i}\int_{\mathbf{x}}\mathrm{q}(\mathbf{x};\theta)\log\frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{p}(\mathbf{x})}\mathrm{d}\mathbf{x}\\
&= \frac{\partial}{\partial\theta_j}\int_{\mathbf{x}}\left[\left(\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\right)\left(1+\log\frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{p}(\mathbf{x})}\right)\right]\mathrm{d}\mathbf{x}\\
&= \int_{\mathbf{x}}\left[\left(\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\right)\left(\frac{\frac{\partial}{\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)}{\mathrm{q}(\mathbf{x};\theta^*)}\right)+\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)\right)\left(1+\log\frac{\mathrm{q}(\mathbf{x};\theta^*)}{\mathrm{p}(\mathbf{x})}\right)\right]\mathrm{d}\mathbf{x}\\
(*) &= \int_{\mathbf{x}}\frac{\left(\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\right)\left(\frac{\partial}{\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)\right)}{\mathrm{q}(\mathbf{x};\theta^*)}\mathrm{d}\mathbf{x}+\int_{\mathbf{x}}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)\right)\log\frac{\mathrm{q}(\mathbf{x};\theta^*)}{\mathrm{p}(\mathbf{x})}\mathrm{d}\mathbf{x}\\
&= \mathcal{F}(\theta^*)+M(\theta^*).
\end{aligned}\tag{A.5}$$

In line $(*)$ we used the fact that $\int_{\mathbf{x}}\nabla^2_\theta\mathrm{q}(\mathbf{x};\theta)\mathrm{d}\mathbf{x} = \nabla^2_\theta\int_{\mathbf{x}}\mathrm{q}(\mathbf{x};\theta)\mathrm{d}\mathbf{x} = \nabla^2_\theta 1 = 0$. $\mathcal{F}$ is the Fisher Information Matrix, and we have defined $M(\theta) = \int_{\mathbf{x}}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)\right)\log\frac{\mathrm{q}(\mathbf{x};\theta)}{\mathrm{p}(\mathbf{x})}\mathrm{d}\mathbf{x}$.

Following a similar derivation, the Hessian of $\mathrm{KL}(\mathrm{m}_T^*(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))$ is

$$
\begin{aligned}
\frac{\partial^2}{\partial\theta_j\partial\theta_i}\mathrm{KL}(\mathrm{m}_T^*(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x})) &= \frac{\partial^2}{\partial\theta_j\partial\theta_i}\int_{\mathbf{x}}\left(\frac{T-1}{T}\mathrm{q}(\mathbf{x};\theta^*)+\frac{1}{T}\mathrm{q}(\mathbf{x};\theta)\right)\log\frac{\left(\frac{T-1}{T}\mathrm{q}(\mathbf{x};\theta^*)+\frac{1}{T}\mathrm{q}(\mathbf{x};\theta)\right)}{\mathrm{p}(\mathbf{x})}\mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial\theta_j}\int_{\mathbf{x}}\left[\frac{1}{T}\left(\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\right)+\frac{1}{T}\left(\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\right)\log\frac{\left(\frac{T-1}{T}\mathrm{q}(\mathbf{x};\theta^*)+\frac{1}{T}\mathrm{q}(\mathbf{x};\theta)\right)}{\mathrm{p}(\mathbf{x})}\right]\mathrm{d}\mathbf{x} \\
&= \frac{1}{T}\frac{\partial}{\partial\theta_j}\int_{\mathbf{x}}\left[\left(\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\right)\left(1+\log\frac{\left(\frac{T-1}{T}\mathrm{q}(\mathbf{x};\theta^*)+\frac{1}{T}\mathrm{q}(\mathbf{x};\theta)\right)}{\mathrm{p}(\mathbf{x})}\right)\right]\mathrm{d}\mathbf{x} \\
&= \frac{1}{T}\int_{\mathbf{x}}\left[\left(\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\right)\left(\frac{\frac{1}{T}\frac{\partial}{\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)}{\mathrm{m}_T^*(\mathbf{x};\theta)}\right)+\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)\right)\left(1+\log\frac{\mathrm{m}_T^*(\mathbf{x};\theta)}{\mathrm{p}(\mathbf{x})}\right)\right]\mathrm{d}\mathbf{x} \\
(**) &= \frac{1}{T^2}\int_{\mathbf{x}}\frac{\left(\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\right)\left(\frac{\partial}{\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)\right)}{\mathrm{q}(\mathbf{x};\theta^*)}\mathrm{d}\mathbf{x}+\frac{1}{T}\int_{\mathbf{x}}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)\right)\log\frac{\mathrm{q}(\mathbf{x};\theta^*)}{\mathrm{p}(\mathbf{x})}\mathrm{d}\mathbf{x} \\
&= \frac{1}{T^2}\mathcal{F}(\theta^*)+\frac{1}{T}M(\theta^*) \\
&= \frac{1}{T}\frac{\partial^2}{\partial\theta_j\partial\theta_i}\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))+\mathcal{F}(\theta)\left(\frac{1-T}{T^2}\right)
\end{aligned}
\tag{A.6}
$$

Here, in $(**)$, we additionally used the fact that $\mathrm{m}_T^*(\mathbf{x};\theta^*)=\mathrm{q}(\mathbf{x};\theta^*)$. We then wrote the final line in terms of the Hessian of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ in (A.5).

To summarize, near the variational limit we have that the KL error is approximately

$$
\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta^*)||\mathrm{p}(\mathbf{x}))+\frac{T}{2\lambda}\mathrm{Tr}\Big(\underbrace{\left(\nabla_\theta^2\mathrm{KL}(\mathrm{q}||\mathrm{p})\right)}_{(A.5)}{}^{-1}\underbrace{\left(\nabla_\theta^2\mathrm{KL}(\mathrm{m}_T^*||\mathrm{p})\right)}_{(A.6)}\Big).
$$

Plugging in (A.5) and (A.6), this is

$$
\begin{aligned}
\text{KL error} &\approx \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta^*)||\mathrm{p}(\mathbf{x}))+\frac{1}{2\lambda}\mathrm{Tr}\left(\mathbf{I}+\frac{1-T}{T}(\mathcal{F}+M)^{-1}\mathcal{F}\right) \\
&= \mathrm{KL}(\mathrm{q}(\mathbf{x};\theta^*)||\mathrm{p}(\mathbf{x}))+\frac{d}{2\lambda}-\frac{1}{2\lambda}\mathrm{Tr}\left(\frac{T-1}{T}(\mathcal{F}+M)^{-1}\mathcal{F}\right)
\end{aligned}
$$

where $\mathbf{I}$ is the identity matrix. Consider the case where $T=1$: the KL error simplifies to $\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta^*)||\mathrm{p}(\mathbf{x}))+\frac{d}{2\lambda}$ where d is the dimensionality of θ. Therefore when $T=1$, KL error is only reduced by further increasing λ. This is an intuitive result: we cannot reduce bias compared to VI when using a single component, and any stochasticity only adds variance.

Now consider the case where $T\geq 2$. We are interested in cases where KL error *increases* with λ near the VI limit. This is equivalent to asking when the following inequality holds:

$$
\mathrm{Tr}\left((\mathcal{F}+M)^{-1}\mathcal{F}\right)>\frac{T}{T-1}\overbrace{\mathrm{Tr}(\mathbf{I})}^{d},
$$

Recall from (A.5) that $\mathcal{F}+M$ is the Hessian of $\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))$, and note that the Fisher Information Matrix is equivalent to the Hessian with respect to θ of $\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{q}(\mathbf{x};\theta^*))$, so we can rewrite this inequality as

$$
\mathrm{Tr}\left((\nabla_\theta^2\mathrm{KL}(\mathrm{q}||\mathrm{p}))^{-1}\nabla_\theta^2\mathrm{KL}(\mathrm{q}||\mathrm{q}^*)\right)>\frac{T}{T-1}\mathrm{Tr}(\mathbf{I}).
$$

Assuming θ^* is a local minimum of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ (which follows from the assumption that θ^* is the unique minimum), both of these are positive definite matrices encoding how sharply curved the $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ or $\mathrm{KL}(\mathrm{q}||\mathrm{q}^*)$ objectives are.

If p is in the same family as q, then $\mathrm{q}*=\mathrm{p}$ and this inequality becomes $\mathrm{Tr}\left(\mathbf{I}\right)>\frac{T}{T-1}\mathrm{Tr}(\mathbf{I})$, which is false for all finite T and approaches equality as $T\to\infty$. This again captures the intuitive idea that we cannot improve on VI by reducing λ when the single-component q is already unbiased.

Conversely, we can view the ratio

$$\frac{\mathrm{Tr}\left((\nabla_\theta^2 \mathrm{KL}(\mathrm{q}||\mathrm{p}))^{-1} \nabla_\theta^2 \mathrm{KL}(\mathrm{q}||\mathrm{q}^*)\right)}{\mathrm{Tr}(\mathbf{I})} \tag{A.7}$$

as an indication of *how poorly matched* $\mathrm{q}(\mathbf{x};\theta^*)$ is to $\mathrm{p}(\mathbf{x})$, locally around the single-component VI solution. We **conjecture** that this ratio is always greater than 1 whenever $\mathrm{p}(\mathbf{x})$ is heavier-tailed than $\mathrm{q}(\mathbf{x};\theta^*)$. Since $\frac{T}{T-1}$ approaches 1 from above in the limit of large T, this implies that there will be some finite T_0 where $\frac{T_0}{T_0-1}$ is less than the ratio in (A.7), and that such a T_0 will be reached sooner the worse $\mathrm{q}(\mathbf{x};\theta^*)$ locally approximates p. ■

B ADDITIONAL EXPERIMENTS

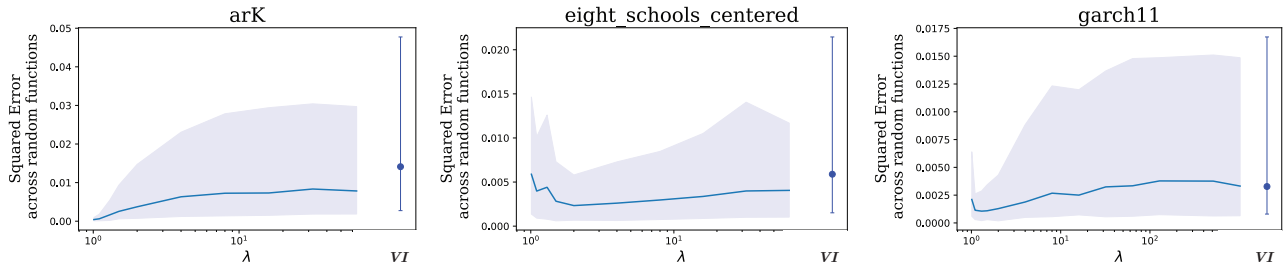

Figure B.1: Further examples. To verify that the results in the main manuscript generalize other datasets, we ran our algorithm at multiple values of λ on three problems from the posteriordb dataset [Magnusson et al., 2021]. Rather than show convergence for one random $f(\mathbf{x})$, as in the main text, we generated 200 random functions over the unconstrained parameter space of each model, using the Fourier synthesis method described in section C below, with $\alpha = -1$. We then calculated the expectation $\mathbb{E}_{\mathrm{m}_T(\mathbf{x})}[f(\mathbf{x})]$ after subsampling mixtures of $T = 100$ components, and plotted the distribution of squared error relative to ground truth expectations (based on very long runs of Stan's default NUTS implementation). Error bars reflect the combined effect of two sources of variability: one from the random choice of $f(\mathbf{x})$ and one from the random subsampling of mixtures. Blue line is the median squared error across different fs and different mixtures, and shading shows $[25\%, 75\%]$ quantiles. The blue dot to the right represents the expectation calculated by a mean-field VI approximation (a diagonal Gaussian) using the automatic differentiation variational inference (ADVI) package built in to Stan [Kucukelbir et al., 2017], run with its default parameters.

C NUMERICAL DETAILS

We implemented (10) in Stan [Carpenter et al., 2017]. For q, we used the family of multivariate Gaussians with diagonal covariance, parameterized as $\theta = [\mu_1, \ldots, \mu_n, \log \sigma_1, \ldots, \log \sigma_n]$ where n is the number of unconstrained parameters (i.e the dimensionality of $\mathbf{x}$). In this parameterization, $\frac{1}{2} \log \mathcal{F}(\theta)$ is simply $-\sum_{i=1}^{n} \log \sigma_i$. We sampled θ from $\psi(\theta)$ using Stan's default implementation of the No U-Turn Sampler (NUTS) with automatic step-size adaptation [Hoffman and Gelman, 2014], and we set the mass equal to λ times the identity matrix. NUTS requires both $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ and its gradient, which we computed using Monte Carlo samples from q and the reparameterization trick. The reparameterized samples were frozen for each trajectory of NUTS and resampled between trajectories.

All code to generate the figures in this paper is available publicly online; the repository URL will be shared after the double-blind review process is complete. Python libraries used include NumPy, SciPy, PyTorch, and Matplotlib [Harris et al., 2020, Virtanen et al., 2020, Paszke et al., 2019, Hunter, 2007].

C.1 FIGURE DETAILS

We used two toy distributions in our results:

- The "banana" distribution over $\mathbb{R}^2$, defined as

$$\log \mathrm{p}(x, y) = -(y - (x/2)^2)^2 - (x/2)^2 \,.$$

- The "Laplace mixture" distribution over $\mathbb{R}^1$, defined as

$$\mathrm{p}(x) \propto 0.4 e^{\frac{|x+1.5|}{0.75}} + 0.6 e^{\frac{|x-1.5|}{0.75}} \,.$$

We also tested our method on three reference problems taken from posteriordb [Magnusson et al., 2021], a database of reference problems for testing and validating inference methods. These were `arK`, `eigh schools centered`, and `garch11`. Results for these additional problems are shown in Figure B.1.

In our experiments, all functions integrated are sums of sinusoids,

$$f(\mathbf{x}) = \sum_{\omega=1}^{N} a \sin(\omega \mathbf{t}^T \mathbf{x} + \phi_\omega)$$

where $\mathbf{t}$ is a random unit vector. This is a convenient target distribution as the integral of a sinusoid under a Gaussian is known analytically:

$$\int_{\mathbf{x}} \sin(\omega \mathbf{t}^\top \mathbf{x} + \phi_\omega) \mathcal{N}(\mu, \Sigma) = \sin(\omega \mathbf{t}^\top \mu + \phi_\omega) \exp\left(-\frac{\omega^2}{2} \mathbf{t}^T \Sigma \mathbf{t}\right)$$

The capability for exact integration of $\int_{\mathbf{x}} \mathrm{m}_T(\mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}$ ensures that no additional variance is introduced in plots; all variance is due to the selection of the components q. In general this integral can be computed with MC methods or, in low enough dimensions, Gaussian quadrature.

In our experiments (Figures 3 and 4) we used $N = 100$ sinusoidal components in $f(\mathbf{x})$, and calculated bias using $T = 5,000$ components thinned from 4 MCMC chains of length $50,000$. To calculate variance, we subsampled $T = 10$ components from these chains, and computed variance over these random instantiations of $\mathrm{m}_{10}(\mathbf{x})$. The NUTS samples over $\mathbf{x}$ treated as ground truth derive from 4 chains of length 1,000,000.

References

S Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016. ISBN 9784431559788. URL https://books.google.com/books?id=UkSFCwAAQBAJ.

Julian Besag, Peter Green, David Higdon, and Kerrie Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–44, 1995. ISSN 08834237. doi: 10.1214/ss/1177010123.

Mark Braverman and Abhishek Bhowmick. Convexity/concavity of mutual information, September 2011. URL https://www.cs.princeton.edu/courses/archive/fall11/cos597D/L04.pdf.

Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76 (1), 2017. ISSN 15487660. doi: 10.18637/jss.v076.i01.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

James P. Hobert and George Casella. The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association*, 91(436):1461–1473, 1996. ISSN 1537274X. doi: 10.1080/01621459.1996.10476714.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

Alp Kucukelbir, David M. Blei, Andrew Gelman, Rajesh Ranganath, and Dustin Tran. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18:1–45, 2017. ISSN 15337928.

M. Magnusson, Paul-Christian Bürkner, and Aki Vehtari. posteriordb: A database of Bayesian posterior inference, 2021. URL https://github.com/stan-dev/posteriordb.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.