Interpolating Between Sampling and Variational Inference with Infinite Stochastic Mixtures

Richard D. Lange[1] **Ari S. Benjamin**[1] **Ralf M. Haefner**[*2] **Xaq Pitkow**[*3]

[1]Dept. of Neurobiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[2]Dept. of Brain and Cognitive Sciences, University of Rochester, Rochester, New York, USA
[3]Baylor College of Medicine, Rice University, Houston, Texas, USA
[*]equal contribution

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 analytically 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].

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 mixture

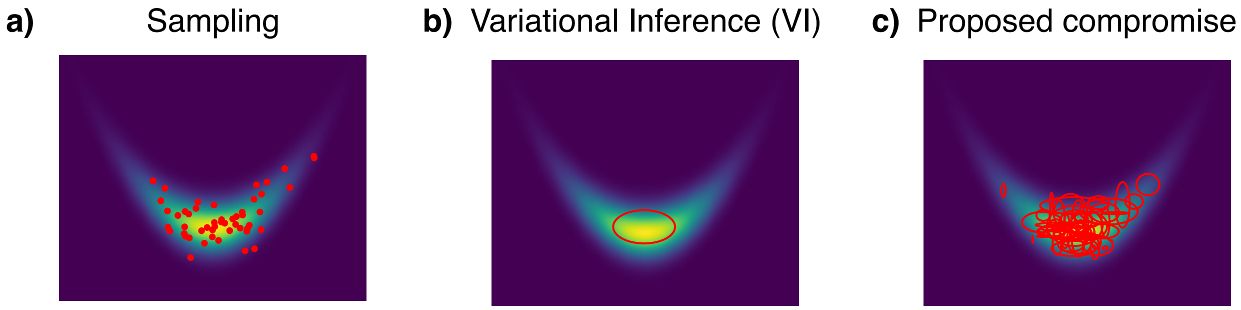

Figure 1: Conceptual introduction on a simple 2D example – the "banana" distribution. **a)** Sampling methods approximate the underlying $\mathrm{p}(\mathbf{x})$ with a stochastic set of representative points, $\mathbf{x} \sim \mathrm{p}(\mathbf{x})$. **b)** Variational Inference (VI) methods begin by selecting an approximating distribution family, $\mathrm{q}(\mathbf{x};\theta)$, here a Gaussian with diagonal covariance 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 from a "mixing distribution" $\psi(\theta)$, i.e. $\theta \sim \psi(\theta)$.

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 $\{\theta^{(1)},\dots,\theta^{(T)}\}$:

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

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, Zobay, 2014].

We will study properties of mixing distributions, or 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 components are distributed according to $\mathrm{p}(\mathbf{x})$. This requires

[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 $\mathrm{p}^*(\mathbf{x})$.

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 and Mutual Information 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 $\psi(\theta)$ can be placed as a point in this space, but in general many $\psi(\theta)$'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 $\psi(\theta)$. This objective may alternatively be viewed as the Lagrangian of a constrained optimization problem over the mixing density $\psi(\theta)$, 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. 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, 2014, Jaakkola and Jordan, 1998, Gershman et al., 2012,

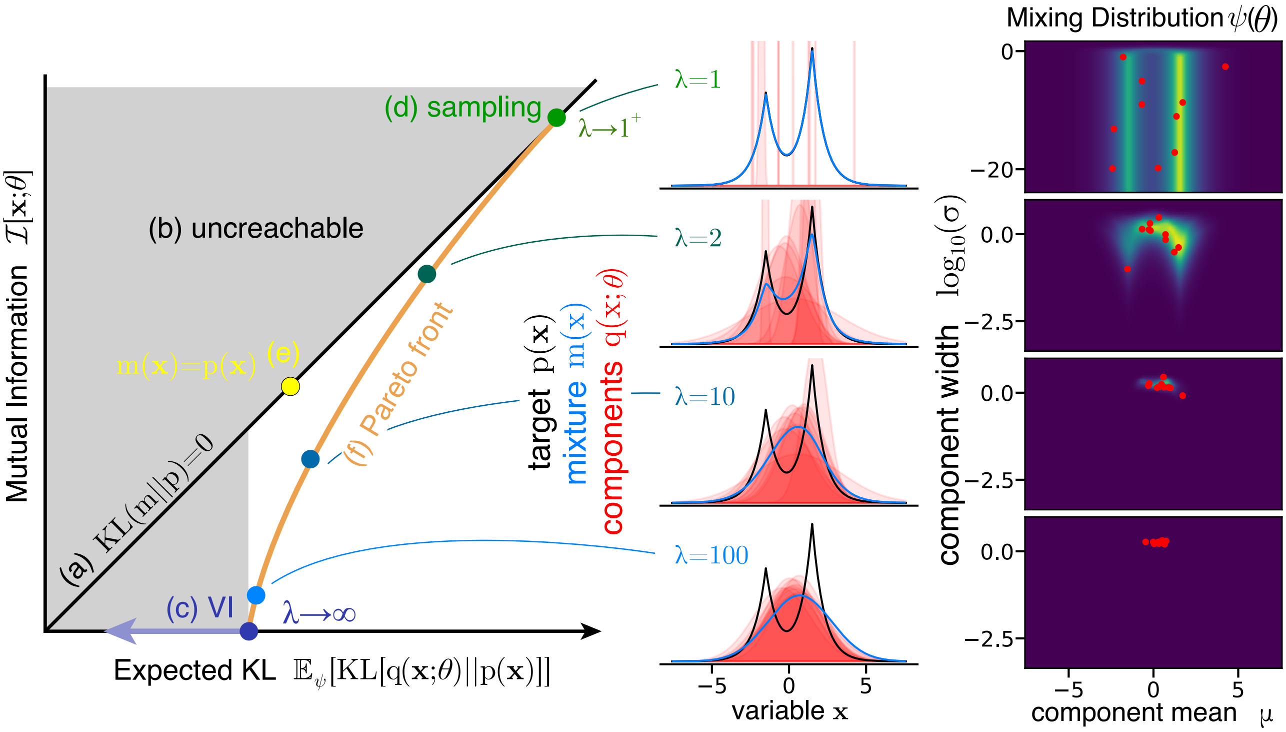

Figure 2: *Left*: Understanding mixtures in terms of Mutual Information and Expected KL. **a)** The quality of any infinite mixture (in terms of $\mathrm{KL}(\mathrm{m}||\mathrm{p})$) is given by its distance from the y=x line (black diagonal line). **b)** Two unreachable regions are shaded in gray: 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 $\theta = \{\mu, \log\sigma\}$ (translucent red components). *Right*: Varying λ controls the mixing distribution over θ (image). Red points correspond to the Gaussian components in the middle.

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 replicates 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, Zobay, 2014, Guo et al., 2016, Miller et al., 2017, Kolchinsky and Tracey, 2017, Nalisnick and Smyth, 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, leading to linear scaling with T.

4.1 STAM'S INEQUALITY

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

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

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

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

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

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

as a proxy for the intractable $\mathcal{I}[\mathbf{x};\theta]$ in (5), having used $\log|\mathcal{F}^{-1}| = -\log|\mathcal{F}|$.

Note that $\mathcal{I}_{\mathcal{F}}[\mathbf{x};\theta]$ has not been proven to be a strict *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) we suspect that the inequality in (7) is nonetheless strict, since we neglect the prior information contained in $\psi(\theta)$ and therefore overestimate the conditional entropy $\mathcal{H}[\hat{\theta}|\theta]$. 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,

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

having dropped additive constants. This now resembles a maximum-entropy problem with an expected-value constraint, which has the following simple closed-form solution:

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

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

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 $\lambda = 1$, the resulting mixture recovers the behavior of "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 it is* ***unbiased*** *and it consists of* ***non-overlapping components****. An* ***unbiased*** *mixture is one where* $\mathrm{m}(\mathbf{x}) = \mathrm{p}(\mathbf{x})$*. A mixture consists of* T ***non-overlapping components*** *if* $\sum_{t=1}^{T}\mathrm{q}(\mathbf{x};\theta_t) \approx \max_t \mathrm{q}(\mathbf{x};\theta_t)$ *with high probability.*

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, though computing and differentiating through $\mathcal{F}(\theta)$ for non-Gaussian component families poses additional challenges.

4.3 IMPLEMENTATION

Equation (10) provides a closed-form unnormalized log probability density, which is straightforward to sample from using any of a large number of existing sampling methods. For example, discrete Langevin dynamics are

$$\theta^{(t+1)} = \underbrace{\theta^{(t)} - \gamma\lambda\nabla_\theta\mathrm{KL}(\mathrm{q}||\mathrm{p})}_{(i)} + \underbrace{\frac{\gamma}{2}\nabla_\theta\log|\mathcal{F}|}_{(ii)} + \underbrace{\sqrt{2\gamma}\eta_t}_{(iii)}$$

where γ is the step size and η_t is unit isotropic Gaussian noise. This update rule is remarkably simple: (i) is equivalent to gradient descent of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$, as done in ADVI [Kucukelbir et al., 2017], (ii) biases the updates towards regions where $|\mathcal{F}(\theta)|$ is large (i.e. narrower components), and (iii) adds noise.

For our experiments below, we implemented sampling from (10) in Stan [Carpenter et al., 2017], an open-source framework for probabilistic models and approximate inference algorithms. We set q to be a multivariate Gaussian with diagonal (axis-aligned) covariance, and 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 a variety of off-the-shelf sampling methods. We computed the $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ term using 200 random samples from q per evaluation, using the reparameterization trick to compute the gradient $\nabla_\theta \mathrm{KL}(\mathrm{q}||\mathrm{p})$ and resampling the reparameterization noise only once per NUTS trajectory. This incurred a high cost in terms of number of function evaluations per sample of θ, but this cost can in principle be significantly reduced by using a sampler that accepts stochastic function evaluations [Korattikara et al., 2014, Ma et al., 2015]. All comparisons to existing methods were with Stan's built-in NUTS sampler (over $\mathbf{x}$) and its built-in Automatic-Differentiation VI (ADVI) [Kucukelbir et al., 2017].

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. 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, $\mathrm{m}_T(\mathbf{x})$. For ADVI, we measured variance across runs with different random initializations. In Figure 3, we show empirically that by increasing λ one can interpolate between the low bias but high variance solution, equivalent to sampling, and the low variance but high bias solution, equivalent to VI. Between these extremes, our method smoothly interpolates both bias and variance. Further implementation details can be found in Appendix B.

Computing bias and variance requires choosing a class of functions $f(\mathbf{x})$. We construct random smooth functions 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: $a_\omega = \omega^{-\alpha}$. Note that by choosing a random direction $\mathbf{t}$ for each frequency, it is easy to apply this definition of a random function to arbitrarily high-dimensional inference problems. An example of a 2-dimensional $f(\mathbf{x})$ is shown in Figure 3b with $\alpha = -1.5$. We vary the smoothness of the integrated function in Figure 4 by varying α [Stein and Shakarchi, 2011].

We also tested our algorithm on three reference problems from posteriordb [Magnusson et al., 2021], as well as on a 32-dimensional regression problem with synthetic data and known ground-truth parameters, shown in Supplemental 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 $\frac{\partial \mathrm{bias}^2}{\partial \lambda} = -\frac{\partial \mathrm{var}}{\partial \lambda}$. Any factor that increases the variance but not the bias for a fixed number of components T will push the optimal λ towards higher values (closer to VI).

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, 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 and computed their expectations them over random mixtures. The resulting MSE, bias, and variance are shown in Figure 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 ($\alpha \to -1$). With all else held equal, it is better to trade some variance for bias when the integrand changes quickly with $\mathbf{x}$.

In Figures 3 and 4, we evaluated bias and variance on the "banana" distribution. Similar results on higher-dimensional problems can be found in Supplemental Figure B.1. For illustration purposes, we chose T so that the variance of NUTS was the same order of magnitude as the bias of ADVI, and estimated bias and variance by randomly subsampling sets of size T from much longer chains. This approach provides theoretical insights on bias/variance trade-offs for T *independent* samples, but in practice bias will be higher due to burn-in time, variance will be higher due to sampler autocorrelations, and each of these may depend nontrivially on λ. Because MCMC samplers are most effective when they have been tuned to the problem at hand, a challenge for future work will be to adapt the sampler parameters on the fly as λ changes.

Another factor that affects the optimal λ is the computational

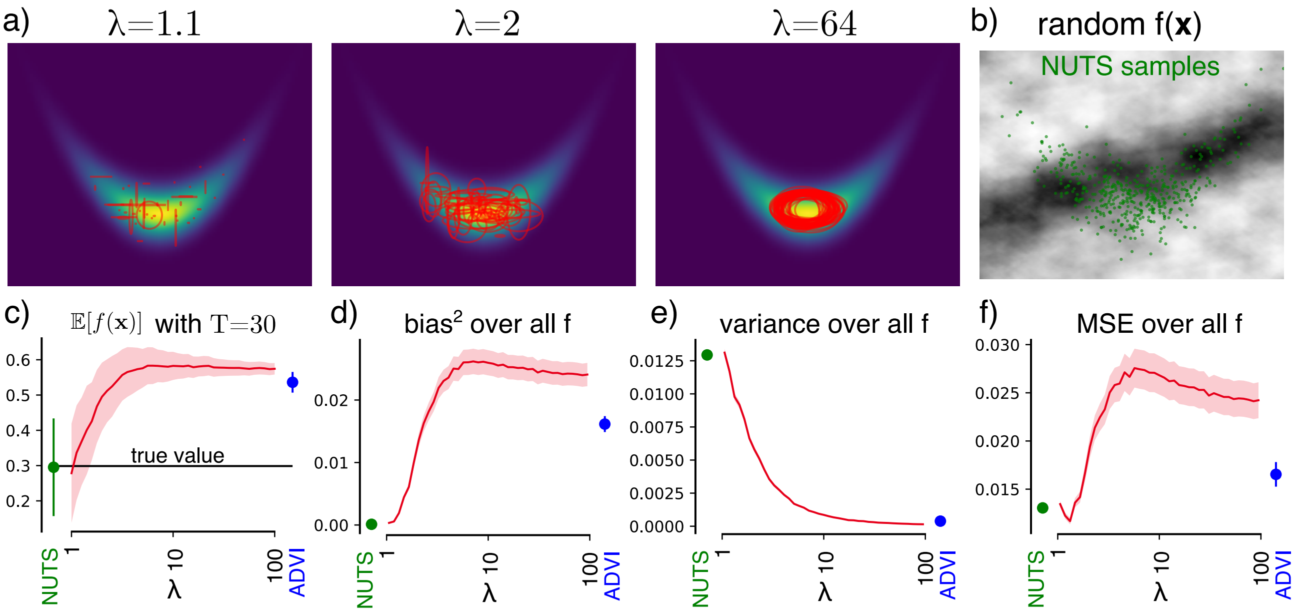

Figure 3: λ controls a bias/variance tradeoff, interpolating between sampling and VI. **a-c)** Behavior of our method on the "banana" distribution for low, medium, and high values of λ. **b)** Example $f(\mathbf{x})$ with $\alpha = -1.5$, constructed using a random mixture of sinusoids of different frequencies and directions. Green points are values of $\mathbf{x}$ sampled using NUTS, shown for reference. **c)** The expected value of $f(\mathbf{x})$ from (b) using our method, compared with NUTS and ADVI. Error bars for NUTS and ours indicate standard deviation across runs with $T = 30$ samples each. At low λ, our method provides an unbiased but high variance estimate of $\mathbb{E}[f(\mathbf{x})]$, matching NUTS, while at high λ it provides a bias near that of ADVI and a vanishing variance. Variance of ADVI is across 10 runs with random initializations. **d-f)** We repeated the analysis in (c) across many random fs (all $\alpha = -1.5$) and report the mean ± standard error of bias^2, variance, and MSE. MSE for our method is minimal around $\lambda \approx 1.3$ (for $T = 30, \alpha = -1.5$).

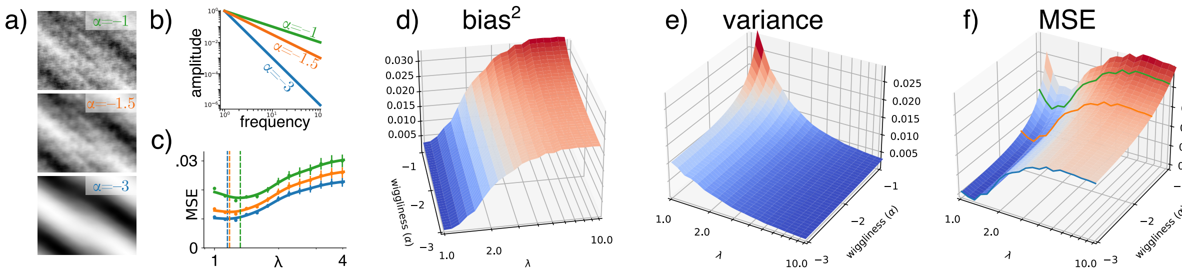

Figure 4: Interactions between λ and integrand wiggliness for the "banana" distribution with $T = 30$. **a)** Example random integrands, $f(\mathbf{x})$, of varying degrees of wiggliness, as in Figure 3b. **b)** Wiggliness of f is governed by α, the slope of amplitude versus frequency of its component sinusoids in a log-log plot. **c)** The λ with the smallest MSE for a fixed number samples depends on the integrand's smoothness. **d)** Bias vanishes near $\lambda = 1$. **e)** Variance is higher for smaller λ and more wiggly integrands. **f)** MSE is the sum of bias^2 and variance. Overlaid lines correspond to slices shown in panel (c).

budget. In our experiments we set a fixed T to demonstrate our algorithm's properties. However, if the time budget is not known in advance, a practitioner may wish to decrease λ adaptively over time. Since variance is $\mathcal{O}(T^{-1})$, for sufficiently large T error is dominated by bias, and so the optimal λ will decay towards 1. This would result in VI-like behavior for small T and sampling-like behavior for large T. How quickly λ should decay will depend on the particular problem, specifically, on $\frac{\partial \mathrm{Bias}^2}{\partial \lambda}$, and on how efficiently one can produce T *independent* samples of θ.

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 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, we define **KL bias** as the KL divergence from the infinite mixture $\mathrm{m}(\mathbf{x})$ to the true distribution, and **KL variance** as 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 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. *Proof: see Appendix A.2.*

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 $\psi(\theta)$ we derived based on the approximate $\mathcal{L}_{\mathcal{F}}$ objective. For this theorem to apply, we must further show that both conditions of "sampling" (Definition 1) are met by $\psi(\theta)$ 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 our method. However, this result is slightly more subtle, as there are three cases where one should expect VI to be (locally) 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. Similarly, if q is not in the same family but the VI solution is sufficiently close to p, then a mixture of nearby qs will add variance to m [Lindsay, 1983], potentially making the match to p worse. Third, if T is small – in the most extreme case, if $T = 1$ – then reducing λ will again only add variance without reducing bias. With these three cases 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* $\mathrm{q}(\mathbf{x};\theta^*)$ *is poorly matched to* $\mathrm{p}(\mathbf{x})$, *in the sense that* *Tr* $\left((\nabla^2_\theta \mathrm{KL}(\mathrm{q}||\mathrm{p}))^{-1}\mathcal{F}\right) > |\theta|$, *and that* λ *is sufficiently large to use a Laplace approximation to* $\psi(\theta)$ *around* θ^*. *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.*

Here, $|\theta|$ is the dimensionality of θ. The requirement that "$\mathrm{q}(\mathbf{x};\theta^*)$ is poorly matched to $\mathrm{p}(\mathbf{x})$" is expressed in terms of the curvature of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ around θ^*; if this curvature is small, then many "nearby" qs will also fit p well, and a mixture of them can improve on VI despite adding variance. On the other hand, the case where this curvature is not small corresponds to the earlier intuition that VI cannot be improved upon if the VI solution is already close to p. For further details, see the full proof in 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 in terms of trading off bias and variance. 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. Our method also provably reduces the bias of VI under certain intuitive conditions.

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, Zobay, 2014, Guo et al., 2016, Nalisnick and Smyth, 2017, 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 [Salimans et al., 2015, 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 from a point to a distribution, our approach allows each sample to cover more space with less variance and greater efficiency.

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 such mixtures [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]).

The trade-offs described in this section are summarized in Table C.1 in the Appendix.

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. Our proof of Theorem 2 requires an assumption about the curvature of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ near the VI solution; identifying cases where this assumption holds may also be an interesting future direction.

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. Other sampling algorithms besides NUTS may also be beneficial, such as ULA, which is known to have favorable scaling properties in higher dimensions [Durmus and Moulines, 2017].

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].

Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-21-1-0422, NSF CAREER IOS-1552868, and support from the McNair Foundation to XP, as well as support from the NIH R01 EY028811 to RMH. Thanks also to Daniel Lee, who acted as our guide to the Stan codebase.

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.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551–1587, 2017. doi: 10.1214/16-AAP1238. URL `https://doi.org/10.1214/16-AAP1238`.

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. *arXiv*, 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.