Interpolating Between Sampling and Variational Inference with Infinite Stochastic Mixtures (Supplementary material)

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

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

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

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

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

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

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 assumption 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 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.*

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, consider a coordinate system where 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 (e.g. additionally parameterizing the orientation of Σ with a rotation matrix). 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 $\theta \sim \psi(\theta)$ (in practice, we set a lower bound on $\log\sigma$ for numerical stability). Since components will become arbitrarily narrow with high probability, we have that $\mathrm{q}(\mathbf{x};\theta^{(j)}) << \mathrm{q}(\mathbf{x};\theta^{(i)})$ in the region where $\mathrm{q}(\mathbf{x};\theta^{(i)})$ has appreciable mass. This means that the mixture will consist of **non-overlapping components** when $\lambda = 1$.

The fact that each component shrinks towards a Dirac delta with high probability then implies that the mixture will be unbiased. To see this, 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

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

In other words, when components are narrow, the distribution of means $\boldsymbol{\mu}$ according to ψ will match the true distribution p. 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** when $\lambda = 1$. ■

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: Our approach will be to calculate the variational derivatives of KL bias and KL variance 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

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

[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.

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

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

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 KL bias with respect to ψ, evaluated at θ is[3]

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

To get the sensitivity of KL bias to λ at the point $\lambda = 1$, 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 KL 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 KL variance with respect to ψ, where

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

using the shorthand $\mathbb{E}_{1..T}[\ldots]$ to denote an expectation over independent draws of $\{\theta^{(t)}\} \sim \psi(\theta)$, for each of $t = \{1..T\}$. Applying the assumption of **non-overlapping components**, the sum inside the log is dominated by its maximum. We can therefore approximate KL variance as

$$\text{KL variance} \approx \mathbb{E}_{1..T}\left[\int_{\mathbf{x}}\frac{1}{T}\sum_{i=1}^{T}\mathrm{q}(\mathbf{x};\theta^{(i)})\log\frac{\frac{1}{T}\mathrm{q}(\mathbf{x};\theta^{(i)})}{\mathrm{m}(\mathbf{x})}\mathrm{d}\mathbf{x}\right] ,$$

since the maximum of the sum over j inside the log will be the ith component from outside the log. This step is most applicable for small to moderate T, since when T grows sufficiently large, even narrow components will overlap each other with appreciable probability. Let us continue assuming that T is sufficiently small and that components are sufficiently non-overlapping. By symmetry, we can remove the sum over i and simplify the outer expectation to a single θ:

$$\ldots = \mathbb{E}_{i}\left[\int_{\mathbf{x}}\mathrm{q}(\mathbf{x};\theta^{(i)})\log\frac{\frac{1}{T}\mathrm{q}(\mathbf{x};\theta^{(i)})}{\mathrm{m}(\mathbf{x})}\mathrm{d}\mathbf{x}\right] ,$$

[3]This section is best viewed in color. Our notation uses red θ to indicate the value where the variational derivative is evaluated, which is distinct from from black θs, which are integrated out.

which simplifies to

$$\text{KL variance} \approx \mathcal{I}[\mathbf{x};\theta] - \log T\,.$$

The variational derivative of $\mathcal{I}[\mathbf{x};\theta]$ with respect to ψ is

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

Taking the inner product with $\frac{\mathrm{d}}{\mathrm{d}\lambda}\psi$, and applying the assumptions from the definition of sampling,

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

In other words, this says that the change in the (upper bound on) KL variance is *non-positive*, and its magnitude is 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 theorem. ■

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^*. As long as $\nabla^2\mathrm{KL}(\mathrm{q}||\mathrm{p})$ is positive definite, which is guaranteed by the assumption that θ^* is unique, the accuracy of this Laplace approximation can be made arbitrarily good by considering larger and larger λ.

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 ψ 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: As λ grows, the Laplace approximation in (A.4) becomes increasingly accurate, and increasingly narrow. Thus, for sufficiently large λ, we can accurately 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[4] 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 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

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

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

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

Next we will calculate the Hessian of $\mathrm{KL}(\mathrm{m}^*_T(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))$. Note that the derivatives are with respect to θ, not θ^*. First, the Hessian of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ is

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

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

[4]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.

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}}\left(\frac{\frac{\partial}{\partial\theta_i}\mathrm{q}(\mathbf{x};\theta)\frac{\partial}{\partial\theta_j}\mathrm{q}(\mathbf{x};\theta)}{\mathrm{q}(\mathbf{x};\theta^*)}\right)\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}\left(\mathcal{F}(\theta^*)+M(\theta^*)\right)+\left(\frac{1}{T^2}-\frac{1}{T}\right)\mathcal{F}(\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

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

and we found that (A.6) could be written in terms of (A.5). To reduce clutter temporarily, let $\mathbf{H}=\nabla_\theta^2\mathrm{KL}(\mathrm{q}||\mathrm{p})$. Combining terms, we have

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

where $\mathbf{I}$ is the identity matrix and $d=\mathrm{Tr}(\mathbf{I})$ is the dimensionality of θ. Consider the case where $T=1$: the KL error simplifies to $\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta^*)||\mathrm{p}(\mathbf{x}))+\frac{d}{2\lambda}$. 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, as this would imply that using a finite λ would improve on VI. This is equivalent to asking when the following inequality holds:

$$\frac{\mathrm{Tr}\left(\mathbf{H}^{-1}\mathcal{F}\right)}{d}>\frac{T}{T-1}\,.$$

Recall that $\mathbf{H}$ was defined as the Hessian of $\mathrm{KL}(\mathrm{q}(\mathbf{x};\theta)||\mathrm{p}(\mathbf{x}))$, so this is

$$\frac{\mathrm{Tr}\left((\nabla_\theta^2\mathrm{KL}(\mathrm{q}||\mathrm{p}))^{-1}\mathcal{F}\right)}{d}>\frac{T}{T-1}\,.$$

The Fisher Information Matrix can also be derived from a local quadratic approximation to $\mathrm{KL}(\mathrm{q}||\mathrm{q})$; this means that in the case where the VI solution is exact, or $\mathrm{q}(\mathbf{x};\theta^*) = \mathrm{p}(\mathbf{x})$, the trace term becomes $\mathrm{Tr}(\mathcal{F}^{-1}\mathcal{F})$, and the inequality is $1 > \frac{T}{T-1}$. This inequality is not satisfied by any positive integer T, and so this expression captures the intuitive condition that VI cannot be improved upon by reducing λ – for any finite T – if the VI solution is already exact.

Conversely, the ratio

$$R \equiv \frac{\mathrm{Tr}\left((\nabla^2_\theta \mathrm{KL}(\mathrm{q}||\mathrm{p}))^{-1}\mathcal{F}\right)}{d}$$

may be thought of as quantifying the extent to which $\mathrm{p}(\mathbf{x})$ is over-dispersed relative to the VI solution $\mathrm{q}(\mathbf{x};\theta^*)$. If the curvature of $\mathrm{KL}(\mathrm{q}||\mathrm{p})$ is low, then many "nearby" qs would fit p almost as well, and this will be reflected in this ratio being larger than 1. Then, the minimum T for which reducing λ improves KL error relative to VI can be found by solving the above inequality, giving which gives

$$T_0 = \left\lceil \frac{R}{R-1} \right\rceil , \tag{A.7}$$

so that for all $T > T_0$, we have the desired property that $\frac{\mathrm{d}}{\mathrm{d}\lambda}$KL error > 0, which implies that using some finite $\lambda < \infty$ will reduce error relative to VI. ■

B NUMERICAL DETAILS

All code to generate the figures in this paper is available at `https://github.com/wrongu/sampling-variational-demos`.

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.

We used two toy distributions in the main paper:

- 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

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

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`, `eight schools centered`, and `garch11`. These problems are 7–, 10–, and 4–dimensional problems, respectively. Finally, we synthesized ground-truth data from a hierarchical regression problem with 30 regressors (a total of 32 parameters) to test how our algorithm would scale to an even higher dimensional problem. Results for all of these additional experiments 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. Because variance scales like $1/T$, we estimated variance for other values of T simply by scaling the variance from the $T = 10$ case. We chose $T = 30$ (scaling variance from $T = 10$ by a factor of $1/3$) for the plots using the banana distribution in the main paper because this gave roughly equal magnitude to the variance of NUTS and the bias of ADVI, which makes the effects of the trade-off between them most pronounced. In Figure B.1, we chose T separately for each problem using the same strategy.

C DETAILED COMPARISON TO PRIOR WORK

Author (Year)	**Component parameters**	**Component family**	**Auxiliary Optimization**	**Time; Space Complexity**	**Recovers Sampling**
Jaakkola and Jordan [1998] "Mixture of Mean Field"	optimized	mean field	minor	$\mathcal{O}(T^2)$; $\mathcal{O}(T)$	$T \to \infty$
Salimans et al. [2015] "Bridging the Gap"	sampled (implicit)	flexible	NN optimization (k steps)	$\mathcal{O}(k+T)$; $\mathcal{O}(T)$	unknown
Gershman et al. [2012] "Nonparametric VI"	optimized	Guassian	none	$\mathcal{O}(T^2)$; $\mathcal{O}(T)$	$T \to \infty$
Zobay [2014] "VI with Gaussian Mixtures"	optimized	Gaussian	none	$\mathcal{O}(T^2)$; $\mathcal{O}(T)$	$T \to \infty$
Guo et al. [2016], Miller et al. [2017] "Boosting VI"	optimized	flexible	none	$\mathcal{O}(T^2)$; $\mathcal{O}(T)$	no
Nalisnick and Smyth [2017] "Stein Mixtures"	optimized	flexible	none	$\mathcal{O}(T^2)$; $\mathcal{O}(T)^*$	$T \to \infty$
Yin and Zhou [2018] "SIVI"	sampled (implicit)	Gaussian*	NN optimization (k steps)	$\mathcal{O}(kT^2)$; $\mathcal{O}(T)^*$	unknown
Acerbi [2018, 2020] "VBMC"	optimized	Gaussian	fit GP (k steps)	$\mathcal{O}(kT^2)$; $\mathcal{O}(T)$	no
Ours	sampled (closed-form)	Gaussian*	none	$\mathcal{O}(T)$; $\mathcal{O}(T)^*$	$\lambda \to 1$

Table C.1: Comparison of our proposed algorithm to a number of existing methods that, in some way or another, use a mixture of "simple" component distributions for approximate inference. **Component parameters:** how are the parameters of the individual mixture components chosen? In terms of minimizing bias and variance, it is best to jointly optimize the location of all T components together. The approach of Salimans et al. [2015] and Yin and Zhou [2018] is similar to ours in that mixture components are stochastically sampled, though ours is the only stochastic method for which we have the mixing distribution $\psi(\theta)$ explicitly in closed form. **Component family:** what is the allowable form of $\mathrm{q}(\mathbf{x}; \theta)$? Methods marked "Gaussian*" (including ours) are in principle applicable to a wider class of distributions, but so far only demonstrated empirically using Gaussian mixtures. **Auxiliary optimization:** Are there additional parameters of the inference process itself that need to be fit or optimized at inference-time? The advantage of our method is that, since we derived $\psi(\theta)$ in closed form, we can begin sampling from it immediately without further optimizations. **Time; Space Complexity:** Note that all methods for which component parameters are "optimized" incur a $\mathcal{O}(T^2)$ cost in time complexity, since the optimal location of each component depends on the location of other components. Methods that require auxiliary optimization (such as training a neural network or NN) incur additional runtime costs. Methods marked $\mathcal{O}(T)^*$ space-complexity can in principle be "streamed," in which case they use constant $\mathcal{O}(1)$ space. **Recovers sampling:** some methods "look like" sampling (unbiased and with narrow components) only in the $T \to \infty$ limit, which is infeasible given $\mathcal{O}(T^2)$ time complexity. Ours is the only method we know of for which sampling-like behavior can be recovered, independent of T, by setting $\lambda = 1$.

References

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

Luigi Acerbi. Variational Bayesian Monte Carlo with Noisy Likelihoods. *arXiv*, 2020. ISSN 10495258.

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.

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.

Fangjian Guo, Xiangyu Wang, Kai Fan, Tamara Broderick, and David B. Dunson. Boosting Variational Inference. *arXiv*, 2016. URL http://arxiv.org/abs/1611.05559.

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.

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.

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.

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.

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.

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

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.

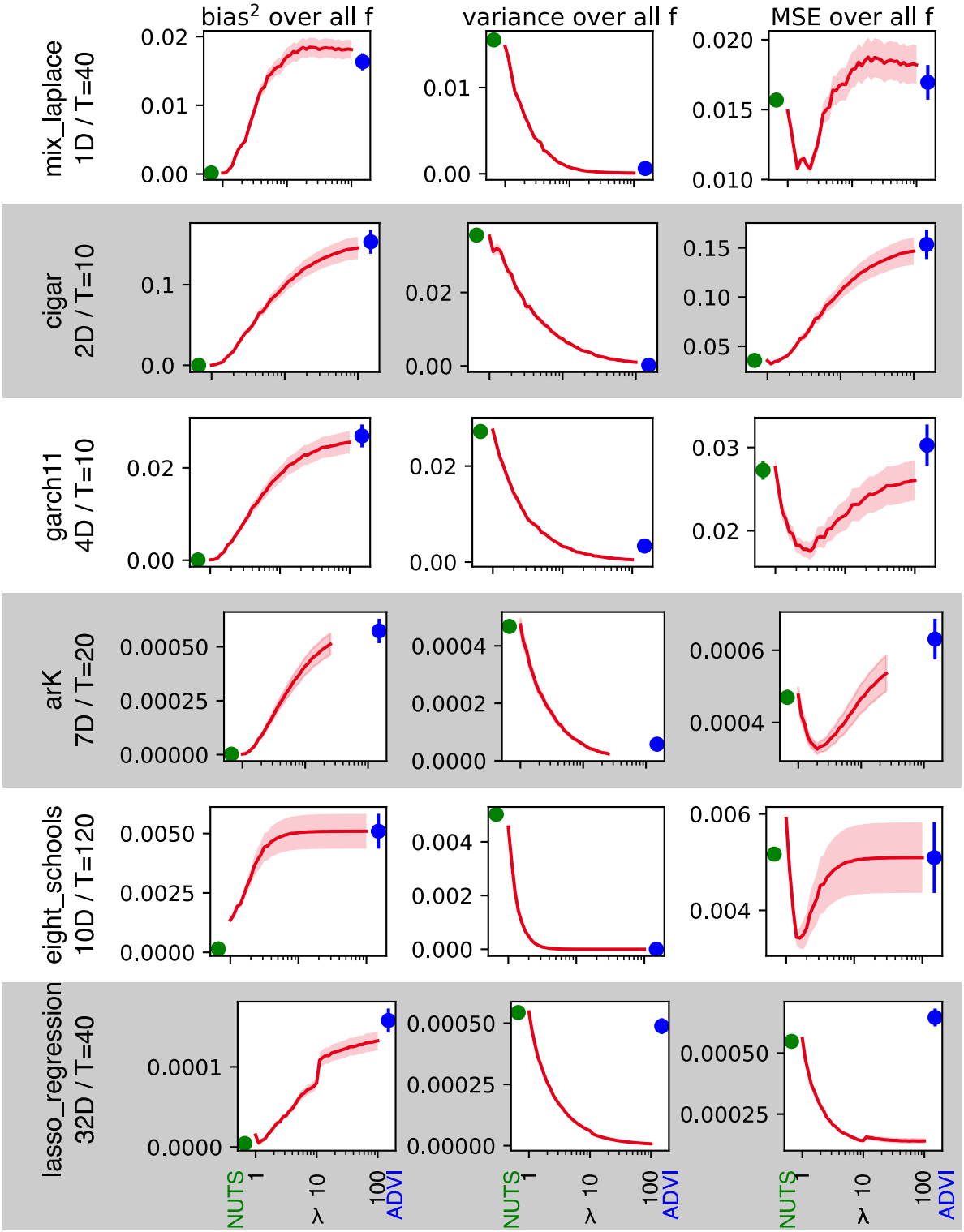

Figure B.1: Bias/variance decomposition across many random fs (all with $\alpha = -1.5$) for additional test problems, plotted in the same format as Figure 3d-f. Each row corresponds to a different inference problem. We chose T separately for each problem such that the variance of NUTS and bias of ADVI were of similar orders of magnitude. **`mix laplace`** is the 1D mixture of two Laplace distributions shown in Figure 2 of the main paper. **`cigar`** is a 2D Gaussian with a correlation of 0.99, oriented along the $y = x$ axis . **`garch11`**, **`eight_schools`**, and **`arK`** refer to problems taken from posteriordb [Magnusson et al., 2021]. **`lasso regression`** is a regression problem with 30 regressors and a double-exponential (Laplace) prior on each regression weight. We report the variance of ADVI using all default settings, which fails to converge reliably for this problem, and hence has significant variance. Optimizing the ADVI settings for this problem could potentially result in much lower variance.