Yale University, Department of Statistics

Variational inference review

August 30, 2019

Contents

1 Introduction

This note gives some background information on variational inference. Consider the posterior inference problem: we observe data $x_1, \ldots, x_n \sim_{i.i.d.} \mathbb{P}_\theta$, a distribution with density $p(x|\theta)$. We place a prior $p(\theta)$ on $\theta \in \mathbb{R}^d$, and we estimate the posterior $p(\theta|x)$. Variational methods for this problem posit a family $\mathcal{Q}$ for the posterior density and optimize a measure of divergence between $q \in \mathcal{Q}$ and the true $p(\theta|x)$.

Variational methods are useful tools for estimating posterior distributions in high-dimensional settings. Markov chain Monte Carlo methods are often used, but these can be infeasible for large datasets. It is common to choose q to minimize the evidence lower bound,

$\mathcal{L}(q) := \mathbb{E}_q \log p(x, z) + H(q)$. As we show in Lemma **??**, this is equivalent to minimizing TK, and there are other equivalences and interesting interpretations.

A mean-field family, in which $q(\theta) = \prod_{i \leq d} q_i(\theta_i)$, is often posited for $\mathcal{Q}$. When the likelihood and prior are conjugate, this results in a simple coordinate ascent algorithm in which the updates are computable in closed form. In non-conjugate models, the Laplace method, which assumes a Gaussian family for $\mathcal{Q}$ and expands $\log p(x, \theta)$ to a quadratic term, is a popular approach; see Section **??** for more details.

In these and other standard methods, restrictive families are chosen for $\mathcal{Q}$ because of their computational convenience. But such methods have serious weaknesses. For instance, restricting $\mathcal{Q}$ to be a mean-field family means the posterior estimate will fail to capture dependence among the posterior coordinates and will underestimate the posterior variance; see Blei *et al.* (2017) for more discussion. The Gaussian form imposed on q in the Laplace method means the estimate will fail to capture multi-modality in the posterior.

Recent literature has developed methods to allow $\mathcal{Q}$ to be a richer family. For instance, the authors of Ranganath *et al.* (2016) use a dual representation of a divergence and a generative form of q to avoid the use of the density itself in the variational optimization. Their form of q is a multi-layer network; although it is still parametric, it is quite flexible. However, their algorithm requires optimization over both the parameters of q and the function in the dual representation. It would be ideal to have a method that does not require such a procedure. The authors of Gershman *et al.* (2012) propose "nonparametric variational inference," which uses a Gaussian-mixture form of q and the Delta method to approximate the model. The mixture of Gaussians for q resembles a kernel density estimator from nonparametric estimation. Moreover, this family for q, along with the Delta method, makes the optimization process fairly simple. The Gaussian-mixture family for q is also related to the "mixture of mean-field" family proposed by Bishop *et al.* (1998). See Section 5 for more details.

Ideally, we would be able to do "completely nonparametric" variational inference. That is, we would like to assume $\log q(\theta) \propto f(\theta)$ where $f(\theta) = \sum_{j=0}^{\infty} \gamma_j \phi_j(\theta)$, with ϕ_j being an orthogonal basis and γ_j being the terms we estimate. But we haven't found a way to do this. If we optimize $KL(q||p(\theta|x))$, we need to approximate $\mathbb{E}_q \log p(x, \theta) - \mathbb{E}_q \log q$, but we don't know how to take expectations with respect to such a family q. This same problem appears with basically any usable divergence; see Section **??** for more discussion.

Table 1: MCMC vs variational methods for posterior inference.

Method	Some details	Advantages
MCMC	Sample from an approximate posterior	We are guaranteed to converge since are sampling from MC with stationary dist equal to what we want. Slow because of sampling; not feasible for high dimensions/large data. Not parallelizeable.
Variational methods	Optimize. Estimate exactly an approximation of p	Typically much faster than MCMC, can handle larger data.

Table 2: Classical variational inference (minimizing $KL(q||p)$. Let $\mathcal{L}(q) = \mathbb{E}_q \log p(x,z) + H(q)$.

Fact	Notes
$\mathcal{L}(q) \leq \log p(x)$	Thus we call $\mathcal{L}(q)$ the "evidence lower bound."
$\mathcal{L}(q) \stackrel{c}{=} -KL(q\|\|p(z\|x))$	Minimize the KL divergence between our estimate and the truth.
$\mathcal{L}(q) = \mathbb{E}_q \log p(x,z) + H(q)$	Minimize cross-entropy (max joint density expectation under q) maximize entropy (which is large variance if q Gaussian). Bias-variance tradeoff; make likelihood large but prevent overfitting by penalizing if variance of q too small.
$\mathcal{L}(q) = \mathbb{E}_q \log p(x\|z) - KL(q(z)\|\|p(z))$	Max likelihood while minimizing KL between q and prior; again, bias-variance tradeoff.

Lemma 1.1. *The inequalities in Table 1 on page 3 hold.*

Proof. Use Jensen:

$$\begin{aligned}\log p(x) &= \log \mathbb{E}_{q(z)} \left(\frac{p(x,z)}{q(z)} \right) \\ &\geq \mathbb{E}_{q(z)} \left(\log p(x,z) - \log q(z) \right)\end{aligned}$$

And

$$\begin{aligned} KL\left(q||p(z|x)\right) &= KL\left(q||p(x,z)\right) + \log p(x) \\ &= -\mathbb{E}_q \log p(x,z) + \mathbb{E}_q \log q + \log p(x) \\ &= -\mathcal{L}(q) + \log p(x) \end{aligned}$$

Now $p(x)$ is constant with respect to q, so minimizing this in q is equivalent to minimizing $KL\left(q||p(x,z)\right)$. And

$$\begin{aligned} \mathcal{L}(q) &= \mathbb{E}_q \log p(x|z) + \mathbb{E}_q \log p(z) - \mathbb{E}_q \log q(z) \\ &= \mathbb{E}_q \log p(x|z) - KL(q(z)||p(z)) \end{aligned}$$

□

2 Algorithms

Here are some more details on what variational algorithms actually look like. In the typical mean-field version, we estimate a separate q_i for all $i \in [n]$, i.e., every variable z_i has its own density. I sometimes refer to this as "non-parametric." Alternatively, we might do a mean-field parametric form, in which we have e.g. $q_\phi(z_i)$ where q is the same for all z. One term I'll use is:

Definition 1 (Factorized parametric density/amortized mean-field density)**.** *Suppose we have one z_i per x_i. The factorized parametric density is $q(z|x) = \prod_{i \leq n} q_\phi(z_i|x_i)$ where ϕ is the same for all $i \in [n]$.*

$$\begin{aligned} q_\phi &= N(\lambda_{\phi_1}, V_{\phi_2}) \text{ where e.g. for some non-linear } \sigma, \\ \lambda_{\phi_1}(x) &= \sigma(Ax + b) \\ V_{\phi_2}(x) &= \sigma(Cx + d) \end{aligned}$$

That is, $\lambda_{\phi_1}, V_{\phi_1}$ are potentially non-linear, parametric functions whose parameters are estimated using the data. So for z_u, the estimated posterior is Gaussian with mean

$$\hat{\lambda}_{\phi_1}(x_u) = \sigma(\hat{A}x_u + \hat{b})$$

and so on. Note that $\hat{A} = \hat{A}(X)$, a function of all the data, and similarly for $\hat{b}, \hat{C}, \hat{d}$.

Table 3: Alternatives for variational algorithms.

Object	Examples of choices
Form of q	E.g., parametric (as in VAE's) or non-parametric (as in mean-field)
Form of prior	Parametric (as in CTM) or non-parametric (as in VAE's)
Integral approximation	Laplace/Delta, linear approximation, sampling. See Section 6.

Table 4: Ordinary and Variational EM.

EM	V-EM
$M(\theta\|\theta^t) = \mathbb{E}_{p(z\|x,\theta^t)} \log p_\theta(x,z)$	$\hat{q} = \text{argmax}_q \, \mathcal{L}(q,\theta)$
$\hat{\theta} = \text{argmax}_\theta \, M(\theta\|\theta^t)$	$\hat{\theta} = \text{argmax}_\theta \, \mathbb{E}_{\hat{q}} \log p_\theta(x,z) = \text{argmax}_\theta \, \mathcal{L}(\hat{q},\theta)$
	If $\theta = \theta_2, \hat{\theta}_2 = \text{argmax}_{\theta_2} \, \mathbb{E}_q \log p_{\theta_2}(x\|z)$

3 Mean-field and conjugacy

Most literature uses a mean-field family for $\mathcal{Q}$, that is, the family $\mathcal{Q}$ consists of densities of the form

$$q(\theta) = \prod_{i \leq d} q_i(\theta_i) \tag{1}$$

Then the optimal q_i is:

$$q_i^*(\theta_i) \propto \exp \mathbb{E}_{-i} \log p(x,\theta)$$

See Lemma 3.1 for a proof. Moreover, in cases of conjugacy, the expectations can be computed exactly, resulting in a simple algorithm known as coordinate-ascent variational inference. This arises in several models of the form θ_i, z_j etc.

While the mean-field assumption is convenient, it has several weaknesses. It contributes to underestimation of the posterior marginal variances, and it by definition fails to capture any covariance among the posterior variables.

Lemma 3.1. *Assume we have latent variables $z_1, \ldots, z_m$, and that $q(z) = \prod_{j \leq m} q(z_j)$.*

$$q^*(z_i) \propto \exp\left(\mathbb{E}_{z_{-i}} \log p(x,z)\right) \tag{2}$$

$$\tag{3}$$

Proof. I will actually do the proof in a simple case where we have two latent variables, z and θ. This is just to simplify notation.

$$\mathcal{L}(q(z), q(\theta)) = \mathbb{E}_{q(\theta)} \left(\mathbb{E}_{q(z)} \log p(x,z,\theta) - \log q(\theta) - \mathbb{E}_{q(z)} \log q(z)\right)$$

There is a restraint that $\int q(\theta) \, d\theta = 1$ and similarly for $q(z)$. Using Lagrange multipliers, the objective is

$$\mathcal{L}(q) - \lambda_1 \int q(\theta) \, d\theta - \lambda_2 \int q(z) \, dz$$

Now inside the integral, we have that the derivative with respect to $q(\theta)$ is

$$\mathbb{E}_{q(z)} \log p(x,z,\theta) - \log q(\theta) - 1 - \lambda_1 = 0$$

[To be correct, we actually need to use the Euler-Lagrange formula. But it boils down, in this case, to setting the derivative with respect to $q(\theta)$ inside the integral to equal zero.] So

$$q(\theta) \propto \exp \mathbb{E}_{q(z)} \log p(x, z, \theta)$$

And similarly for $q(z)$. □

Lemma 3.2. *In the setting of Wang & Blei (2013), assume $q(z, \theta) = q(z)q(\theta)$. Then*

$$q^*(\theta) \propto \exp\left(\mathbb{E}_{q(z)} \log p(z, \theta)\right) \tag{4}$$
$$q^*(z) \propto \exp\left(\mathbb{E}_{q(\theta)} \log p(x, z|\theta)\right) \tag{5}$$

Proof. This follows directly from the mean-field optima, just via:

$$\begin{aligned} q^*(\theta) &\propto \exp \mathbb{E}_{q(z)} \log p(x, z, \theta) \\ &= \exp \mathbb{E}_{q(z)} \left(\log p(x|z) + \log p(z, \theta)\right) \\ &= \exp\left(\mathbb{E}_{q(z)} \log p(x|z)\right) \exp\left(\mathbb{E}_{q(z)} \log p(z, \theta)\right) \end{aligned}$$

And similarly for $q^*(z)$. □

4 Variational autoencoders

Variational autoencoders optimize the usual variational objective, $\mathcal{L}(q, \theta_2) = \mathbb{E}_q \log p_\theta(x|z) - KL(q_\phi(z)||p(z))$ in the variational parameters ϕ and the model parameters θ. One major difference between variational autoencoders and the typical mean-field variational inference as that VAE's assume a parametric (Gaussian) form for q and then use deep networks for the parameters, $\lambda(x), V(x)$. Usually V is assumed to be diagonal and they estimate $\log v_1(x), \ldots, \log v_d(x)$. The variational autoencoder is like a variational-EM algorithm, just with a non-parametric prior, a parametric non-linear model, a parametric variational density, and reparametrization-plus-sampling to approximate integrals. They use gradient ascent for the optimization. They optimize simultaneously in the model and posterior parameters, θ, ϕ, respectively. In variational-EM, we iterate over optimizing in ϕ, then in θ, but this should be the same thing (dynamic programming).

Table 5: VAE's.

Name	Other name	Details
Prior	$p(z)$ is an $N(0, I_d)$ density	
Decoder	posterior estimate $q_\phi(z)$	parametric
Encoder	model $p_{\theta_2}(x\|z)$	$p_{\theta_2}(x\|z) = g\left(x, \eta_{\theta_2}(z)\right)$ where g is the final layer in the network and $\eta_{\theta_2}(z)$ is some non-linear function or deep network. E.g., for continuous data, we might let $g(x, \eta_{\theta_2}(z)) = \phi(x - \eta_{\theta_2}(z))$. For discrete data, we might let $g(x, \eta_{\tilde{\theta}_2}(z)) = softmax(\eta_{\tilde{\theta}_2}(z)'\beta_x)$. Note that then, $\theta_2 = (\tilde{\theta}_2, \beta)$.

4.1 Using the Laplace method in VAE's - should we?

As mentioned above, we often run into the problem that we can't compute an integral like $\mathbb{E}_q \log p_{\theta_2}(x|z)$ analytically. As discussed in Section 6, there are many ways of getting around this. It seems most tutorials and papers on VAE's just use a Monte Carlo approximation of the integral (and reparametrize so they can take gradients with respect to the q parameters). But the Laplace method is another reasonable alternative. Here is what we would do. Let

$$f_\theta(Z) = \log p_\theta(X, Z)$$

Let $q = q_\phi$ be the $N(\lambda_\phi(X), V_\phi(X))$ density. This is going to be our variational density for all the z_u's. Since the density factorizes nicely,

$$f_\theta(Z) = \sum_u \log p_\theta(x_u, z_u)$$

Let $f(z_u) = \log p_\theta(x_u, z_u)$. For each $u \in [U]$, let $z_{u0} = \text{argmax}_{z_u} f_\theta(z_u)$. Note that $\text{argmax}_Z f_\theta(Z) = (z_{10}, \ldots, z_{U0}) \in \mathbb{R}^{d \times U}$. The objective is:

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_q \sum_u f_\theta(z_u) + \frac{U}{2} \log|V_\phi| \\
&\approx \mathbb{E}_q \sum_u \left(f_\theta(z_{0u}) + (z_u - z_{0u})' \nabla^2 f_\theta(z_{0u})(z_u - z_{0u})/2 + \log|V_\phi| \right)/2 \\
&= \frac{1}{2} \sum_u \left(f_\theta(z_u) + (\lambda_\phi(x_u) - z_{0u})' \nabla^2 f_\theta(z_{0u})(\lambda_\phi(x_u) - z_{0u}) + tr\left(V_\phi(x_u) \nabla^2 f_\theta(z_{0u})\right) + \log|V_\phi(x_u)| \right)
\end{aligned}$$

Now we optimize this in ϕ. Then we optimize it in θ. We could alternatively just sample from $q_{\hat{\phi}}$ and optimize $\mathbb{E}_{\hat{q}} \log p_\theta(x, z)$. It should be about the same either way. If we do the

former, note that the parts that the optimization in ϕ and θ involve, respectively, the terms:

$$\mathcal{L}(\phi) = \frac{1}{2}\sum_u \left((\lambda_\phi(x_u) - z_{0u})'\nabla^2 f_\theta(z_{0u})(\lambda_\phi(x_u) - z_{0u}) + tr\left(V_\phi(x_u)\nabla^2 f_\theta(z_{0u})\right) + \log|V_\phi(x_u)|\right) \tag{6}$$

and (7)

$$\mathcal{L}(\theta) = \frac{1}{2}\sum_u \left(f_\theta(z_u) + (\lambda_\phi(x_u) - z_{0u})'\nabla^2 f_\theta(z_{0u})(\lambda_\phi(x_u) - z_{0u}) + tr\left(V_\phi(x_u)\nabla^2 f_\theta(z_{0u})\right)\right) \tag{8}$$

The former is - almost - equal to:

$$\sum_u KL\left(N(\lambda_\phi(x_u), V_\phi(x_u)) \;||N(z_{0u}, -\nabla^2 f_\theta(z_0 u)^{-1})\;\right)$$

But what would be the point of doing this Laplace approximation here for VAE's? I'm not sure there is a good reason to do this. It's somewhat circular. In the typical Laplace method, we arrive at the mean-field variational approximation $N(z_{0u}, -\nabla^2 f(z_{0u})^{-1})$. Why wouldn't we just use this instead of the λ_ϕ? Somehow the two methods seem incompatible? At the end of the day, if we optimize (6) in ϕ, we are going to arrive at

$$\begin{aligned}\lambda_\phi(x_u) &\approx z_{0u}\\ V_\phi(x_u) &\approx -\nabla^2 f(z_{0u})^{-1}\end{aligned}$$

i.e., essentially the mean-field density.

5 Nonparametric variational inference

In Gershman *et al.* (2012), the authors propose to assume a Gaussian mixture form for q:

$$q(\theta) = \sum_{k\leq K}\frac{1}{K}q_k(\theta)$$

where $q_k(\theta)$ is the $N(\mu_k, \sigma_k^2 I_d)$ density and K is chosen in advance. To compute a bound on the KL divergence, first note that

$$\begin{aligned}\mathbb{E}_q \log q &= \int \frac{1}{K}\sum_{k\leq K} q_k(\theta)\log q(\theta)\\ &= \frac{1}{K}\sum_{k\leq K}\int q_k(\theta)\log q(\theta)\end{aligned}$$

And $\int q_k(\theta)\log q(\theta) \leq \log\int q_k(\theta)q(\theta)$ by Jensen. So

$$\begin{aligned}\mathbb{E}_q \log p(x,\theta) - \mathbb{E}_q \log q &= \frac{1}{K}\sum_{k\leq K}\mathbb{E}_{q_k(\theta)}\log p(x,\theta) - \mathbb{E}_q \log q\\ &\geq \frac{1}{K}\sum_{k\leq K}\mathbb{E}_{q_k(\theta)}\log p(x,\theta) - \frac{1}{K}\sum_{k\leq K}\log \mathbb{E}_{q_k} q\end{aligned}$$

Now by some calculations,

$$\mathbb{E}_{q_k} q(\theta) = \sum_{j \leq K} \exp\left(-\frac{(\mu_j - \mu_k)^2}{(\sigma_k^2 + \sigma_k^2)}\right)$$

They expand each term in the first sum in a quadratic expansion around μ_k; this is the Delta method. For each $k \in [K]$, write

$$\log p(x, \theta) = f(\theta) \approx f(\mu_k) + f'(\mu_k)(\theta - \mu_k) + \frac{1}{2} f''(\mu_k)(\theta - \mu_k)^2$$

Since $\mathbb{E}_{q_k}(\theta - \mu_k) = 0$ and $\mathbb{E}_{q_k}(\theta - \mu_k)^2 = \sigma_k^2$, the objective they optimize is

$$KL(q||p) \approx \frac{1}{K} \sum_{k \in [K]} \left(f(\mu_k) + \frac{\sigma_k^2}{2} f''(\mu_k) - \log \sum_{j \leq K} \exp\left(-\frac{(\mu_j - \mu_k)^2}{(\sigma_k^2 + \sigma_k^2)}\right)\right)$$

[Write how they do the full algorithm. Some kind of gradient ascent.]

6 Methods for integrals

6.1 Reparametrization trick

Definition 2 (Reparameterization trick)**.** *Suppose we wish to find the parameters θ to maximize $\mathbb{E}_{z \sim q_\theta} f(z)$ where the distribution q depends on θ. For some reason, e.g., that f has a complex form, we can't directly calculating this integral. So we might generate $z_i \sim_{i.i.d.} q$ for $i \in [N]$, and use the Monte Carlo approximation of the integral:*

$$\mathbb{E}_{z \sim q} f(z) \approx \frac{1}{N} \sum_{i \leq N} f(z_i)$$

But we obviously can't compute the derivatives of this with respect to θ. Now suppose for a simple example that $q_\theta(z) = N(\theta, 1)$. Letting $\xi \sim N(0, 1)$, we have $z = \mu + \xi$. We now approximate the integral via

$$\mathbb{E}_{z \sim q} f(z) = \mathbb{E}_{\xi \sim N(0,1)} f(\mu + \xi) \approx \frac{1}{N} \sum_{i \leq N} f(\theta + \xi_i)$$

where $\xi_i \sim_{i.i.d.} N(0, 1)$ for $i \in [N]$. Now as long as f is differentiable in θ, we can take the derivatives of this with respect to θ and do gradient ascent.

6.2 KDE Trick

Suppose we assume data from a distribution with density q can be sampled via:

$$z_0 \sim N(0, I_k) \tag{9}$$
$$z = h_B(z_0) \tag{10}$$

where h_B is a function with parameters B. We don't have a form for q, unless we know the map h is invertible and we have the inverse; this won't be the case if, for instance, z has more dimensions than z_0. But we can draw $y_{01}, \dots, y_{0M} \sim_{i.i.d.} N(0, I_k)$, and let

$$q(z) \approx \sum_{m \leq M} K(z, h_B(y_{0m}))$$

We refer to this as the KDE trick.

Given the above definitions, one way to estimate the KL-variational objective:

$$\mathbb{E}_q \log p(x, \theta) - \mathbb{E}_q \log q$$

is to use a generative form for q. Then the reparametrization trick can be used to estimate the first integral, and the KDE trick, along with the reparametrization trick, can be used to estimate the second. As long as the kernel and $\log p(x, \theta)$ are differentiable, we can use gradient ascent to optimize.

6.3 Linear approximation

See Blei & Lafferty (2007) for more details. Suppose we want to compute $\mathbb{E}_q \log b(z)$, where $b(z)$ is such that we can analytically compute $\mathbb{E}_q b(z)$. So the only problem is the log. We use the Taylor/linear approximation of the log, i.e., we use that for any $x, \zeta > 0$,

$$\log x = \log(x/\zeta) + \log \zeta \leq (x/\zeta) - 1 + \log \zeta$$

With this and Jensen, we proceed.

6.4 Laplace and other approximation methods

An alternative to CAVI that can be used when the variable are non-conjugate is the Laplace method, as proposed in Wang & Blei (2013). A Gaussian form is assumed for q. Via a Taylor expansion of $\log p(x, \theta)$, a simple update for q can be performed that only requires calculating a mode and the Hessian of $\log p(x, \theta)$ at θ.

While the Laplace method is useful for handling non-conjugacy of the posterior variables. Moreover, it does not require a mean-field assumption, and if calculating the full Hessian is feasible, covariance among the posterior variables can be captured. However, this method still requires a Gaussian form for q.

Definition 3 (Laplace method)**.** *See Wang & Blei (2013) and my summary of it for more general information. We will assume $q(z)$ is the $N(\lambda, V)$ density, and we will optimize over λ, V. We don't assume V is diagonal. Let $f_\theta(z) := \log p_\theta(x, z)$. Then the evidence lower bound can be written*

$$\mathcal{L}(q) = \mathbb{E}_{q(z)} f(z) - \frac{1}{2} \log |V|^{-1}$$

We can't compute $\mathbb{E}_q f(z)$. *But Taylor expand* $f(z)$ *around* $\hat{z} = \text{argmax}_z f(z)$. *We obtain*

$$\begin{aligned}\mathbb{E}_{q(z)} f(z) &\approx \mathbb{E}_{q(z)} \left(f(\hat{z}) + \frac{1}{2}(z-\hat{z})\nabla^2 f(\hat{z})(z-\hat{z}) \right) \\ &= f(\hat{z}) + \frac{1}{2}\left((\lambda-\hat{z})'\nabla^2 f(\hat{z})(\lambda-\hat{z}) + tr(V\nabla^2 f(\hat{z})) + \log|V| \right)\end{aligned}$$

That is, we were actually able to compute the expectation when we assumed q *was the* $N(\lambda, V)$ *density. Now we can maximize this objective in* λ, V, *which yields the optima of:*

$$\begin{aligned}\hat{\lambda}(x) &= \hat{z}(x) \\ \hat{V}(x) &= -\nabla^2 f(\hat{z}(x))^{-1}\end{aligned}$$

Thus, to update $q(z)$, *we need only find* $\hat{z}$ *(usually by an iterative procedure) and the Hessian of* f.

7 Posterior inference vs. density estimation

Table 6: In this work, we often borrow methods from density estimation and try to apply them to posterior estimation. Here are some issues. H is some operator on densities. Careful with density estimation vs. sampling

Task/thing to calculate	Density estimation	Posterior estimation
Sampling	Have samples from the density	Don't have samples from the po
Form of function	Don't have	Have unnormalized form
$\mathbb{E}_p g$	Have samples from p so fine.	Usually can't just compute. Fo have posterior samples. Gri hard/slow. So do importance tegral, maybe using a variation Problem if it's underdispersed. iterating, but must always be ab for q, we're fine if it is e.g. some express as a deep function of s instance.
$\mathbb{E}_q g$	Need to be able to sample from q	Just need to be able to sample fr or sampling form. Sampling for because only g here.
$\mathbb{E}_q H(p)$	Don't have form of p; can't do	Have unnormalized p, so as lon depend on the normalizer, we're
$\mathbb{E}_p H(q)$	Have samples from p so fine. If want flexible sampling form for q, can use KDE trick.	Same IS tricks for $\mathbb{E}_p$. Can us Problem if iterate and update re portance sampling because KD to underdispersion.

Table 7: Comparison of my favorite optimization methods for posterior inference.

Method and usable objective	Discussion
$KL(\mathbb{Q}\|\|\mathbb{P}) = \mathbb{E}_q \log p(x,z) - \mathbb{E}_q \log q$	Easy iterative algorithm for exponential families Underestimates variance; see via form of optimization. Must pick family for q such that can compute $\mathbb{E}_q \log q$ and can sa $\mathbb{E}_q \log p(x,z)$. Or other tricks, as in Section 6, work on this latter We could though use a neural-net q because can sample and can d
$SM_p(\mathbb{P},\mathbb{Q}) = \mathbb{E}_p\left(\frac{1}{2}\psi_q^2 + \triangle q\right)$	Does not underestimate variance Don't need normalizers because of $\nabla \log$ and because don't nee integral. Exponential family form: allows for nonparametric form and only Fast unless system big. But can parallelize. Requires integral over p. Can do IS, but with what reference dist? If use variational estimate, underdispersed. Can iterate and do IS with previous iteration. . If do exponential family way, can't sample from exponential famil If do neural net way, can sample, but don't have explicit density for family form of solution. Can still use KDE. More complicated bu of potential underdispersion of reference distribution because of n One solution: Gaussian: both an exponential family and we can s
$KL(\mathbb{P}\|\|\mathbb{Q}) = \mathbb{E}_p \log q$	Does not underestimate variance. We could use methods from Tabl of EP?
$SM_q(\mathbb{P},\mathbb{Q})$	None, must play minimax game? Does not underestimate variance we run into the issue of $\mathbb{E}_q g(q)$ here; could solve same way as in th
$JS(\mathbb{P},\mathbb{Q})$	None! Requires full form of p. Must play the minimax game
$TV(\mathbb{P},\mathbb{Q})$	$\downarrow$
$W_1(\mathbb{P},\mathbb{Q})$	$\downarrow$

References

Bishop, Christopher, Lawrence, N., Jaakkola, T., & Jordan, M. I. 1998 (January). Approximating posterior distributions in belief networks using mixtures. *Pages 416–422 of: Advances in Neural Information Processing Systems*, vol. 10.

Blei, David M., & Lafferty, John D. 2007. A correlated topic model of science. *Annals of*

Applied Statistics, **1**(1), 17–35.

Blei, David M., Kucukelbir, Alp, & McAuliffe, Jon D. 2017. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, **112**(518), 859–877.

Gershman, Samuel J., Hoffman, Matthew D., & Blei, David M. 2012. Nonparametric Variational Inference. *In: Proceedings of the 29th International Conference on Machine Learning*.

Ranganath, Rajesh, Tran, Dustin, Altosaar, Jaan, & Blei, David. 2016. Operator Variational Inference. *Pages 496–504 of:* Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., & Garnett, R. (eds), *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc.

Wang, Chong, & Blei, David M. 2013. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, **14**, 1005–1031.