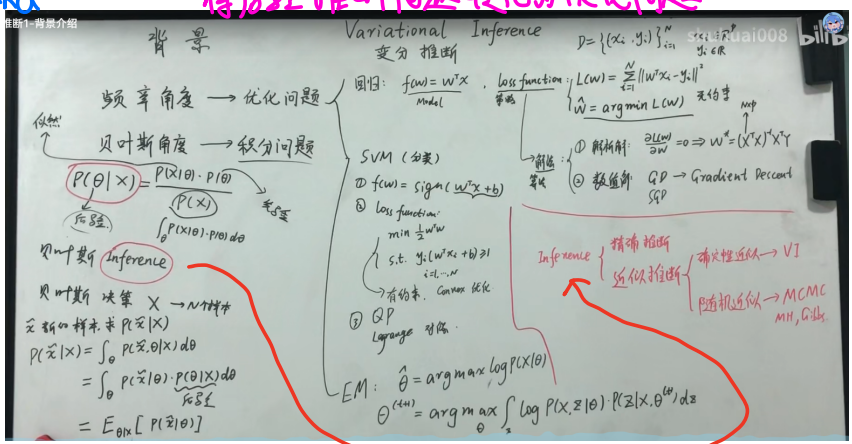
Variational Inference (变分推断)

Background

将后验推断问题转化为优化问题



VI 推导

X: observed data

Z: latent variable + parameter

(X, Z): complete data

$p(\theta|X)$

KL 散度: $-\int p(x) \ln \frac{q(x)}{p(x)} dx \geq 0$

⇒ 分布 $p(x)$ 和 $q(x)$ 之间的 KL 散度

$$\log p(X) = \log p(X, Z) - \log p(Z|X)$$

$$= \log \frac{p(X,Z)}{q(Z)} - \log \frac{p(Z|X)}{q(Z)}$$

对两边求 $q(Z)$ 期望:

左: $\int_Z \log p(X) \cdot q(Z) dZ = \log p(X) \int_Z q(Z) dZ = \log p(X)$

右: $\int_Z \log \frac{p(X,Z)}{q(Z)} \cdot q(Z) - \int_Z \log \frac{p(Z|X)}{q(Z)} q(Z) dZ$

ELBO　　KL 散度

$(\log 1 = 0)$

$$\log p(X) = \mathcal{L}(q) + KL(q||p) \geq 0$$

若 $q(Z) \approx p(Z|X)$ ⇒ KL 散度 = 0

(posterior)

⇓

又∵ $\log p(X)$ 恒定

∴ 即相当于对 $\mathcal{L}(q)$ 求最大值

⇒ $\tilde{q}(Z) = \underset{q(Z)}{\operatorname{argmax}} \mathcal{L}(q)$

Main idea: VI re-frames the computation of an integral (the marginal from exact inference) as the optimization of its lower bound.

KL: not symmetric!

⇒ $KL(p||q) \neq KL(q||p)$

4. 假设 $q(z)$ 可以划分为相互独立的 M 个组，即 $q(z)=\prod_{i=1}^{M} q_i z_i$ (mean field)

$$L(q)=\int_z \log\frac{p(x,z)}{q(z)}\cdot q(z)$$ 先固定其他分量，求 $q_j z_j$

$$=\int_z [\log p(x,z)-\log q(z)]\, q(z)$$

$$=\underbrace{\int_z q(z)\log p(x,z)}_{①}\underbrace{-\int_z q(z)\log q(z)\,dz}_{②}$$

① $\int_z q(z)\log p(x,z)\,dz=\int_z \prod_{i=1}^{M} q_i z_i \log p(x,z)\,dz$

变为上帝视角，先积分 $q_j z_j$，固定剩余 z

$$=\int_{z_j} q_j z_j\left(\int_{z-z_j}\prod_{i\neq j}^{M} q_i z_i \log p(x,z)\,dz_1 z_2\cdots dm\right)dz_j \quad (i\neq j)$$

$$=\int_{z_j} q_j z_j\left(\int_{z-z_j}\log p(x,z)\prod_{i\neq j}^{M} q_i z_i\,dz_i\right)dz_j$$

$$=\boxed{\int_{z_j} q_j z_j\left(E_{\prod_{i\neq j}^{M} q_i z_i}\log p(x,z)\right)dz_j} \quad ①$$

$\log \hat{p}(x,z_j)$

② $\int_z q(z)\log q(z)\,dz=\int_z \prod_{i=1}^{M} q_i(z_i)\sum_{i=1}^{M}\log q_i(z_i)\,dz$

$$=\int_z \prod_{i=1}^{M} q_i(z_i)\,(\log q_1 z_1+\log q_2 z_2+\cdots\log q_m z_m)\,dz \quad ✡$$

看第一项：

$$=\int_z \prod_{i=1}^{M} q_i z_i \log q_1 z_1\,dz$$

设定化技巧

$$=\int_z q_1 z_1 q_2 z_2\cdots q_m z_m \log q_1 z_1\,dz$$

$$=\int_{z_1} q_1 z_1\log q_1 z_1\,dz_1\cdots\underbrace{\int_{z_m} q_m z_m\,dz_m}_{=1}$$

$$= \int_{z_1} q_1(z_1) \log q_1(z_1)\, dz_1$$

$$\therefore \star = \sum_{i=1}^{m} \int_{z_i} q_i(z_i) \log q_i(z_i)\, dz_i$$

$$= \boxed{\int_{z_j} q_j(z_j) \log q_j(z_j)\, dz_j + C} \quad ②$$

$$① - ②: \int_{z_j} q_j(z_j) \log \frac{\hat{p}(x, z_j)}{q_j(z_j)}\, dz_j$$

$$= -KL(q_j \,\|\, \hat{p}(x, z_j)) \le 0$$

即 $q_j(z_j) = \hat{p}(x, z_j)$ 时 $\mathcal{L}(q)$ 取最大

(Coordinate ascent)

Algorithm 1: Coordinate ascent variational inference (CAVI)

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$
Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$
while *the ELBO has not converged* **do**
  for $j \in \{1, \ldots, m\}$ **do**
    Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$
  end
  Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] + \mathbb{E}[\log q(\mathbf{z})]$
end
return $q(\mathbf{z})$

缺点：
1. 假设太强
2. 实际中很多时候无法计算

EM is coordinate ascent on the ELBO

② BBVI

$q_\phi(z)$

ELBO: $E_{q_\phi(z)}\left[\log \frac{p_\theta(x_i^{(i)}, z)}{q_\phi(z)}\right]$

x_i 表示对 i 个样本

$$= E_{q_\phi(z)}\left[\log p_\theta(x_i^{(i)}, z) - \log q_\phi(z)\right]$$

$$= \mathcal{L}(\phi)$$

$$\log p_\theta(x^{(i)}) = \underbrace{ELBO}_{\mathcal{L}(\phi)} + KL(q(\|p) \geq \mathcal{L}(\phi)$$

找一个 $\hat{\phi}$ 使 $L(\phi)$ 最大 → KL 接近 0 → $q \simeq p$

目标

target: $\hat{\phi} = \arg\max_\phi \mathcal{L}(\phi)$

想用梯度下降方法:

△ $\nabla_x \int p(y|x)\, h(x,y)\, dy = \int \nabla_x (p(y|x) h(x,y))\, dy$

$$\nabla_\phi \mathcal{L}(\phi) = \nabla_\phi E_{q_\phi(z)}\left[\log p_\theta(x^{(i)}, z) - \log q_\phi(z)\right]$$

$$= \nabla_\phi \int \underbrace{q_\phi(z)}_{A} \underbrace{\left[\log p_\theta(x^{(i)}, z) - \log q_\phi(z)\right]}_{B} dz$$

类似莱布尼茨: A 导 B 不导

$$= \int \nabla_\phi q_\phi(z) \left[\log p_\theta(x^{(i)}, z) - \log q_\phi(z)\right] dz \quad ① + *$$

A 不导 B 导

$$\int q_\phi(z) \nabla_\phi \left[\log p_\theta(x^{(i)}, z) - \log q_\phi(z)\right] dz \quad ②$$

先看求 2 项:

$$\int q_\phi(z) \underbrace{\nabla_\phi \left[\log p_\theta(x^{(i)}, z) - \log q_\phi(z)\right]}_{0} dz$$

$$= -\int q_\phi(z)\, \nabla_\phi \log q_\phi(z)\, dz$$

$$= -\int \cancel{q_\phi(z)} \frac{1}{\cancel{q_\phi(z)}} \cdot \nabla_\phi q_\phi(z)\, dz$$

$$= -\int \nabla_\phi q_\phi(z)\, dz$$

$$= -\nabla_\phi \int q_\phi(z)\, dz$$

$$= 0$$

$$* \Rightarrow \int \nabla_\phi q_{\phi(z)} \left[\log p_\theta(x^{(i)}, z) - \log q_\phi(z)\right] dz$$

设法写成期望形式：log-derivative trick $\frac{\partial}{\partial x} p(y|x) = p(y|x) \frac{\partial}{\partial x} \log p(y|x)$

$$\int q_{\phi(z)} \nabla_\phi \log q_{\phi(z)} \left[\log p_\theta(x^{(i)}, z) - \log q_\phi(z)\right] dz$$

$$\nabla_\phi L(\phi) = E_{q_{\phi(z)}}\left[\nabla_\phi \log q_{\phi(z)} \left[\log p_\theta(x^{(i)}, z) - \log q_\phi(z)\right] dz\right.$$

这个期望可以用 MC 采样来近似，从而得到梯度。然后用 gd 得到参数 $\theta^{new} = \theta^{old} + \lambda \cdot \nabla_\theta$

→ Rejection Sampling

Importantance sampling

$$z^{(l)} \sim q_\phi(z) \quad l = 1, 2, \cdots L$$

$$\simeq \frac{1}{L} \sum_{l=1}^{L} \nabla_\phi \log q_\phi(z^{(l)}) \left[\log p_\theta(x^{(i)}, z^{(l)}) - \log q_\phi(z^{(l)})\right]$$

有先对数项，导致单样本梯度很大

Reparameterization trick !

考虑：

$$\nabla_\phi L(\phi) = \nabla_\phi \mathbb{E}_{q_\phi}[\log p_\theta(x^i, z) - \log q_\phi(z)] \tag{15}$$

我们取：$z = g_\phi(\varepsilon, x^i), \varepsilon \sim p(\varepsilon)$，于是对后验：$z \sim q_\phi(z|x^i)$，有 $|q_\phi(z|x^i)dz| = |p(\varepsilon)d\varepsilon|$。代入上面的梯度中：

$$\begin{aligned}
\nabla_\phi L(\phi) &= \nabla_\phi \mathbb{E}_{q_\phi}[\log p_\theta(x^i, z) - \log q_\phi(z)] \\
&= \nabla_\phi L(\phi) = \nabla_\phi \int [\log p_\theta(x^i, z) - \log q_\phi(z)] q_\phi dz \\
&= \nabla_\phi \int [\log p_\theta(x^i, z) - \log q_\phi(z)] p_\varepsilon d\varepsilon \\
&= \mathbb{E}_{p(\varepsilon)}[\nabla_\phi [\log p_\theta(x^i, z) - \log q_\phi(z)]] \\
&= \mathbb{E}_{p(\varepsilon)}[\nabla_z [\log p_\theta(x^i, z) - \log q_\phi(z)] \nabla_\phi z] \\
&= \mathbb{E}_{p(\varepsilon)}[\nabla_z [\log p_\theta(x^i, z) - \log q_\phi(z)] \nabla_\phi g_\phi(\varepsilon, x^i)]
\end{aligned} \tag{16}$$

对这个式子进行蒙特卡洛采样，然后计算期望，得到梯度。

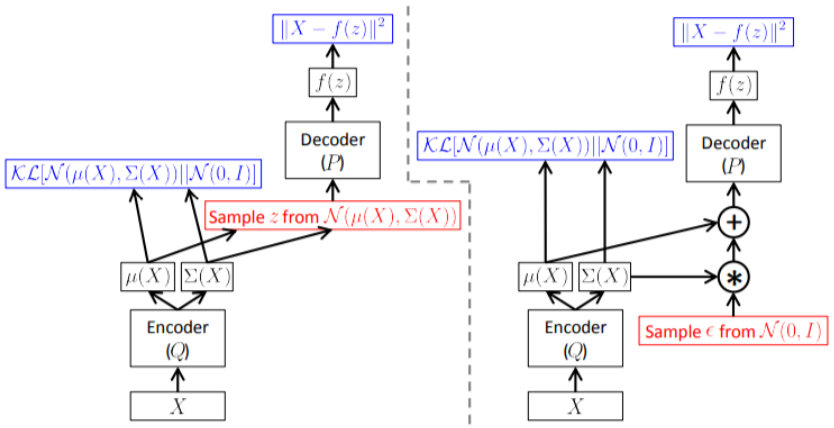


Figure 4: A training-time variational autoencoder implemented as a feed-forward neural network, where $P(X|z)$ is Gaussian. Left is without the "reparameterization trick", and right is with it. Red shows sampling operations that are non-differentiable. Blue shows loss layers. The feedforward behavior of these networks is identical, but backpropagation can be applied only to the right network.

Reparameterization trick

- Consider differentiation of complex expectation

$$\frac{\partial}{\partial x} \int p(y|x)h(x,y)dy$$

- Express y as a deterministic function $g(.)$ of random ϵ and x and perform change-of-variables rule

parameter free

$$\int p(y|x)h(x,y)dy = \int r(\epsilon)h(x, g(\epsilon, x))d\epsilon$$

- Then stochastic differentiation is simply

$$\frac{\partial}{\partial x} \int p(y|x)h(x,y)dy = \frac{\partial}{\partial x} \int r(\epsilon)h(x, g(\epsilon, x))d\epsilon \approx$$
$$\frac{d}{dx}h(x, g(x, \hat{\epsilon})) = \frac{\partial}{\partial x}h(x, g(x, \hat{\epsilon})) + \frac{\partial}{\partial g}h(x, g(x, \hat{\epsilon}))\frac{\partial}{\partial x}g(x, \hat{\epsilon})$$

where $\hat{\epsilon} \sim r(\epsilon)$

Examples of repameterization

$p(x\|y)$	$r(\epsilon)$	$g(\epsilon, y)$
$\mathcal{N}(x\|\mu, \sigma^2)$	$\mathcal{N}(\epsilon\|0, 1)$	$x = \sigma\epsilon + \mu$
$\mathcal{G}(x\|1, \beta)$	$\mathcal{G}(\epsilon\|1, 1)$	$x = \beta\epsilon$
$\mathcal{E}(x\|\lambda)$	$\mathcal{U}(\epsilon\|0, 1)$	$x = -\frac{\log \epsilon}{\lambda}$
$\mathcal{N}(x\|\mu, \Sigma)$	$\mathcal{N}(\epsilon\|0, I)$	$x = A\epsilon + \mu$, where $AA^T = \Sigma$

- Not all continuous distributions can be effectively reparameterized
- Discrete distiributions **cannot** be reparameterized

Reparameterization trick for ELBO

- Return to the first term in our ELBO

$$\int q(Z|X,\phi)\log p(X|Z,\theta)dZ$$

- To get stochastic gradient w.r.t. ϕ first apply mini-batching

$$\frac{\partial}{\partial\phi}\int q(Z|X,\phi)\log p(X|Z,\theta)dZ \approx n\frac{\partial}{\partial\phi}\int q(z_i|x_i,\phi)\log p(x_i|z_i,\theta),$$

- Now perform reparameterization trick $z_i = g(\epsilon, x_i, \phi)$

$$n\frac{\partial}{\partial\phi}\int q(z_i|x_i,\phi)\log p(x_i|z_i,\theta) = n\frac{\partial}{\partial\phi}\int r(\epsilon)\log p(x_i|g(\epsilon,x_i,\phi)z_i,\theta)d\epsilon \approx$$
$$n\frac{\partial}{\partial\phi}\log p(x_i|g(\hat{\epsilon},x_i,\phi)z_i,\theta), \quad \hat{\epsilon}\sim r(\epsilon)$$

VAE: final algorithm

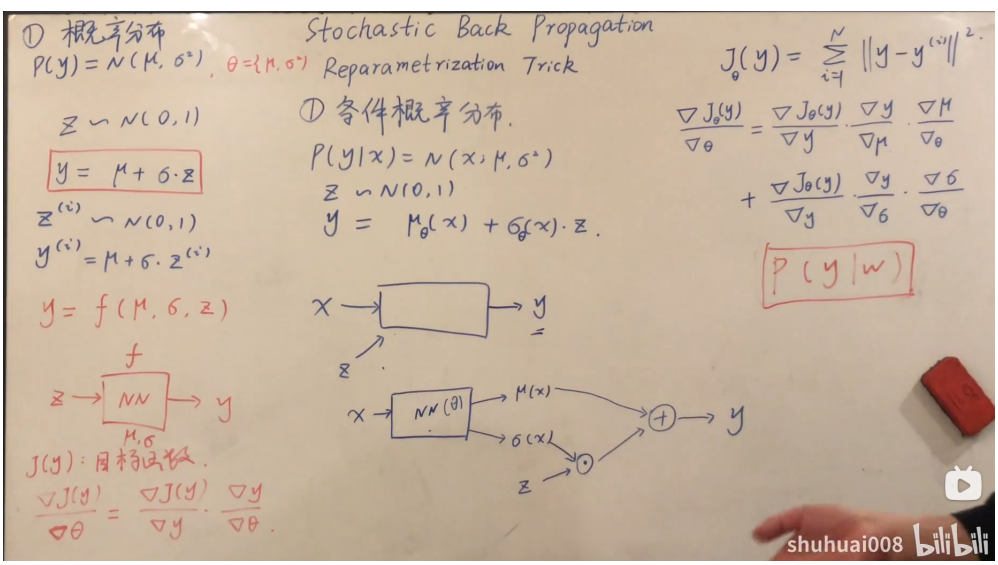
- Input: Training data X, dimension of latent space d
- Pick random $i \sim \mathcal{U}\{1,\ldots,n\}$ and compute stochastic gradients of ELBO w.r.t. θ and ϕ
  - Differentiate w.r.t. θ

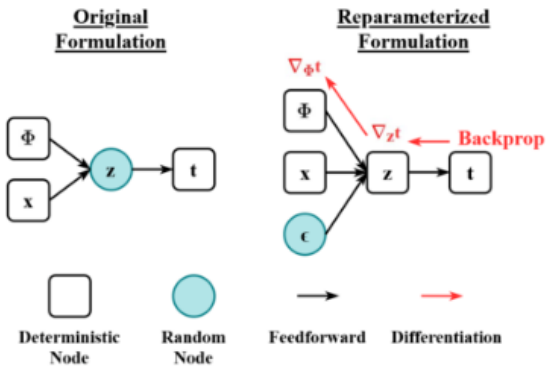
$$\text{stoch.grad}_\theta\mathcal{L}(\phi,\theta) = n\frac{\partial}{\partial\theta}\log p(x_i|z_i^*,\theta),$$

    where $z_i^* \sim q(z_i|x_i,\phi)$
  - Differentiate w.r.t. ϕ

$$\text{stoch.grad}_\phi\mathcal{L}(\phi,\theta) = n\frac{\partial}{\partial\phi}\log p(x_i|g(\hat{\epsilon},x_i,\phi),\theta) - \frac{\partial}{\partial\phi}KL(q(z_i|x_i,\phi)||p(z_i)),$$

    where $\hat{\epsilon}\sim r(\epsilon)$
- Update θ and ϕ according to selected stochastic optimization method





The scheme of the reparameterization trick. The randomness variable ϵ is injected into the latent space $\mathbf{z}$ as external input. In this way, it is possible to backpropagate the gradient without involving stochastic variable during the update.

02

$v = \begin{bmatrix} \mu \\ \sigma \end{bmatrix}$ w/ mean μ, variance σ

Idea: if $q(w|v)$ is Gaussian, consider two ways to sample from it

(1) draw $w \sim \text{Normal}(\mu, \sigma^2)$

(2) draw $\varepsilon \sim N(0,1)$ — standardized r.v.
$w \leftarrow \sigma\varepsilon + \mu$ — plus a transform

Can prove these two procedures produce same distribution.

Key idea: define $w \triangleq t(v, \varepsilon)$
deterministic mapping of parameters v and a std. r.v. ε w/ distr. $p(\varepsilon)$

Now, for any function $f(w,v)$ that depends on w and v,

we can rewrite expectations

$$\mathbb{E}_{q(w|v)}[f(w,v)] = \mathbb{E}_{p(\varepsilon)}[f(t(v,\varepsilon), v)]$$

why? now gradient w.r.t. v can go inside

$$\nabla_v \mathbb{E}_{q(w|v)}[f(w,v)] = \nabla_v \mathbb{E}_{p(\varepsilon)}[f(t(v,\varepsilon),v)] = \mathbb{E}_{p(\varepsilon)}[\nabla_v f(t(v,\varepsilon),v)]$$
chain rule →

03

chain rule ↘

$$\nabla_v \mathbb{E}_{q(w|v)}[f(w,v)] = \mathbb{E}_{p(\varepsilon)}\left[\frac{\partial f(w,v)}{\partial w}\frac{\partial w}{\partial v} + \frac{\partial f(w,v)}{\partial v}\right]$$

Given S samples, can compute as

$$= \frac{1}{S}\sum_s \frac{\partial f(w^s,v)}{\partial w}\bigg|_{w=w^s}\frac{\partial w}{\partial v}\bigg|_{w=w^s} + \frac{\partial f(w^s,v)}{\partial v}$$

Illustration:

What is gradient wrt μ of expected value of normal? Very easy w/ reparameterization (1D)

First way:

$$\nabla_\mu \mathbb{E}_{w\sim N(\mu,\sigma^2)}[w]$$
$$= \nabla_\mu \int p(w|\mu,\sigma^2)\, w\, dw$$
$$= \nabla_\mu \int \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(w-\mu)^2} w\, dw$$

HARD!

Second way:

w can be parameterized as $w = \mu + \sigma\varepsilon$, $\varepsilon \sim N(0,1)$

$$\nabla_\mu \mathbb{E}_{\varepsilon\sim N(0,1)}[\mu + \sigma\varepsilon]$$
$$= \nabla_\mu \mu + \nabla_\mu \underbrace{\mathbb{E}[\sigma\varepsilon]}_{=0} = 1$$
const wrt μ

· exponential family

· 指数族分布可以写成统一的形式：

$$p(x|\eta) = h(x)\exp(\eta^T\phi(x) - A(\eta))$$

$$= \frac{1}{\exp(A(\eta))} h(x)\exp(\eta^T\phi(x))$$

其中，η 是参数向量，$x \in \mathbb{R}^p$

$A(\eta)$：log partition function（可看作归一化因子）

$\phi(x)$：充分统计量（包含样本点的所有信息，有完备性，可理解为"人大代表"）

Explanation:

$p(x|\theta) = \frac{1}{z}\hat{p}(x|\theta)$ → 归一化因子

时左边式子，$\exp(A(\eta))$ 看作 z，$h(x)\exp(\eta^T\phi(x))$ 看作 $\hat{p}(x|\theta)$，则 $A(\eta) = \log(z)$

properties：

$$A'(\eta) = E[\phi(x)]$$

$$A''(\eta) = Var[\phi(x)]$$

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 518, 859–877, Review
https://doi.org/10.1080/01621459.2017.1285773

Variational Inference: A Review for Statisticians

David M. Blei[a], Alp Kucukelbir[b], and Jon D. McAuliffe[c]

[a]Department of Computer Science and Statistics, Columbia University, New York, NY; [b]Department of Computer Science, Columbia University, New York, NY; [c]Department of Statistics, University of California, Berkeley, CA

ABSTRACT

One of the core problems of modern statistics is to approximate difficult-to-compute probability densities. This problem is especially important in Bayesian statistics, which frames all inference about unknown quantities as a calculation involving the posterior density. In this article, we review variational inference (VI), a method from machine learning that approximates probability densities through optimization. VI has been used in many applications and tends to be faster than classical methods, such as Markov chain Monte Carlo sampling. The idea behind VI is to first posit a family of densities and then to find a member of that family which is close to the target density. Closeness is measured by Kullback–Leibler divergence. We review the ideas behind mean-field variational inference, discuss the special case of VI applied to exponential family models, present a full example with a Bayesian mixture of Gaussians, and derive a variant that uses stochastic optimization to scale up to massive data. We discuss modern research in VI and highlight important open problems. VI is powerful, but it is not yet well understood. Our hope in writing this article is to catalyze statistical research on this class of algorithms. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2016
Revised December 2016

KEYWORDS
Algorithms; Computationally intensive methods; Statistical computing

1. Introduction

One of the core problems of modern statistics is to approximate difficult-to-compute probability densities. This problem is especially important in Bayesian statistics, which frames all inference about unknown quantities as a calculation about the posterior. Modern Bayesian statistics relies on models for which the posterior is not easy to compute and corresponding algorithms for approximating them.

In this article, we review variational inference (VI), a method from machine learning for approximating probability densities (Jordan et al. 1999; Wainwright and Jordan 2008). Variational inference is widely used to approximate posterior densities for Bayesian models, an alternative strategy to Markov chain Monte Carlo (MCMC) sampling. Compared to MCMC, variational inference tends to be faster and easier to scale to large data—it has been applied to problems such as large-scale document analysis, computational neuroscience, and computer vision. But variational inference has been studied less rigorously than MCMC, and its statistical properties are less well understood. In writing this article, our hope is to catalyze statistical research on variational inference.

First, we set up the general problem. Consider a joint density of latent variables $\mathbf{z} = z_{1:m}$ and observations $\mathbf{x} = x_{1:n}$,

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z})p(\mathbf{x} \mid \mathbf{z}).$$

In Bayesian models, the latent variables help govern the distribution of the data. A Bayesian model draws the latent variables from a prior density $p(\mathbf{z})$ and then relates them to the observations through the likelihood $p(\mathbf{x} \mid \mathbf{z})$. Inference in a Bayesian model amounts to conditioning on data and computing the posterior $p(\mathbf{z} \mid \mathbf{x})$. In complex Bayesian models, this computation often requires approximate inference.

For decades, the dominant paradigm for approximate inference has been MCMC (Hastings 1970; Gelfand and Smith 1990). In MCMC, we first construct an ergodic Markov chain on $\mathbf{z}$ whose stationary distribution is the posterior $p(\mathbf{z} \mid \mathbf{x})$. Then, we sample from the chain to collect samples from the stationary distribution. Finally, we approximate the posterior with an empirical estimate constructed from (a subset of) the collected samples.

MCMC sampling has evolved into an indispensable tool to the modern Bayesian statistician. Landmark developments include the Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970), the Gibbs sampler (Geman and Geman 1984), and its application to Bayesian statistics (Gelfand and Smith 1990). MCMC algorithms are under active investigation. They have been widely studied, extended, and applied; see Robert and Casella (2004) for a perspective.

However, there are problems for which we cannot easily use this approach. These arise particularly when we need an approximate conditional faster than a simple MCMC algorithm can produce, such as when datasets are large or models are very complex. In these settings, variational inference provides a good alternative approach to Bayesian inference.

Rather than use sampling, the main idea behind variational inference is to use optimization. First, we posit a *family* of approximate densities $\mathcal{Q}$. This is a set of densities over the latent variables. Then, we try to find the member of that family that

CONTACT Alp Kucukelbir alp@cs.columbia.edu Department of Computer Science, Columbia University, New York, NY 10027.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

MCMC → sample from Markov chain >> posterior
VI → optimization

minimizes the Kullback-Leibler (KL) divergence to the exact posterior,

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z})\in\mathcal{Q}} \text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z}\,|\,\mathbf{x})\right). \quad (1)$$

Finally, we approximate the posterior with the optimized member of the family $q^*(\cdot)$.

Variational inference thus turns the inference problem into an optimization problem, and the reach of the family $\mathcal{Q}$ manages the complexity of this optimization. One of the key ideas behind variational inference is to choose $\mathcal{Q}$ to be flexible enough to capture a density close to $p(\mathbf{z}\,|\,\mathbf{x})$, but simple enough for efficient optimization.[1]

We emphasize that MCMC and variational inference are different approaches to solving the same problem. MCMC algorithms sample a Markov chain; variational algorithms solve an optimization problem. MCMC algorithms approximate the posterior with samples from the chain; variational algorithms approximate the posterior with the result of the optimization.

Comparing variational inference and MCMC. When should a statistician use MCMC and when should she use variational inference? We will offer some guidance. MCMC methods tend to be more computationally intensive than variational inference but they also provide guarantees of producing (asymptotically) exact samples from the target density (Robert and Casella 2004). Variational inference does not enjoy such guarantees—it can only find a density close to the target—but tends to be faster than MCMC. Because it rests on optimization, variational inference easily takes advantage of methods like stochastic optimization (Robbins and Monro 1951; Kushner and Yin 1997) and distributed optimization. Some MCMC methods can also exploit these innovations (Welling and Teh 2011; Ahmed et al. 2012).

Thus, variational inference is suited to large datasets and scenarios where we want to quickly explore many models; MCMC is suited to smaller datasets and scenarios where we happily pay a heavier computational cost for more precise samples. For example, we might use MCMC in a setting where we spent 20 years collecting a small but expensive dataset, where we are confident that our model is appropriate, and where we require precise inferences. We might use variational inference when fitting a probabilistic model of text to one billion text documents and where the inferences will be used to serve search results to a large population of users. In this scenario, we can use distributed computation and stochastic optimization to scale and speed up inference, and we can easily explore many different models of the data.

Dataset size is not the only consideration. Another factor is the geometry of the posterior distribution. For example, the posterior of a mixture model has multiple modes, each corresponding to a label permutation of the components. Gibbs sampling, if the model permits, is a powerful approach to sampling from such target distributions; it quickly focuses on one of the modes. For mixture models where Gibbs sampling is not an option, variational inference may perform better than a more general MCMC technique (e.g., Hamiltonian Monte Carlo), even for small datasets (Kucukelbir et al. 2015). Exploring the interplay between model complexity and inference (and between variational inference and MCMC) is an exciting avenue for future research (see Section 5.4).

The relative accuracy of variational inference and MCMC is still unknown. We do know that variational inference generally underestimates the variance of the posterior density; this is a consequence of its objective function. But, depending on the task at hand, underestimating the variance may be acceptable. Several lines of empirical research have shown that variational inference does not necessarily suffer in accuracy, for example, in terms of posterior predictive densities (Blei and Jordan 2006; Braun and McAuliffe 2010; Kucukelbir et al. 2017); other research focuses on where variational inference falls short, especially around the posterior variance, and tries to more closely match the inferences made by MCMC (Giordano, Broderick, and Jordan 2015). In general, a statistical theory and understanding around variational inference is an important open area of research (see Section 5.2). We can envision future results that outline which classes of models are particularly suited to each algorithm and perhaps even theory that bounds their accuracy. More broadly, variational inference is a valuable tool, alongside MCMC, in the statistician's toolbox.

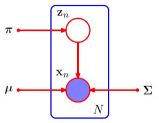
It might appear to the reader that variational inference is only relevant to Bayesian analysis. Indeed, both variational inference and MCMC have had a significant impact on applied Bayesian computation and we will be focusing on latent variable Bayesian models here. We emphasize, however, that these techniques also apply more generally to computation about intractable densities. MCMC is a tool for simulating from densities and variational inference is a tool for approximating densities. One need not be a Bayesian to have use for variational inference.

Research on variational inference. The development of variational techniques for Bayesian inference followed two parallel, yet separate, tracks. Peterson and Anderson (1987) is arguably the first variational procedure for a particular model: a neural network. This article, along with insights from statistical mechanics (Parisi 1988), led to a flurry of variational inference procedures for a wide class of models (Saul, Jaakkola, and Jordan 1996; Jaakkola and Jordan 1996, 1997; Ghahramani and Jordan 1997; Jordan et al. 1999). In parallel, Hinton and Van Camp (1993) proposed a variational algorithm for a similar neural network model. Neal and Hinton (1998, first published in 1993) made important connections to the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), which then led to a variety of variational inference algorithms for other types of models (Waterhouse, MacKay, and Robinson 1996; MacKay 1997; Barber and Bishop 1998).

Modern research on variational inference focuses on several aspects: tackling Bayesian inference problems that involve massive data; using improved optimization methods for solving Equation (1) (which is usually subject to local minima); developing generic variational inference algorithms that are easy to apply to a wide class of models; and increasing the accuracy of variational inference, for example, by stretching the boundaries of $\mathcal{Q}$ while managing complexity in optimization.

Organization of this article. Section 2 describes the basic ideas behind the simplest approach to variational inference: mean-field inference and coordinate-ascent optimization. Section 3

[1] We focus here on KL$(q\|p)$-based optimization, also called Kullback–Leibler variational inference (Barber 2012). Wainwright and Jordan (2008) emphasized that any procedure that uses optimization to approximate a density can be termed "variational inference." This includes methods like expectation propagation (Minka 2001), belief propagation (Yedidia, Freeman, and Weiss 2001), or even the Laplace approximation. We briefly discuss alternative divergence measures in Section 5.



works out the details for a Bayesian mixture of Gaussians, an example model familiar to many readers. Sections 4.1 and 4.2 describe variational inference for the class of models where the joint density of the latent and observed variables is in the exponential family—this includes many intractable models from modern Bayesian statistics and reveals deep connections between variational inference and the Gibbs sampler by Gelfand and Smith (1990). Section 4.3 expands on this algorithm to describe stochastic variational inference (Hoffman et al. 2013), which scales variational inference to massive data using stochastic optimization (Robbins and Monro 1951). Finally, with these foundations in place, Section 5 gives a perspective on the field—applications in the research literature, a survey of theoretical results, and an overview of some open problems.

2. Variational Inference

The goal of variational inference is to approximate a conditional density of latent variables given observed variables. The key idea is to solve this problem with optimization. We use a family of densities over the latent variables, parameterized by free "variational parameters." The optimization finds the member of this family, that is, the setting of the parameters, which is closest in KL divergence to the conditional of interest. The fitted variational density then serves as a proxy for the exact conditional density. (All vectors defined below are column vectors, unless stated otherwise.)

2.1. The Problem of Approximate Inference

Let $\mathbf{x} = x_{1:n}$ be a set of observed variables and $\mathbf{z} = z_{1:m}$ be a set of latent variables, with joint density $p(\mathbf{z}, \mathbf{x})$. We omit constants, such as hyperparameters, from the notation.

The inference problem is to compute the conditional density of the latent variables given the observations, $p(\mathbf{z} \mid \mathbf{x})$. This conditional can be used to produce point or interval estimates of the latent variables, form predictive densities of new data, and more.

We can write the conditional density as

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}. \tag{2}$$

The denominator contains the marginal density of the observations, also called the *evidence*. We calculate it by marginalizing out the latent variables from the joint density,

$$p(\mathbf{x}) = \int p(\mathbf{z}, \mathbf{x})\, d\mathbf{z}. \tag{3}$$

For many models, this evidence integral is unavailable in closed form or requires exponential time to compute. The evidence is what we need to compute the conditional from the joint; this is why inference in such models is hard.

Note we assume that all unknown quantities of interest are represented as latent random variables. This includes parameters that might govern all the data, as found in Bayesian models, and latent variables that are "local" to individual data points.

Bayesian mixture of Gaussians. Consider a Bayesian mixture of unit-variance univariate Gaussians. There are K mixture components, corresponding to K Gaussian distributions with means $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$. The mean parameters are drawn independently from a common prior $p(\mu_k)$, which we assume to be a Gaussian $\mathcal{N}(0, \sigma^2)$; the prior variance σ^2 is a hyperparameter. To generate an observation x_i from the model, we first choose a cluster assignment c_i. It indicates which latent cluster x_i comes from and is drawn from a categorical distribution over $\{1, \ldots, K\}$. (We encode c_i as an indicator K-vector, all zeros except for a one in the position corresponding to x_i's cluster.) We then draw x_i from the corresponding Gaussian $\mathcal{N}(c_i^\top \boldsymbol{\mu}, 1)$.

The full hierarchical model is

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2), && k = 1, \ldots, K, && (4)\\ c_i &\sim \text{categorical}(1/K, \ldots, 1/K), && i = 1, \ldots, n, && (5)\\ x_i \mid c_i, \boldsymbol{\mu} &\sim \mathcal{N}\left(c_i^\top \boldsymbol{\mu}, 1\right) && i = 1, \ldots, n. && (6) \end{aligned}$$

For a sample of size n, the joint density of latent and observed variables is

$$p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(x_i \mid c_i, \boldsymbol{\mu}). \tag{7}$$

The latent variables are $\mathbf{z} = \{\boldsymbol{\mu}, \mathbf{c}\}$, the K class means and n class assignments.

Here, the evidence is

$$p(\mathbf{x}) = \int p(\boldsymbol{\mu}) \prod_{i=1}^{n} \sum_{c_i} p(c_i) p(x_i \mid c_i, \boldsymbol{\mu})\, d\boldsymbol{\mu}. \tag{8}$$

The integrand in Equation (8) does not contain a separate factor for each μ_k. (Indeed, each μ_k appears in all n factors of the integrand.) Thus, the integral in Equation (8) does not reduce to a product of one-dimensional integrals over the μ_k's. The time complexity of numerically evaluating the K-dimensional integral is $\mathcal{O}(K^n)$.

If we distribute the product over the sum in (8) and rearrange, we can write the evidence as a sum over all possible configurations $\mathbf{c}$ of cluster assignments,

$$p(\mathbf{x}) = \sum_{\mathbf{c}} p(\mathbf{c}) \int p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(x_i \mid c_i, \boldsymbol{\mu})\, d\boldsymbol{\mu}. \tag{9}$$

Here, each individual integral is computable, thanks to the conjugacy between the Gaussian prior on the components and the Gaussian likelihood. But there are K^n of them, one for each configuration of the cluster assignments. Computing the evidence remains exponential in K, hence intractable.

2.2. The Evidence Lower Bound

In variational inference, we specify a family $\mathcal{Q}$ of densities over the latent variables. Each $q(\mathbf{z}) \in \mathcal{Q}$ is a candidate approximation to the exact conditional. Our goal is to find the best candidate, the one closest in KL divergence to the exact conditional.[2] Inference now amounts to solving the following optimization problem,

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z}) \in \mathcal{Q}} \text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})\right). \tag{10}$$

[2] The KL divergence is an information-theoretical measure of proximity between two densities. It is asymmetric—that is, $\text{KL}(q\|p) \neq \text{KL}(p\|q)$—and nonnegative. It is minimized when $q(\cdot) = p(\cdot)$.

Once found, $q^*(\cdot)$ is the best approximation of the conditional, within the family $\mathcal{Q}$. The complexity of the family determines the complexity of this optimization.

However, this objective is not computable because it requires computing the logarithm of the evidence, $\log p(\mathbf{x})$ in Equation (3). (That the evidence is hard to compute is why we appeal to approximate inference in the first place.) To see why, recall that KL divergence is

$$\text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})\right) = \mathbb{E}\left[\log q(\mathbf{z})\right] - \mathbb{E}\left[\log p(\mathbf{z} \mid \mathbf{x})\right], \quad (11)$$

where all expectations are taken with respect to $q(\mathbf{z})$. Expanding the conditional,

$$\text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})\right) = \mathbb{E}[\log q(\mathbf{z})] - \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] + \log p(\mathbf{x}). \quad (12)$$

This reveals its dependence on $\log p(\mathbf{x})$.

Because we cannot compute the KL, we optimize an alternative objective that is equivalent to the KL up to an added constant,

$$\text{ELBO}(q) = \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] - \mathbb{E}\left[\log q(\mathbf{z})\right]. \quad (13)$$

This function is called the evidence lower bound (ELBO) (for reasons explained in the text following). The ELBO is the negative KL divergence of Equation (12) plus $\log p(\mathbf{x})$, which is a constant with respect to $q(\mathbf{z})$. Maximizing the ELBO is equivalent to minimizing the KL divergence.

Examining the ELBO gives intuitions about the optimal variational density. We rewrite the ELBO as a sum of the expected log-likelihood of the data and the KL divergence between the prior $p(\mathbf{z})$ and $q(\mathbf{z})$,

$$\begin{aligned}\text{ELBO}(q) &= \mathbb{E}\left[\log p(\mathbf{z})\right] + \mathbb{E}\left[\log p(\mathbf{x} \mid \mathbf{z})\right] - \mathbb{E}\left[\log q(\mathbf{z})\right] \\ &= \mathbb{E}\left[\log p(\mathbf{x} \mid \mathbf{z})\right] - \text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z})\right).\end{aligned}$$

Which values of $\mathbf{z}$ will this objective encourage $q(\mathbf{z})$ to place its mass on? The first term is an expected likelihood; it encourages densities that place their mass on configurations of the latent variables that explain the observed data. The second term is the negative divergence between the variational density and the prior; it encourages densities close to the prior. Thus, the variational objective mirrors the usual balance between likelihood and prior.

Another property of the ELBO is that it lower-bounds the (log) evidence, $\log p(\mathbf{x}) \geq \text{ELBO}(q)$ for any $q(\mathbf{z})$. This explains the name. To see this notice that Equations (12) and (13) give the following expression of the evidence,

$$\log p(\mathbf{x}) = \text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})\right) + \text{ELBO}(q). \quad (14)$$

The bound then follows from the fact that $\text{KL}(\cdot) \geq 0$ (Kullback and Leibler 1951). In the original literature on variational inference, this was derived through Jensen's inequality (Jordan et al. 1999).

The relationship between the ELBO and $\log p(\mathbf{x})$ has led to using the variational bound as a model selection criterion. This has been explored for mixture models (Ueda and Ghahramani 2002; McGrory and Titterington 2007) and more generally (Beal and Ghahramani 2003). The premise is that the bound is a good approximation of the marginal likelihood, which provides a basis for selecting a model. Though this sometimes works in practice, selecting based on a bound is not justified in theory. Other research has used variational approximations in the log predictive density to use VI in cross-validation-based model selection (Nott et al. 2012).

Finally, many readers will notice that the first term of the ELBO in Equation (13) is the expected complete log-likelihood, which is optimized by the EM algorithm (Dempster, Laird, and Rubin 1977). The EM algorithm was designed for finding maximum likelihood estimates in models with latent variables. It uses the fact that the ELBO is equal to the log-marginal-likelihood $\log p(\mathbf{x})$ (i.e., the log evidence) when $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x})$. EM alternates between computing the expected complete log-likelihood according to $p(\mathbf{z} \mid \mathbf{x})$ (the E step) and optimizing it with respect to the model parameters (the M step). Unlike variational inference, EM assumes the expectation under $p(\mathbf{z} \mid \mathbf{x})$ is computable and uses it in otherwise difficult parameter estimation problems. Unlike EM, variational inference does not estimate fixed model parameters—it is often used in a Bayesian setting where classical parameters are treated as latent variables. Variational inference applies to models where we cannot compute the exact conditional of the latent variables.[3]

2.3. The Mean-Field Variational Family

We described the ELBO, the variational objective function in the optimization of Equation (10). We now describe a variational family $\mathcal{Q}$, to complete the specification of the optimization problem. The complexity of the family determines the complexity of the optimization; it is more difficult to optimize over a complex family than a simple family.

In this review, we focus on the *mean-field variational family*, where the latent variables are mutually independent and each governed by a distinct factor in the variational density. A generic member of the mean-field variational family is

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j). \quad (15)$$

Each latent variable z_j is governed by its own variational factor, the density $q_j(z_j)$. In optimization, these variational factors are chosen to maximize the ELBO of Equation (13).

We emphasize that the variational family is not a model of the observed data—indeed, the data $\mathbf{x}$ does not appear in Equation (15). Instead, it is the ELBO, and the corresponding KL minimization problem, which connects the fitted variational density to the data and model.

Notice we have not specified the parametric form of the individual variational factors. In principle, each can take on any parametric form appropriate to the corresponding random variable. For example, a continuous variable might have a Gaussian factor; a categorical variable will typically have a categorical factor. We will see in Sections 4, 4.1, and 4.2 that there are many

[3] Two notes: (a) Variational EM is the EM algorithm with a variational E-step, that is, a computation of an approximate conditional. (b) The coordinate ascent algorithm of Section 2.4 resembles the EM algorithm. The "E step" computes approximate conditionals of local latent variables; the "M step" computes a conditional of the global latent variables.

models for which properties of the model determine optimal forms of the mean-field variational factors $q_j(z_j)$.

Finally, though we focus on mean-field inference in this review, researchers have also studied more complex families. One way to expand the family is to add dependencies between the variables (Saul and Jordan 1996; Barber and Wiegerinck 1999); this is called structured variational inference. Another way to expand the family is to consider mixtures of variational densities, that is, additional latent variables within the variational family (Bishop et al. 1998). Both of these methods potentially improve the fidelity of the approximation, but there is a trade off. Structured and mixture-based variational families come with a more difficult-to-solve variational optimization problem.

Bayesian mixture of Gaussians (continued). Consider again the Bayesian mixture of Gaussians. The mean-field variational family contains approximate posterior densities of the form

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^{K} q\big(\mu_k; m_k, s_k^2\big) \prod_{i=1}^{n} q(c_i; \varphi_i). \tag{16}$$

Following the mean-field recipe, each latent variable is governed by its own variational factor. The factor $q(\mu_k; m_k, s_k^2)$ is a Gaussian distribution on the kth mixture component's mean parameter; its mean is m_k and its variance is s_k^2. The factor $q(c_i; \varphi_i)$ is a distribution on the ith observation's mixture assignment; its assignment probabilities are a K-vector φ_i.

Here, we have asserted parametric forms for these factors: the mixture components are Gaussian with variational parameters (mean and variance) specific to the kth cluster; the cluster assignments are categorical with variational parameters (cluster probabilities) specific to the ith data point. In fact, these are the optimal forms of the mean-field variational density for the mixture of Gaussians.

With the variational family in place, we have completely specified the variational inference problem for the mixture of Gaussians. The ELBO is defined by the model definition in Equation (7) and the mean-field family in Equation (16). The corresponding variational optimization problem maximizes the ELBO with respect to the variational parameters, that is, the Gaussian parameters for each mixture component and the categorical parameters for each cluster assignment. We will see this example through in Section 3.

Visualizing the mean-field approximation. The mean-field family is expressive because it can capture any marginal density of the latent variables. However, it cannot capture correlation between them. Seeing this in action reveals some of the intuitions and limitations of mean-field variational inference.

Consider a two-dimensional Gaussian distribution, shown in violet in Figure 1. This density is highly correlated, which defines its elongated shape.

The optimal mean-field variational approximation to this posterior is a product of two Gaussian distributions. Figure 1 shows the mean-field variational density after maximizing the ELBO. While the variational approximation has the same mean as the original density, its covariance structure is, by construction, decoupled.

Further, the marginal variances of the approximation underrepresent those of the target density. This is a common effect

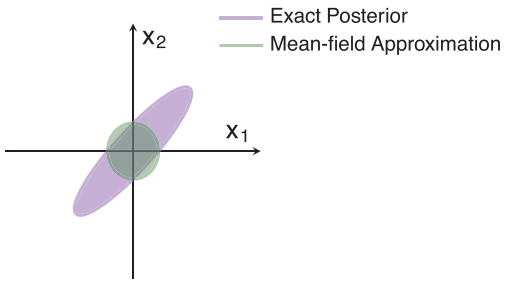


Figure 1. Visualizing the mean-field approximation to a two-dimensional Gaussian posterior. The ellipses show the effect of mean-field factorization. (The ellipses are 2σ contours of the Gaussian distributions.)

in mean-field variational inference and, with this example, we can see why. The KL divergence from the approximation to the posterior is in Equation (11). It penalizes placing mass in $q(\cdot)$ on areas where $p(\cdot)$ has little mass, but penalizes less the reverse. In this example, to successfully match the marginal variances, the circular $q(\cdot)$ would have to expand into territory where $p(\cdot)$ has little mass.

2.4. Coordinate Ascent Mean-Field Variational Inference

Using the ELBO and the mean-field family, we have cast approximate conditional inference as an optimization problem. In this section, we describe one of the most commonly used algorithms for solving this optimization problem, coordinate ascent variational inference (CAVI) (Bishop 2006). CAVI iteratively optimizes each factor of the mean-field variational density, while holding the others fixed. It climbs the ELBO to a local optimum.

The algorithm. We first state a result. Consider the jth latent variable z_j. The *complete conditional* of z_j is its conditional density given all of the other latent variables in the model and the observations, $p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})$. Fix the other variational factors $q_\ell(z_\ell)$, $\ell \neq j$. The optimal $q_j(z_j)$ is then proportional to the exponentiated expected log of the complete conditional,

$$q_j^*(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}. \tag{17}$$

The expectation in Equation (17) is with respect to the (currently fixed) variational density over $\mathbf{z}_{-j}$, that is, $\prod_{\ell \neq j} q_\ell(z_\ell)$. Equivalently, Equation (17) is proportional to the exponentiated log of the joint,

$$q_j^*(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})]\}. \tag{18}$$

Because of the mean-field property—all the latent variables are independent—the expectations on the right-hand side do not involve the jth variational factor. Thus, this is a valid coordinate update.

These equations underlie the CAVI algorithm, presented as Algorithm 1. We maintain a set of variational factors $q_\ell(z_\ell)$. We iterate through them, updating $q_j(z_j)$ using Equation (18). CAVI goes uphill on the ELBO of Equation (13), eventually finding a local optimum. As examples we show CAVI for a mixture of Gaussians in Section 3 and for a nonconjugate linear regression in Appendix A (in the online supplementary materials).

CAVI can also be seen as a "message passing" algorithm (Winn and Bishop 2005), iteratively updating each random variable's variational parameters based on the variational parameters of the variables in its Markov blanket. This perspective

enabled the design of automated software for a large class of models (Wand et al. 2011; Minka et al. 2014). Variational message passing connects variational inference to the classical theories of graphical models and probabilistic inference (Pearl 1988; Lauritzen and Spiegelhalter 1988). It has been extended to nonconjugate models (Knowles and Minka 2011) and generalized via factor graphs (Minka 2005).

Algorithm 1: Coordinate ascent variational inference (CAVI)

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$
Output: A variational density $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$
Initialize: Variational factors $q_j(z_j)$
while *the ELBO has not converged* **do**
  for $j \in \{1, \ldots, m\}$ **do**
    Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$
  end
  Compute $\text{ELBO}(q) = \mathbb{E}[\log p(\mathbf{z}, \mathbf{x})] + \mathbb{E}[\log q(\mathbf{z})]$
end
return $q(\mathbf{z})$

Finally, CAVI is closely related to Gibbs sampling (Geman and Geman 1984; Gelfand and Smith 1990), the classical workhorse of approximate inference. The Gibbs sampler maintains a realization of the latent variables and iteratively samples from each variable's complete conditional. Equation (18) uses the same complete conditional. It takes the expected log, and uses this quantity to iteratively set each variable's variational factor.[4]

Derivation. We now derive the coordinate update in Equation (18). The idea appears in Bishop (2006), but the argument there uses gradients, which we do not. Rewrite the ELBO of Equation (13) as a function of the jth variational factor $q_j(z_j)$, absorbing into a constant the terms that do not depend on it,

$$\text{ELBO}(q_j) = \mathbb{E}_j[\mathbb{E}_{-j}[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})]] - \mathbb{E}_j[\log q_j(z_j)] + \text{const.} \quad (19)$$

We have rewritten the first term of the ELBO using iterated expectation. The second term we have decomposed, using the independence of the variables (i.e., the mean-field assumption) and retaining only the term that depends on $q_j(z_j)$.

Up to an added constant, the objective function in Equation (19) is equal to the negative KL divergence between $q_j(z_j)$ and $q_j^*(z_j)$ from Equation (18). Thus, we maximize the ELBO with respect to q_j when we set $q_j(z_j) = q_j^*(z_j)$.

2.5. Practicalities

Here, we highlight a few things to keep in mind when implementing and using variational inference in practice.

Initialization. The ELBO is (generally) a nonconvex objective function. CAVI only guarantees convergence to a local optimum, which can be sensitive to initialization. Figure 2 shows the ELBO trajectory for 10 random initializations using the

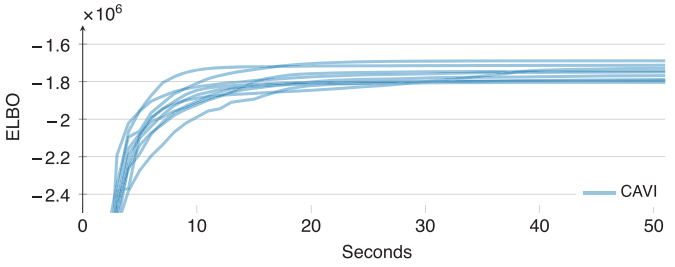


Figure 2. Different initializations may lead CAVI to find different local optima of the ELBO.

Gaussian mixture model. The means of the variational factors were randomly initialized by drawing from a factorized Gaussian calibrated to the empirical mean and variance of the dataset. (This inference is on images; see Section 3.4.) Each initialization reaches a different value, indicating the presence of many local optima in the ELBO. In terms of $\text{KL}(q||p)$, better local optima give variational densities that are closer to the exact posterior.

This is not always a disadvantage. Some models, such as the mixture of Gaussians (Section 3 and Appendix B, in the online supplementary materials) and mixed-membership model (Appendix C, in the online supplementary materials), exhibit many posterior modes due to label switching: swapping cluster assignment labels induces many symmetric posterior modes. Representing one of these modes is sufficient for exploring latent clusters or predicting new observations.

Assessing convergence. Monitoring the ELBO in CAVI is simple; we typically declare convergence once the change in ELBO falls below some small threshold. However, computing the ELBO of the full dataset may be undesirable. Instead, we suggest computing the average log predictive of a small held-out dataset. Monitoring changes here is a proxy to monitoring the ELBO of the full data. (Unlike the full ELBO, held-out predictive probability is not guaranteed to monotonically increase across iterations of CAVI.)

Numerical stability. Probabilities are constrained to live within $[0, 1]$. Precisely manipulating and performing arithmetic on small numbers requires additional care. When possible, we recommend working with logarithms of probabilities. One useful identity is the "log-sum-exp" trick,

$$\log\left[\sum_i \exp(x_i)\right] = \alpha + \log\left[\sum_i \exp(x_i - \alpha)\right]. \quad (20)$$

The constant α is typically set to $\max_i x_i$. This provides numerical stability to common computations in variational inference procedures.

3. A Complete Example: Bayesian Mixture of Gaussians

As an example, we return to the simple mixture of Gaussians model of Section 2.1. To review, consider K mixture components and n real-valued data points $x_{1:n}$. The latent variables are K real-valued mean parameters $\boldsymbol{\mu} = \mu_{1:K}$ and n latent-class assignments $\mathbf{c} = c_{1:n}$. The assignment c_i indicates which latent cluster x_i comes from. In detail, c_i is an indicator K-vector, all zeros except for a one in the position corresponding to x_i's cluster. There is a fixed hyperparameter σ^2, the variance

[4] Many readers will know that we can significantly speed up the Gibbs sampler by marginalizing out some of the latent variables; this is called collapsed Gibbs sampling. We can speed up variational inference with similar reasoning; this is called collapsed variational inference. It has been developed for the same class of models described here (Sung, Ghahramani, and Bang 2008; Hensman, Rattray, and Lawrence 2012). These ideas are outside the scope of our review.

of the normal prior on the μ_k's. We assume the observation variance is one and take a uniform prior over the mixture components.

The joint density of the latent and observed variables is in Equation (7). The variational family is in Equation (16). Recall that there are two types of variational parameters—categorical parameters φ_i for approximating the posterior cluster assignment of the ith data point and Gaussian parameters m_k and s_k^2 for approximating the posterior of the kth mixture component.

We combine the joint and the mean-field family to form the ELBO for the mixture of Gaussians. It is a function of the variational parameters $\mathbf{m}$, $\mathbf{s}^2$, and $\boldsymbol{\varphi}$,

$$\begin{aligned}\text{ELBO}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi}) &= \sum_{k=1}^{K} \mathbb{E}\left[\log p(\mu_k); m_k, s_k^2\right] \\ &+ \sum_{i=1}^{n}\left(\mathbb{E}\left[\log p(c_i); \varphi_i\right] + \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}, \mathbf{s}^2\right]\right) \\ &- \sum_{i=1}^{n} \mathbb{E}\left[\log q\big(c_i; \varphi_i\big)\right] - \sum_{k=1}^{K} \mathbb{E}\left[\log q(\mu_k; m_k, s_k^2)\right]. \quad (21)\end{aligned}$$

In each term, we have made explicit the dependence on the variational parameters. Each expectation can be computed in closed form.

The CAVI algorithm updates each variational parameter in turn. We first derive the update for the variational cluster assignment factor; we then derive the update for the variational mixture component factor.

3.1. The Variational Density of the Mixture Assignments

We first derive the variational update for the cluster assignment c_i. Using Equation (18),

$$q^*(c_i; \varphi_i) \propto \exp\{\log p(c_i) + \mathbb{E}[\log p(x_i \mid c_i, \boldsymbol{\mu}); \mathbf{m}, \mathbf{s}^2]\}. \quad (22)$$

The terms in the exponent are the components of the joint density that depend on c_i. The expectation in the second term is over the mixture components $\boldsymbol{\mu}$.

The first term of Equation (22) is the log prior of c_i. It is the same for all possible values of c_i, $\log p(c_i) = -\log K$. The second term is the expected log of the c_ith Gaussian density. Recalling that c_i is an indicator vector, we can write

$$p(x_i \mid c_i, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(x_i \mid \mu_k)^{c_{ik}}.$$

We use this to compute the expected log probability,

$$\begin{aligned}&\mathbb{E}\big[\log p(x_i \mid c_i, \boldsymbol{\mu})\big] \\ &= \sum_k c_{ik} \mathbb{E}\left[\log p(x_i \mid \mu_k); m_k, s_k^2\right] \quad (23) \\ &= \sum_k c_{ik} \mathbb{E}\big[-(x_i - \mu_k)^2/2; m_k, s_k^2\big] + \text{const.} \quad (24) \\ &= \sum_k c_{ik}\big(\mathbb{E}\left[\mu_k; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2; m_k, s_k^2\right]/2\big) + \text{const.} \quad (25)\end{aligned}$$

In each line we remove terms that are constant with respect to c_i. This calculation requires $\mathbb{E}[\mu_k]$ and $\mathbb{E}[\mu_k^2]$ for each mixture component, both computable from the variational Gaussian on the kth mixture component.

Thus, the variational update for the ith cluster assignment is

$$\varphi_{ik} \propto \exp\left\{\mathbb{E}\left[\mu_k; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2; m_k, s_k^2\right]/2\right\}. \quad (26)$$

Notice it is only a function of the variational parameters for the mixture components.

3.2. The Variational Density of the Mixture-Component Means

We turn to the variational density $q(\mu_k; m_k, s_k^2)$ of the kth mixture component. Again we use Equation (18) and write down the joint density up to a normalizing constant,

$$q(\mu_k) \propto \exp\left\{\log p(\mu_k) + \sum_{i=1}^{n} \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}_{-k}^2\right]\right\}. \quad (27)$$

We now calculate the unnormalized logarithm of this coordinate-optimal $q(\mu_k)$. Recall φ_{ik} is the probability that the ith observation comes from the kth cluster. Because c_i is an indicator vector, we see that $\varphi_{ik} = \mathbb{E}[c_{ik}; \varphi_i]$. Now

$$\begin{aligned}&\log q(\mu_k) \\ &= \log p(\mu_k) + \sum_i \mathbb{E}\left[\log p(x_i|c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}_{-k}^2\right] + \text{const.} \quad (28) \\ &= \log p(\mu_k) + \sum_i \mathbb{E}\left[c_{ik} \log p(x_i \mid \mu_k); \varphi_i\right] + \text{const.} \quad (29) \\ &= -\mu_k^2/2\sigma^2 + \sum_i \mathbb{E}\left[c_{ik}; \varphi_i\right] \log p(x_i \mid \mu_k) + \text{const.} \quad (30) \\ &= -\mu_k^2/2\sigma^2 + \sum_i \varphi_{ik}\left(-(x_i - \mu_k)^2/2\right) + \text{const.} \quad (31) \\ &= -\mu_k^2/2\sigma^2 + \sum_i \varphi_{ik} x_i \mu_k - \varphi_{ik}\mu_k^2/2 + \text{const.} \quad (32) \\ &= \left(\sum_i \varphi_{ik} x_i\right)\mu_k - \left(1/2\sigma^2 + \sum_i \varphi_{ik}/2\right)\mu_k^2 + \text{const.} \quad (33)\end{aligned}$$

This calculation reveals that the coordinate-optimal variational density of μ_k is an exponential family with sufficient statistics $\{\mu_k, \mu_k^2\}$ and natural parameters $\{\sum_{i=1}^n \varphi_{ik} x_i, -1/2\sigma^2 - \sum_{i=1}^n \varphi_{ik}/2\}$, that is, a Gaussian. Expressed in terms of the variational mean and variance, the updates for $q(\mu_k)$ are

$$m_k = \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}, \quad s_k^2 = \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}. \quad (34)$$

These updates relate closely to the complete conditional density of the kth component in the mixture model. The complete conditional is a posterior Gaussian given the data assigned to the kth component. The variational update is a weighted complete conditional, where each data point is weighted by its variational probability of being assigned to component k.

Algorithm 2: CAVI for a Gaussian mixture model

Input: Data $x_{1:n}$, number of components K, prior variance of component means σ^2
Output: Variational densities $q(\mu_k; m_k, s_k^2)$ (Gaussian) and $q(z_i; \varphi_i)$ (K-categorical)
Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s^2_{1:K}$, and $\boldsymbol{\varphi} = \varphi_{1:n}$
while *the ELBO has not converged* **do**
  for $i \in \{1, \ldots, n\}$ **do**
    Set $\varphi_{ik} \propto \exp\{\mathbb{E}[\mu_k; m_k, s_k^2]x_i - \mathbb{E}[\mu_k^2; m_k, s_k^2]/2\}$
  end
  for $k \in \{1, \ldots, K\}$ **do**

$$\text{Set } m_k \longleftarrow \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

$$\text{Set } s_k^2 \longleftarrow \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

  end
  Compute $\text{ELBO}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$
end
return $q(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

3.3. CAVI for the Mixture of Gaussians

Algorithm 2 presents coordinate-ascent variational inference for the Bayesian mixture of Gaussians. It combines the variational updates in Equation (22) and Equation (34). The algorithm requires computing the ELBO of Equation (21). We use the ELBO to track the progress of the algorithm and assess when it has converged.

Once we have a fitted variational density, we can use it as we would use the posterior. For example, we can obtain a posterior decomposition of the data. We assign points to their most likely mixture assignment $\hat{c}_i = \arg\max_k \varphi_{ik}$ and estimate cluster means with their variational means m_k.

We can also use the fitted variational density to approximate the predictive density of new data. This approximate predictive is a mixture of Gaussians,

$$p(x_{\text{new}} \mid x_{1:n}) \approx \frac{1}{K} \sum_{k=1}^{K} p(x_{\text{new}} \mid m_k), \tag{35}$$

where $p(x_{\text{new}} \mid m_k)$ is a Gaussian with mean m_k and unit variance.

3.4. Empirical Study

We present two analyses to demonstrate the mixture of Gaussians algorithm in action. The first is a simulation study; the second is an analysis of a dataset of natural images.

Simulation study. Consider two-dimensional real-valued data $\mathbf{x}$. We simulate $K = 5$ Gaussians with random means, covariances, and mixture assignments. Figure 3 shows the data; each point is colored according to its true cluster. Figure 3 also illustrates the initial variational density of the mixture components—each is a Gaussian, nearly centered, and with a wide variance; the subpanels plot the variational density of the components as the CAVI algorithm progresses.

The progression of the ELBO tells a story. We highlight key points where the ELBO develops "elbows," phases of the maximization where the variational approximation changes its shape. These "elbows" arise because the ELBO is not a convex function in terms of the variational parameters; CAVI iteratively reaches better plateaus.

Finally, we plot the logarithm of the Bayesian predictive density as approximated by the variational density. Here, we report the average across held-out data. Note this plot is smoother than the ELBO.

Image analysis. We now turn to an experimental study. Consider the task of grouping images according to their color profiles. One approach is to compute the color histogram of the images. Figure 4 shows the red, green, and blue channel histograms of two images from the imageCLEF data (Villegas, Paredes, and Thomee 2013). Each histogram is a vector of length 192; concatenating the three color histograms gives a 576-dimensional representation of each image, regardless of its original size in pixel-space.

We use CAVI to fit a Gaussian mixture model with 30 clusters to image histograms. We randomly select two sets of 10,000 images from the imageCLEF collection to serve as training and testing datasets. Figure 5 shows similarly colored images assigned to four randomly chosen clusters. Figure 6 shows the average log predictive accuracy of the testing set as a function of time. We compare CAVI to an implementation in Stan (Stan Development Team 2015), which uses a Hamiltonian Monte Carlo-based sampler (Hoffman and Gelman 2014). (Details are in Appendix B.) CAVI is orders of magnitude faster than this sampling algorithm.[5]

4. Variational Inference with Exponential Families

We described mean-field variational inference and derived CAVI, a general coordinate-ascent algorithm for optimizing the ELBO. We demonstrated this approach on a simple mixture of Gaussians, where each coordinate update was available in closed form.

The mixture of Gaussians is one member of the important class of models where each complete conditional is in the exponential family. This includes a number of widely used models, such as Bayesian mixtures of exponential families, factorial mixture models, matrix factorization models, certain hierarchical regression models (e.g., linear regression, probit regression, Poisson regression), stochastic blockmodels of networks, hierarchical mixtures of experts, and a variety of mixed-membership models (which we will discuss below).

Working in this family simplifies variational inference: it is easier to derive the corresponding CAVI algorithm, and it enables variational inference to scale up to massive data. In Section 4.1, we develop the general case. In Section 4.2, we discuss conditionally conjugate models, that is, the common Bayesian application where some latent variables are "local" to a data point and others, usually identified with parameters, are "global" to the entire dataset. Finally, in Section 4.3, we describe

[5] This is not a definitive comparison between variational inference and MCMC. Other samplers, such as a collapsed Gibbs sampler, may perform better than Hamiltonian Monte Carlo sampling.

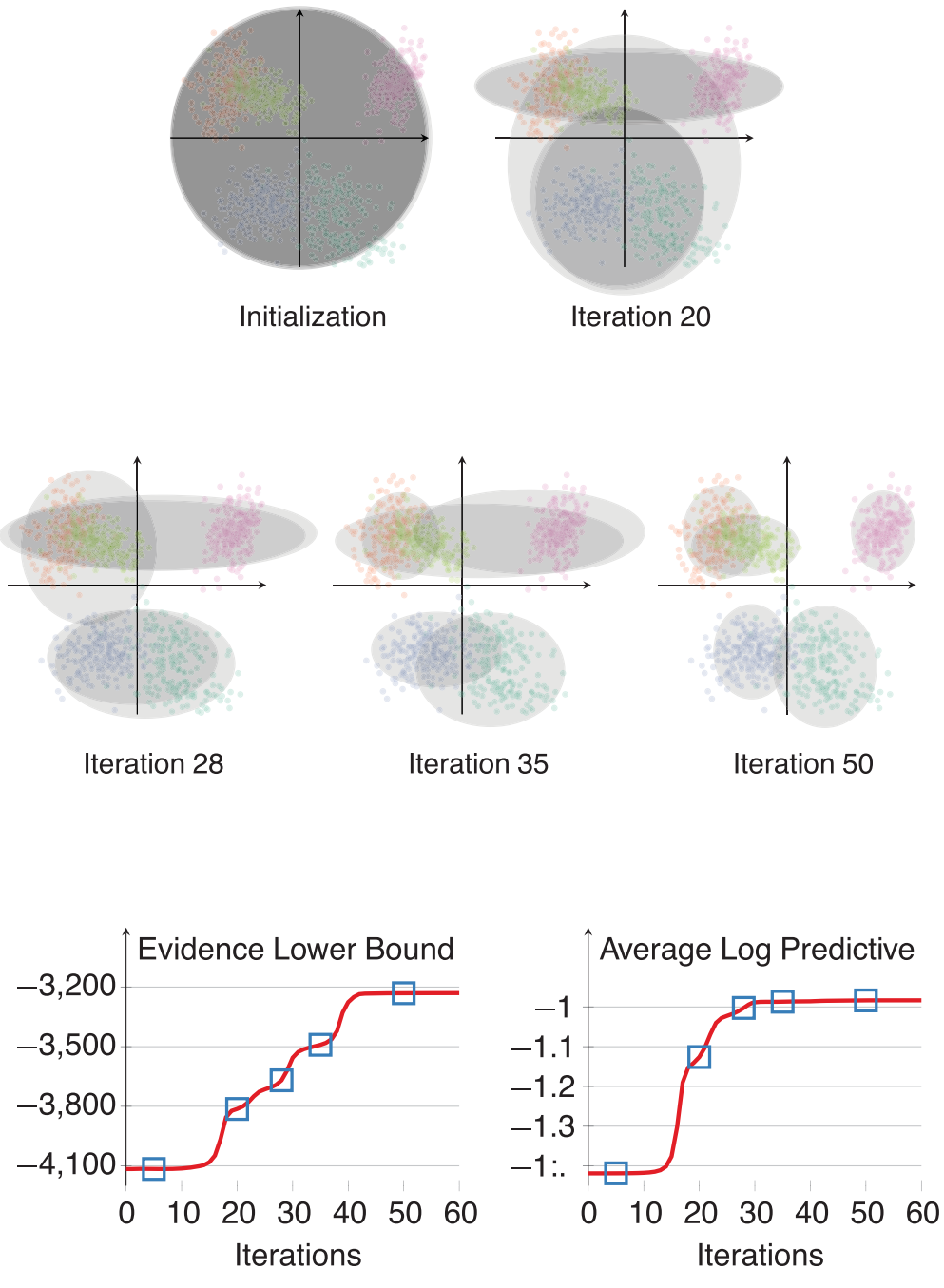


Figure 3. A simulation study of a two-dimensional Gaussian mixture model. The ellipses are 2σ contours of the variational approximating factors.

stochastic variational inference (Hoffman et al. 2013), a stochastic optimization algorithm that scales up variational inference in this setting.

4.1. Complete Conditionals in the Exponential Family

Consider the generic model $p(\mathbf{z}, \mathbf{x})$ of Section 2.1 and suppose each complete conditional is in the exponential family:

$$p(z_j \mid \mathbf{z}_{-j}, \mathbf{x}) = h(z_j) \exp\{\eta_j(\mathbf{z}_{-j}, \mathbf{x})^\top z_j - a(\eta_j(\mathbf{z}_{-j}, \mathbf{x}))\}, \tag{36}$$

where z_j is its own sufficient statistic, $h(\cdot)$ is a base measure, and $a(\cdot)$ is the log normalizer (Brown 1986). Because this is a conditional density, the parameter $\eta_j(\mathbf{z}_{-j}, \mathbf{x})$ is a function of the conditioning set.

Consider mean-field variational inference for this class of models, where we fit $q(\mathbf{z}) = \prod_j q_j(z_j)$. The exponential family assumption simplifies the coordinate update of Equation (17),

$$\begin{aligned} q(z_j) &\propto \exp\left\{\mathbb{E}\left[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})\right]\right\} && (37)\\ &= \exp\left\{\log h(z_j) + \mathbb{E}\left[\eta_j(\mathbf{z}_{-j}, \mathbf{x})\right]^\top z_j \right.\\ &\quad \left. -\mathbb{E}\left[a(\eta_j(\mathbf{z}_{-j}, \mathbf{x}))\right]\right\} && (38)\\ &\propto h(z_j) \exp\left\{\mathbb{E}\left[\eta_j(\mathbf{z}_{-j}, \mathbf{x})\right]^\top z_j\right\}. && (39) \end{aligned}$$

This update reveals the parametric form of the optimal variational factors. Each one is in the same exponential family as its corresponding complete conditional. Its parameter has the same

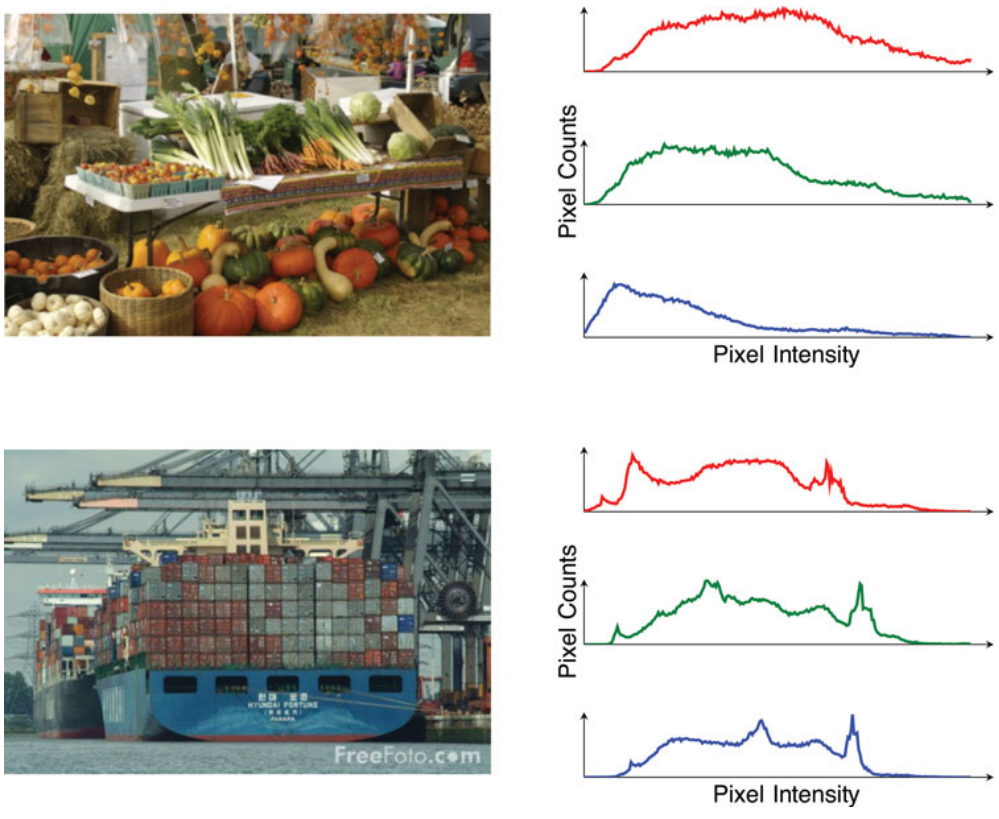


Figure 4. Red, green, and blue channel image histograms for two images from the imageCLEF dataset. The top image lacks blue hues, which is reflected in its blue channel histogram. The bottom image has a few dominant shades of blue and green, as seen in the peaks of its histogram.

Figure 5. Example clusters from the Gaussian mixture model. We assign each image to its most likely mixture cluster. The subfigures show nine randomly sampled images from four clusters; their namings are subjective.

dimension and it has the same base measure $h(\cdot)$ and log normalizer $a(\cdot)$.

Having established their parametric forms, let v_j denote the variational parameter for the jth variational factor. When we update each factor, we set its parameter equal to the expected parameter of the complete conditional,

$$v_j = \mathbb{E}\left[\eta_j(\mathbf{z}_{-j}, \mathbf{x})\right]. \tag{40}$$

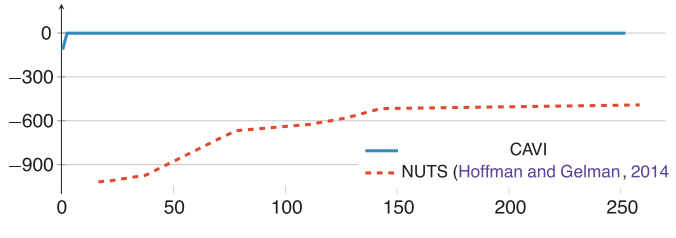


Figure 6. Comparison of CAVI to a Hamiltonian Monte Carlo-based sampling technique. CAVI fits a Gaussian mixture model to 10,000 images in less than a minute.

This expression facilitates deriving CAVI algorithms for many complex models.

4.2. Conditional Conjugacy and Bayesian Models

One important special case of exponential family models are *conditionally conjugate models* with local and global variables. Models like this come up frequently in Bayesian statistics and statistical machine learning, where the global variables are the "parameters" and the local variables are per-data-point latent variables.

Conditionally conjugate models. Let β be a vector of *global latent variables*, which potentially govern any of the data. Let $\mathbf{z}$ be a vector of *local latent variables*, whose ith component only governs data in the ith "context." The joint density is

$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i \mid \beta). \tag{41}$$

The mixture of Gaussians of Section 3 is an example. The global variables are the mixture components; the ith local variable is the cluster assignment for data point x_i.

We will assume that the modeling terms of Equation (41) are chosen to ensure each complete conditional is in the exponential family. In detail, we first assume the joint density of each (x_i, z_i) pair, conditional on β, has an exponential family form,

$$p(z_i, x_i \mid \beta) = h(z_i, x_i) \exp\{\beta^\top t(z_i, x_i) - a(\beta)\}, \quad (42)$$

where $t(\cdot, \cdot)$ is the sufficient statistic.

Next, we take the prior on the global variables to be the corresponding conjugate prior (Diaconis et al. 1979; Bernardo and Smith 1994),

$$p(\beta) = h(\beta) \exp\{\alpha^\top [\beta, -a(\beta)] - a(\alpha)\}. \quad (43)$$

This prior has natural (hyper)parameter $\alpha = [\alpha_1, \alpha_2]^\top$, a column vector, and sufficient statistics that concatenate the global variable and its log normalizer in the density of the local variables.

With the conjugate prior, the complete conditional of the global variables is in the same family. Its natural parameter is

$$\hat{\alpha} = \left[\alpha_1 + \sum_{i=1}^{n} t(z_i, x_i),\ \alpha_2 + n\right]^\top. \quad (44)$$

Turn now to the complete conditional of the local variable z_i. Given β and x_i, the local variable z_i is conditionally independent of the other local variables $\mathbf{z}_{-i}$ and other data $\mathbf{x}_{-i}$. This follows from the form of the joint density in Equation (41). Thus,

$$p(z_i \mid x_i, \beta, \mathbf{z}_{-i}, \mathbf{x}_{-i}) = p(z_i \mid x_i, \beta). \quad (45)$$

We further assume that this density is in an exponential family,

$$p(z_i \mid x_i, \beta) = h(z_i) \exp\left\{\eta(\beta, x_i)^\top z_i - a(\eta(\beta, x_i))\right\}. \quad (46)$$

This is a property of the local likelihood term $p(z_i, x_i \mid \beta)$ from Equation (42). For example, in the mixture of Gaussians, the complete conditional of the local variable is a categorical.

Variational inference in conditionally conjugate models. We now describe CAVI for this general class of models. Write $q(\beta \mid \lambda)$ for the variational posterior approximation on β; we call λ the "global variational parameter." It indexes the same exponential family density as the prior. Similarly, let the variational posterior $q(z_i \mid \varphi_i)$ on each local variable z_i be governed by a "local variational parameter" φ_i. It indexes the same exponential family density as the local complete conditional. CAVI iterates between updating each local variational parameter and updating the global variational parameter.

The local variational update is

$$\varphi_i = \mathbb{E}_\lambda\left[\eta(\beta, x_i)\right]. \quad (47)$$

This is an application of Equation (40), where we take the expectation of the natural parameter of the complete conditional in Equation (45).

The global variational update applies the same technique. It is

$$\lambda = \left[\alpha_1 + \sum_{i=1}^{n} \mathbb{E}_{\varphi_i}\left[t(z_i, x_i)\right],\ \alpha_2 + n\right]^\top. \quad (48)$$

Here, we take the expectation of the natural parameter in Equation (44).

CAVI optimizes the ELBO by iterating between local updates of each local parameter and global updates of the global parameters. To assess convergence, we can compute the ELBO at each iteration (or at some lag), up to a constant that does not depend on the variational parameters,

$$\text{ELBO} = \left(\alpha_1 + \sum_{i=1}^{n} \mathbb{E}_{\varphi_i}\left[t(z_i, x_i)\right]\right)^\top \mathbb{E}_\lambda\left[\beta\right] - (\alpha_2 + n)\mathbb{E}_\lambda\left[a(\beta)\right] - \mathbb{E}\left[\log q(\beta, \mathbf{z})\right]. \quad (49)$$

This is the ELBO in Equation (13) applied to the joint in Equation (41) and the corresponding mean-field variational density; we have omitted terms that do not depend on the variational parameters. The last term is

$$\mathbb{E}\left[\log q(\beta, \mathbf{z})\right] = \lambda^\top \mathbb{E}_\lambda\left[t(\beta)\right] - a(\lambda) + \sum_{i=1}^{n} \varphi_i^\top \mathbb{E}_{\varphi_i}\left[z_i\right] - a(\varphi_i). \quad (50)$$

CAVI for the mixture of Gaussians model (Algorithm 2) is an instance of this method. Appendix C in the online supplement presents another example of CAVI for latent Dirichlet allocation (LDA), a probabilistic topic model.

4.3. Stochastic Variational Inference

Modern applications of probability models often require analyzing massive data. However, most posterior inference algorithms do not easily scale. CAVI is no exception, particularly in the conditionally conjugate setting of Section 4.2. The reason is that the coordinate ascent structure of the algorithm requires iterating through the entire dataset at each iteration. As the dataset size grows, each iteration becomes more computationally expensive.

An alternative to coordinate ascent is gradient-based optimization, which climbs the ELBO by computing and following its gradient at each iteration. This perspective is the key to scaling up variational inference using stochastic variational inference (SVI) (Hoffman et al. 2013), a method that combines natural gradients (Amari 1998) and stochastic optimization (Robbins and Monro 1951).

SVI focuses on optimizing the global variational parameters λ of a conditionally conjugate model. The flow of computation is simple. The algorithm maintains a current estimate of the global variational parameters. It repeatedly (a) subsamples a data point from the full dataset; (b) uses the current global parameters to compute the optimal local parameters for the subsampled data point; and (c) adjusts the current global parameters in an appropriate way. SVI is detailed in Algorithm 3. We now show why it is a valid algorithm for optimizing the ELBO.

The natural gradient of the ELBO. In gradient-based optimization, the *natural gradient* accounts for the geometric structure of probability parameters (Amari 1982, 1998). Specifically, natural gradients warp the parameter space in a sensible way, so that moving the same distance in different directions amounts to equal change in symmetrized KL divergence. The usual Euclidean gradient does not enjoy this property.

In exponential families, we find the natural gradient with respect to the parameter by premultiplying the usual gradient by the inverse covariance of the sufficient statistic, $a''(\lambda)^{-1}$.

This is the inverse Riemannian metric and the inverse Fisher information matrix (Amari 1982).

Conditionally conjugate models enjoy simple natural gradients of the ELBO. We focus on gradients with respect to the global parameter λ. Hoffman et al. (2013) derived the Euclidean gradient of the ELBO,

$$\nabla_\lambda \text{ELBO} = a''(\lambda)(\mathbb{E}_\varphi[\hat{\alpha}] - \lambda), \tag{51}$$

where $\mathbb{E}_\varphi[\hat{\alpha}]$ is in Equation (48). Premultiplying by the inverse Fisher information gives the natural gradient $g(\lambda)$,

$$g(\lambda) = \mathbb{E}_\varphi[\hat{\alpha}] - \lambda. \tag{52}$$

It is the difference between the coordinate updates $\mathbb{E}_\varphi[\hat{\alpha}]$ and the variational parameters λ at which we are evaluating the gradient. In addition to enjoying good theoretical properties, the natural gradient is easier to calculate than the Euclidean gradient. For more on natural gradients and variational inference, see Sato (2001) and Honkela et al. (2008).

We can use this natural gradient in a gradient-based optimization algorithm. At each iteration, we update the global parameters,

$$\lambda_t = \lambda_{t-1} + \epsilon_t g(\lambda_t), \tag{53}$$

where ϵ_t is a step size.

Substituting Equation (52) into the second term reveals a special structure,

$$\lambda_t = (1 - \epsilon_t)\lambda_{t-1} + \epsilon_t \mathbb{E}_\varphi[\hat{\alpha}]. \tag{54}$$

Notice this does not require additional types of calculations other than those for coordinate ascent updates. At each iteration, we first compute the coordinate update. We then adjust the current estimate to be a weighted combination of the update and the current variational parameter.

Though easy to compute, using the natural gradient has the same cost as the coordinate update in Equation (48); it requires summing over the entire dataset and computing the optimal local variational parameters for each data point. With massive data, this is prohibitively expensive.

Stochastic optimization of the ELBO. Stochastic variational inference solves this problem by using the natural gradient in a stochastic optimization algorithm. Stochastic optimization algorithms follow noisy but cheap-to-compute gradients to reach the optimum of an objective function. (In the case of the ELBO, stochastic optimization will reach a local optimum.) In their seminal article, Robbins and Monro (1951) proved results implying that optimization algorithms can successfully use noisy, unbiased gradients, as long as the step size sequence satisfies certain conditions. This idea has blossomed (Kushner and Yin 1997; Spall 2003). Stochastic optimization has enabled modern machine learning to scale to massive data (Le Cun and Bottou 2004).

Our aim is to construct a cheaply computed, noisy, unbiased natural gradient. We expand the natural gradient in Equation (52) using Equation (44):

$$g(\lambda) = \alpha + \left[\sum_{i=1}^{n} \mathbb{E}_{\varphi_i^*}\left[t(z_i, x_i)\right],\ n\right]^\top - \lambda, \tag{55}$$

where φ_i^* indicates that we consider the optimized local variational parameters (at fixed global parameters λ) in Equation (47). We construct a noisy natural gradient by sampling an index from the data and then rescaling the second term,

$$t \sim \text{Unif}(1, \ldots, n) \tag{56}$$

$$\hat{g}(\lambda) = \alpha + n\left[\mathbb{E}_{\varphi_t^*}\left[t(z_t, x_t)\right],\ 1\right]^\top - \lambda. \tag{57}$$

The noisy natural gradient $\hat{g}(\lambda)$ is unbiased: $\mathbb{E}_t\left[\hat{g}(\lambda)\right] = g(\lambda)$. And it is cheap to compute—it only involves a single sampled data point and only one set of optimized local parameters. (This immediately extends to minibatches, where we sample B data points and rescale appropriately.) Again, the noisy gradient only requires calculations from the coordinate ascent algorithm. The first two terms of Equation (57) are equivalent to the coordinate update in a model with n replicates of the sampled data point.

Finally, we set the step size sequence. It must follow the conditions by Robbins and Monro (1951),

$$\sum_t \epsilon_t = \infty \quad ; \quad \sum_t \epsilon_t^2 < \infty. \tag{58}$$

Many sequences will satisfy these conditions, for example, $\epsilon_t = t^{-\kappa}$ for $\kappa \in (0.5, 1]$. The full SVI algorithm is in Algorithm 3.

We emphasize that SVI requires no new derivation beyond what is needed for CAVI. Any implementation of CAVI can be immediately scaled up to a stochastic algorithm.

Probabilistic topic models. We demonstrate SVI with a probabilistic topic model. Probabilistic topic models are mixed-membership models of text, used to uncover the latent "topics" that run through a collection of documents. Topic models have become a popular technique for exploratory data analysis of large collections (Blei 2012).

In detail, each latent topic is a distribution over terms in a vocabulary and each document is a collection of words that comes from a mixture of the topics. The topics are shared across the collection, but each document mixes them with different proportions. (This is the hallmark of a mixed-membership model.) Thus, topic modeling casts topic discovery as a posterior inference problem. Posterior estimates of the topics and topic proportions can be used to summarize, visualize, explore, and form predictions about the documents.

One motivation for topic modeling is to get a handle on massive collections of documents. Early inference algorithms were based on coordinate ascent variational inference (Blei, Ng, and Jordan 2003) and analyzed collections in the thousands or tens of thousands of documents. (Appendix C presents this algorithm). With SVI, topic models scale up to millions of documents; the details of the algorithm are in Hoffman et al. (2013). Figure 7 illustrates topics inferred using the latent Dirichlet allocation model (Blei, Ng, and Jordan 2003) from 1.8M articles from the *New York Times*. This analysis would not have been possible without SVI.

5. Discussion

We described variational inference, a method that uses optimization to make probabilistic computations. The goal is to approximate the conditional density of latent variables $\mathbf{z}$ given observed variables $\mathbf{x}$, $p(\mathbf{z} \mid \mathbf{x})$. The idea is to posit a family of densities $\mathcal{Q}$ and then to find the member $q^*(\cdot)$ that is closest

Algorithm 3: SVI for conditionally conjugate models

Input: Model $p(\mathbf{x}, \mathbf{z})$, data $\mathbf{x}$, and step size sequence ϵ_t
Output: Global variational densities $q_\lambda(\beta)$
Initialize: Variational parameters λ_0
while *TRUE* **do**
  Choose a data point uniformly at random,
  $t \sim \text{Unif}(1, \ldots, n)$
  Optimize its local variational parameters
  $\varphi_t^* = \mathbb{E}_\lambda[\eta(\beta, x_t)]$
  Compute the coordinate update as though x_t were repeated n times,

$$\hat{\lambda} = \alpha + n\mathbb{E}[\varphi_t^* f(z_t, x_t)]$$

  Update the global variational parameter,
  $\lambda_t = (1 - \epsilon_t)\lambda_t + \epsilon_t \hat{\lambda}_t$
end
return λ

in KL divergence to the conditional of interest. Minimizing the KL divergence is the optimization problem, and its complexity is governed by the complexity of the approximating family.

We then described the mean-field family, that is, the family of fully factorized densities of the latent variables. Using this family, variational inference is particularly amenable to coordinate-ascent optimization, which iteratively optimizes each factor. This approach closely connects to the classical Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990). We showed how to use mean-field VI to approximate the posterior density of a Bayesian mixture of Gaussians, discussed the special case of exponential families and conditional conjugacy, and described the extension to stochastic variational inference (Hoffman et al. 2013), which scales mean-field variational inference to massive data.

Figure 7. Topics found in a corpus of 1.8M articles from the New York Times. Reproduced with permission from Hoffman et al. (2013).

5.1. Applications

Researchers in many fields have used variational inference to solve real problems. Here, we focus on example applications of mean-field variational inference and structured variational inference based on the KL divergence. This discussion is not exhaustive; our intention is to outline the diversity of applications of variational inference.

Computational biology. VI is widely used in computational biology, where probabilistic models provide important building blocks for analyzing genetic data. For example, VI has been used in genome-wide association studies (Logsdon, Hoffman, and Mezey 2010; Carbonetto and Stephens 2012), regulatory network analysis (Sanguinetti, Lawrence, and Rattray 2006), motif detection (Xing et al. 2004), phylogenetic hidden Markov models (Jojic et al. 2004), population genetics (Raj, Stephens, and Pritchard 2014), and gene expression analysis (Stegle et al. 2010).

Computer vision and robotics. Since its inception, variational inference has been important to computer vision. Vision researchers frequently analyze large and high-dimensional datasets of images, and fast inference is important to successfully deploy a vision system. Some of the earliest examples included inferring nonlinear image manifolds (Bishop and Winn 2000) and finding layers of images in videos (Jojic and Frey 2001). As other examples, variational inference is important to probabilistic models of videos (Chan and Vasconcelos 2009; Wang and Mori 2009), image denoising (Likas and Galatsanos 2004), tracking (Vermaak, Lawrence, and Pérez 2003; Yu and Wu 2005), place recognition and mapping for robotics (Cummins and Newman 2008; Ramos et al. 2012), and image segmentation with Bayesian nonparametrics (Sudderth and Jordan 2009). Du et al. (2009) used variational inference in a probabilistic model to combine the tasks of segmentation, clustering, and annotation.

Computational neuroscience. Modern neuroscience research also requires analyzing very large and high-dimensional datasets, such as high-frequency time series data or high-resolution functional magnetic imaging data. There have been many applications of variational inference to neuroscience, especially for autoregressive processes (Roberts and Penny 2002; Penny, Kiebel, and Friston 2003; Penny, Trujillo-Barreto, and Friston 2005; Flandin and Penny 2007; Harrison and Green 2010). Other applications of variational inference to neuroscience include hierarchical models of multiple subjects (Woolrich et al. 2004), spatial models (Sato et al. 2004; Zumer et al. 2007; Kiebel et al. 2008; Wipf and Nagarajan 2009; Lashkari et al. 2012; Nathoo et al. 2014), brain-computer interfaces (Sykacek, Roberts, and Stokes 2004), and factor models (Manning et al. 2014; Gershman et al. 2014). There is a software toolbox that uses variational methods for solving neuroscience and psychology research problems (Daunizeau et al. 2014).

Natural language processing and speech recognition. In natural language processing, variational inference has been used for solving problems such as parsing (Liang et al. 2007; Liang, Jordan, and Klein 2009), grammar induction (Kurihara and Sato 2006; Naseem et al. 2010; Cohen and Smith 2010), models of streaming text (Yogatama et al. 2014), topic modeling (Blei, Ng, and Jordan 2003), and hidden Markov models and part-of-speech tagging (Wang and Blunsom 2013). In speech recognition, variational inference has been used to fit complex coupled

hidden Markov models (Reyes-Gomez, Ellis, and Jojic 2004) and switching dynamic systems (Deng 2004).

Other applications. There have been many other applications of variational inference. Fields in which it has been used include economics (Braun and McAuliffe 2010), optimal control and reinforcement learning (Van Den Broek, Wiegerinck, and Kappen 2008; Furmston and Barber 2010), statistical network analysis (Wiggins and Hofman 2008; Airoldi et al. 2008), astronomy (Regier et al. 2015), and the social sciences (Erosheva, Fienberg, and Joutard 2007; Grimmer 2011). General variational inference algorithms have been developed for a variety of classes of models, including shrinkage models (Armagan, Clyde, and Dunson 2011; Armagan and Dunson 2011; Neville, Ormerod, and Wand 2014), general time-series models (Roberts et al. 2004; Barber and Chiappa 2006; Archambeau et al. 2007a, 2007b; Johnson and Willsky 2014; Foti et al. 2014), robust models (Tipping and Lawrence 2005; Wang and Blei 2015), and Gaussian process models (Titsias and Lawrence 2010; Damianou, Titsias, and Lawrence 2011; Hensman, Fusi, and Lawrence 2013).

5.2. Theory

Though researchers have not developed much theory around variational inference, there are several threads of research about theoretical guarantees of variational approximations. As we mentioned in the introduction, one of our purposes for writing this article is to catalyze research on the statistical theory around variational inference.

Below, we summarize a variety of results. In general, they are all of the following type: treat VI posterior means as point estimates (or use M-step estimates from variational EM) and confirm that they have the usual frequentist asymptotics. (Sometimes the research finds that they do not enjoy the same asymptotics.) Each result revolves around a single model and a single family of variational approximations.

You, Ormerod, and Muller (2014) studied the variational posterior for a classical Bayesian linear model. They put a normal prior on the coefficients and an inverse gamma prior on the response variance. They found that, under standard regularity conditions, the mean-field variational posterior mean of the parameters is consistent in the frequentist sense. Ormerod, You, and Muller (2014) built on their earlier work with a spike-and-slab prior on the coefficients and found similar consistency results.

Hall, Ormerod, and Wand (2011a) and Hall et al. (2011b) examined a simple Poisson mixed-effects model, one with a single predictor and a random intercept. They used a Gaussian variational approximation and estimated parameters with variational EM. They proved consistency of these estimates at the parametric rate and showed asymptotic normality with asymptotically valid standard errors.

Celisse et al. (2012) and Bickel et al. (2013) analyzed network data using stochastic blockmodels. They showed asymptotic normality of parameter estimates obtained using a mean-field variational approximation. They highlighted the computational advantages and theoretical guarantees of the variational approach over maximum likelihood for dense, sparse, and restricted variants of the stochastic blockmodel.

Westling and McCormick (2015) studied the consistency of VI through a connection to M-estimation. They focused on a broader class of models (with posterior support in real coordinate space) and analyzed an automated VI technique that uses a Gaussian variational approximation (Kucukelbir et al. 2015). They derived an asymptotic covariance matrix estimator of the variational approximation and showed its robustness to model misspecification.

Finally, Wang and Titterington (2006) analyzed variational approximations to mixtures of Gaussians. Specifically, they considered Bayesian mixtures with conjugate priors, the mean-field variational approximation, and an estimator that is the variational posterior mean. They confirmed that CAVI converges to a local optimum, that the VI estimator is consistent, and that the VI estimate and maximum likelihood estimate (MLE) approach each other at a rate of $\mathcal{O}(1/n)$. Wang and Titterington (2005), showed that the asymptotic variational posterior covariance matrix is "too small"—it differs from the MLE covariance (i.e., the inverse Fisher information) by a positive-definite matrix.

5.3. Beyond Conditional Conjugacy

We focused on models where the complete conditional is in the exponential family. Many models, however, do not enjoy this property. A simple example is Bayesian logistic regression,

$$\beta_k \sim \mathcal{N}(0, 1),$$
$$y_i \mid x_i, \beta \sim \text{Bern}(\sigma(\beta^\top x_i)),$$

where $\sigma(\cdot)$ is the logistic function. The posterior density of the coefficients is not in an exponential family and we cannot apply the variational inference methods we discussed above. Specifically, we cannot compute the expectations in the first term of the ELBO in Equation (13) or the coordinate update in Equation (18).

Exploring variational methods for such models has been a fruitful area of research. An early example is Jaakkola and Jordan (1997, 2000), who developed a variational bound tailored to logistic regression. Blei and Lafferty (2007) later adapted their idea to nonconjugate topic models, and researchers have continued to improve the original bound (Khan et al. 2010; Marlin, Khan, and Murphy 2011; Ermis and Bouchard 2014). In other work, Braun and McAuliffe (2010) derived a variational inference algorithm for the discrete choice model, which also lies outside of the class of conditionally conjugate models. They developed a delta method to approximate the difficult-to-compute expectations. Finally, Wand et al. (2011) used auxiliary variable methods, quadrature, and mixture approximations to handle a variety of likelihood terms that fall outside of the exponential family.

More recently, researchers have generalized nonconjugate inference, seeking recipes that can be used across many models. Wang and Blei (2013) adapted Laplace approximations and the delta method to this end, improving inference in nonconjugate generalized linear models and topic models; this approach is also used by Bugbee, Breidt, and van der Woerd (2016) for

semiparametric regression. Knowles and Minka (2011) generalized the Jaakkola and Jordan (1997, 2000) bound in a message-passing algorithm and Wand (2014) further simplified and extended their approach. Tan and Nott (2013, 2014) applied these message-passing methods to generalized linear mixed models (and also combined them with SVI). Rohde and Wand (2016) unified many of these algorithmic developments and provided practical insights into their numerical implementations.

Finally, there has been a flurry of research on optimizing difficult variational objectives with Monte Carlo (MC) estimates of the gradient. The idea is to write the gradient of the ELBO as an expectation, compute MC estimates of it, and then use stochastic optimization with repeated MC gradients. This first appeared independently in several articles (Ji, Shen, and West 2010; Nott et al. 2012; Paisley, Blei, and Jordan 2012; Wingate and Weber 2013). The newest approaches avoid any model-specific derivations, and are termed "black box" inference methods. As examples, see Kingma and Welling (2014); Rezende, Mohamed, and Wierstra (2014); Ranganath, Gerrish, and Blei (2014); Ranganath, Tran, and Blei (2016); Salimans and Knowles (2014); Titsias and Lázaro-Gredilla (2014); and Tran, Ranganath, and Blei (2016). Kucukelbir et al. (2017) leveraged these ideas toward an automatic VI technique that works on any model written in the probabilistic programming system Stan (Stan Development Team 2015). This is a step toward a derivation-free, easy-to-use VI algorithm.

5.4. Open Problems

There are many open avenues for statistical research in variational inference.

We focused on optimizing $\mathrm{KL}\left(q(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})\right)$ as the variational objective function. A promising avenue of research is to develop variational inference methods that optimize other measures, such as α-divergence measures. As one example, expectation propagation (Minka 2001) is inspired by the KL divergence "in the other direction," between $p(\mathbf{z}\,|\,\mathbf{x})$ and $q(\mathbf{z})$. Other work has developed divergences based on lower bounds that are tighter than the ELBO (Barber and de van Laar 1999; Leisink and Kappen 2001). While alternative divergences may be difficult to optimize, they may give better approximations (Minka 2005; Opper and Winther 2005).

Though it is flexible, the mean-field family makes strong independence assumptions. These assumptions help with scalable optimization, but they limit the expressibility of the variational family. Further, they can exacerbate issues with local optima of the objective and underestimating posterior variances; see Figure 1. A second avenue of research is to develop better approximations while maintaining efficient optimization.

As we mentioned previously, structured variational inference has its roots in the early days of the method (Saul and Jordan 1996; Barber and Wiegerinck 1999). More recently, Hoffman and Blei (2015) used generic structured variational inference in a stochastic optimization algorithm; Kucukelbir et al. (2017), Challis and Barber (2013), and Tan and Nott (2017) took advantage of Gaussian variational families with nondiagonal covariance; Giordano, Broderick, and Jordan (2015) post-processed the mean-field parameters to correct for underestimating the variance; and Ranganath, Tran, and Blei (2016) embedded the mean-field parameters themselves in a hierarchical model to induce variational dependencies between latent variables.

The interface between variational inference and MCMC remains relatively unexplored. de Freitas et al. (2001) used fitted variational distributions as a component of a proposal distribution for Metropolis–Hastings. Hoffman, Blei, and Mimno (2012) and Hoffman and Blei (2015) studied MCMC as a method of approximating coordinate updates, for example, to include structure in the variational family. Salimans, Kingma, and Welling (2015) proposed a variational approximation to the MCMC chain; their method enables an explicit trade off between computational accuracy and speed. Understanding how to combine these two strategies for approximate inference is a ripe area for future research. A principled analysis of when to use (and combine) variational inference and MCMC would have both theoretical and practical impact in the field.

Finally, the statistical properties of variational inference are not yet well understood, especially in contrast to the wealth of analysis of MCMC techniques. There has been some progress; see Section 5.2. A final open research problem is to understand variational inference as an estimator and to understand its statistical profile relative to the exact posterior.

Supplementary Materials

The online supplementary materials contain the appendices for the article.

References

Ahmed, A., Aly, M., Gonzalez, J., Narayanamurthy, S., and Smola, A. (2012), "Scalable Inference in Latent Variable Models," in *International Conference on Web Search and Data Mining*, pp. 123–132. [860]

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [872]

Amari, S. (1982), "Differential Geometry of Curved Exponential Families-Curvatures and Information Loss," *The Annals of Statistics*, 10, 357–385. [869]

——— (1998), "Natural Gradient Works Efficiently in Learning," *Neural Computation*, 10, 251–276. [869]

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007a), "Gaussian Process Approximations of Stochastic Differential Equations," *Workshop on Gaussian Processes in Practice*, 1, 1–16. [872]

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-Taylor, J. (2007b), "Variational Inference for Diffusion Processes," in *Neural Information Processing Systems*, pp. 17–24. [872]

Armagan, A., Clyde, M., and Dunson, D. (2011), "Generalized Beta Mixtures of Gaussians," in *Neural Information Processing Systems*, pp. 523–531. [872]

Armagan, A., and Dunson, D. (2011), "Sparse Variational Analysis of Linear Mixed Models for Large Data Sets," *Statistics & Probability Letters*, 81, 1056–1062. [872]

Barber, D. (2012), *Bayesian Reasoning and Machine Learning*, Cambridge, UK: Cambridge University Press. [860]

Barber, D., and Bishop, C. M. (1998), "Ensemble Learning in Bayesian Neural Networks," in *Generalization in Neural Networks and Machine Learning*, ed. C. M. Bishop, New York: Springer Verlag, pp. 215–237. [860]

Barber, D., and Chiappa, S. (2006), "Unified Inference for Variational Bayesian Linear Gaussian State-Space Models," in *Neural Information Processing Systems*, pp. 81–88. [872]

Barber, D., and de van Laar, P. (1999), "Variational Cumulant Expansions for Intractable Distributions," *Journal of Artificial Intelligence Research*, 10, 435–455. [873]