Advanced Bayesian Methodology

Logistics: presenter sign-ups

- most of you bid—thank you!
- signup form: https://tinyurl.com/stat451signup
- next week: jlederman (Tues) and rakoort (Thurs)
- presentations—your choice, depends on material
  - slides
  - chalk
  - code(?)
  - purpose: foster discussion, dig into details
  - we can meet to discuss

Logistics: reader reports

- purpose: document your own journey through the lit
  - what other papers did you have to read?
  - what other papers did you draw connections to?
  - use Zotero (or equivalent), and build-up a bibliography
- style: loose, unique to your journey
- format: 2-3 paragraphs; will upload LaTeX template to Github
- due at the beginning of Thurs class

Variational Inference: A Review for Statisticians [Blei et al., 2017]

Bayesian modeling

A probabilistic model is a joint distribution

$$P(\boldsymbol{x}, \boldsymbol{z}) = P(\boldsymbol{x} \mid \boldsymbol{z})\, P(\boldsymbol{z}) \tag{1}$$

that relates **latent variables** $\boldsymbol{z} = z_{1:m}$ to **observations** $\boldsymbol{x} = x_{1:n}$.

The posterior distribution characterizes our uncertainty about the latent variables

$$P(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{\overbrace{P(\boldsymbol{x} \mid \boldsymbol{z})}^{\text{likelihood}}\ \overbrace{P(\boldsymbol{z})}^{\text{prior}}}{\underbrace{P(\boldsymbol{x})}_{\text{evidence}}} \tag{2}$$

This is the target of Bayesian inference.

Variational inference

Define family of approximate (variational) densities $\mathcal{Q}$.

Find the member that minimizes:

$$q^*(\boldsymbol{z}) = \arg\min_{q(\boldsymbol{z}) \in \mathcal{Q}} \mathrm{KL}\left(q(\boldsymbol{z}) \,||\, p(\boldsymbol{z} \mid \boldsymbol{x})\right) \tag{3}$$

Here we are minimizing with respect to $q(\boldsymbol{z})$ itself—the term variational inference comes from "calculus of variations".

The evidence lower bound (ELBO)

How do you minimize the KL divergence to *a distribution you cannot form*?

$$\mathrm{KL}\Big(q(\boldsymbol{z}) \mid\mid p(\boldsymbol{z} \mid \boldsymbol{x})\Big) = \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z} \mid \boldsymbol{x})}\right] \tag{4}$$

$$= \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}, \boldsymbol{x})}\right] + \underbrace{\log p(\boldsymbol{x})}_{\text{(log) evidence}} \tag{5}$$

$$\propto_q \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z}, \boldsymbol{x})}\right] \tag{6}$$

Re-arranging terms reveals the *evidence lower bound (ELBO)*:

$$\underbrace{\log p(\boldsymbol{x})}_{\text{(log) evidence}} = \underbrace{\mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{z}, \boldsymbol{x})}{q(\boldsymbol{z})}\right]}_{\text{evidence lower bound (ELBO)}} + \underbrace{\mathrm{KL}\Big(q(\boldsymbol{z}) \mid\mid p(\boldsymbol{z} \mid \boldsymbol{x})\Big)}_{\geq 0} \tag{7}$$

Maximizing the ELBO (which we can form / compute) $\equiv$ minimizing the KL.

ELBO surgery

What solutions for $q(\boldsymbol{z})$ will the ELBO encourage?

$$\text{ELBO}(q) = \underbrace{\mathbb{E}_q[\log p(\boldsymbol{z}, \boldsymbol{x})]}_{\text{exp. complete log like.}} - \underbrace{\mathbb{E}_q[\log q(\boldsymbol{z})]}_{\text{-entropy}} \tag{8}$$

$$= \underbrace{\mathbb{E}_q[\log p(\boldsymbol{x} \mid \boldsymbol{z})]}_{\text{exp. log likelihood}} - \underbrace{\text{KL}(q(\boldsymbol{z}) \,||\, p(\boldsymbol{z}))}_{\text{regularizer}} \tag{9}$$

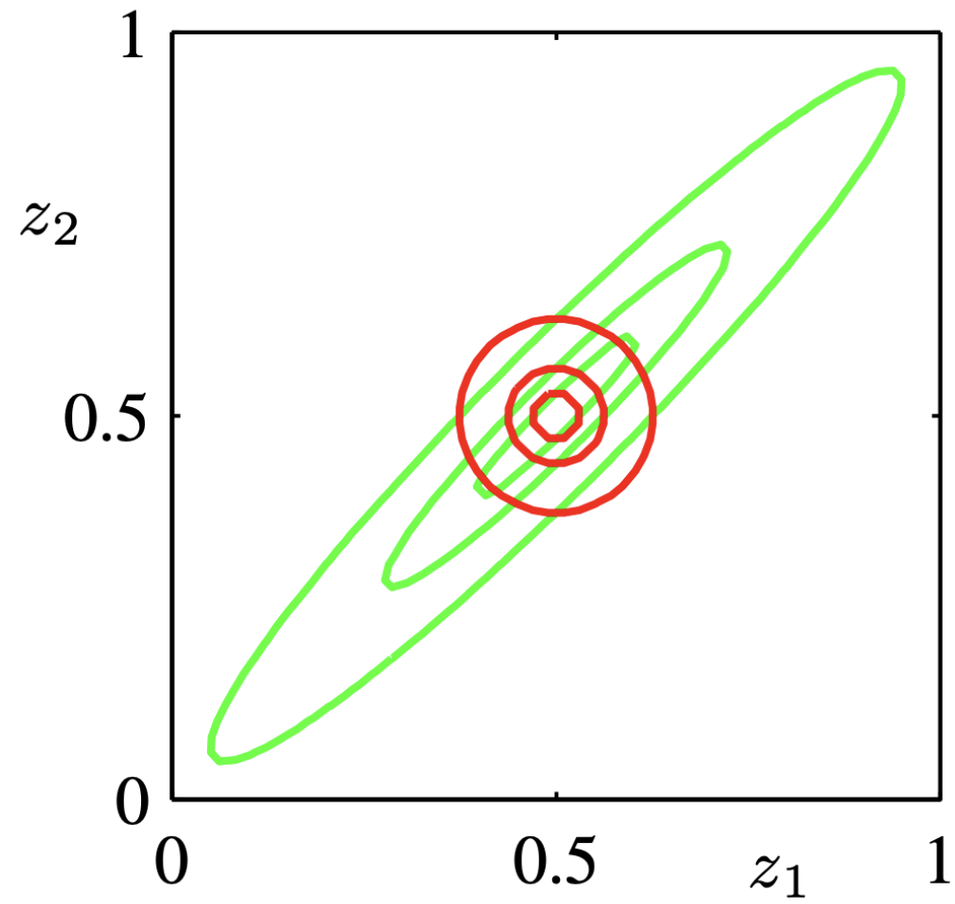
Mean field approximation

Choose a factorized or "mean-field" variational family

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q(z_j) \tag{10}$$

Every latent variable z_j is governed by its own $q(z_j)$ with its own variational parameters.

Figure from [Bishop, 2006, Chap. 10]: Green is exact posterior $P(z_1, z_2 \mid \boldsymbol{x})$. Red is the best-fit mean-field approximation. Notice that there is no correlation between z_1 and z_2.



Very simple variational family facilitates optimization...

Coordinate ascent variational inference

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 $q(\mathbf{z})$

Coordinate ascent variational inference

The result from Bishop [2006] is that the optimal factor is *proportional* to

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{\neg j}\left[\log p(z_j \mid \boldsymbol{z}_{\neg j}, \boldsymbol{x})\right]\right\} \tag{11}$$

Let's be explicit about what this means—the optimal factor is *equal* to

$$q_j^*(z_j) = \frac{1}{\mathcal{C}_j} \times \exp\left\{\mathbb{E}_{\neg j}\left[\log p(z_j \mid \boldsymbol{z}_{\neg j}, \boldsymbol{x})\right]\right\} \tag{12}$$

where we have introduced the *normalizer* (constant w.r.t. z_j)

$$\mathcal{C}_j \triangleq \int \mathrm{d}z_j \, \exp\left\{\mathbb{E}_{\neg j}\left[\log p(z_j \mid \boldsymbol{z}_{\neg j}, \boldsymbol{x})\right]\right\} \tag{13}$$

Coordinate ascent variational inference

Collect all the terms proportional to $q_j(z_j)$:

$$\mathrm{ELBO}(q_j) = \mathbb{E}_j[\mathbb{E}_{\neg j}[\log p(\boldsymbol{z}, \boldsymbol{x})] - \mathbb{E}_j[\log q_j(\boldsymbol{z})]] + \text{cons.} \tag{14}$$

Claim: $q_j^*(z_j) = \arg\max_{q_j} \mathrm{ELBO}(q_j)$.

Proof: Set up the KL divergence to the (supposed) optimal $q^*(z_j)$

$$\mathrm{KL}(q(z_j) \,||\, q^*(z_j)) = \int \mathrm{d}z_j \, q(z_j) \Big[\log q(z_j) - \log q^*(z_j) \Big] \tag{15}$$

On the last slide we said $q^*(z_j) = \exp\{\cdots\}/\mathcal{C}_j$

$$= \int \mathrm{d}z_j \, q(z_j) \Big[\log q(z_j) - \mathbb{E}_{\neg j} \big[\log p(z_j \mid \boldsymbol{z}_{\neg j}, \boldsymbol{x}) \big] + \log \mathcal{C}_j \Big] \tag{16}$$

Factoring out $\log p(\boldsymbol{z}_{\neg j}, \boldsymbol{x})$ and combining it with $\mathcal{C}_j$

$$= \mathbb{E}_j \left[\log q(z_j) \right] - \mathbb{E}_j \left[\mathbb{E}_{\neg j} \big[\log p(z_j, \boldsymbol{z}_{\neg j}, \boldsymbol{x}) \big] \right] + \text{const} \tag{17}$$

This is the negative of the objective

$$= -\mathrm{ELBO}(q_j) \tag{18}$$

Section 3: A complete example: Bayesian mixture of Gaussians

Bayesian mixture of Gaussians

Generative process:

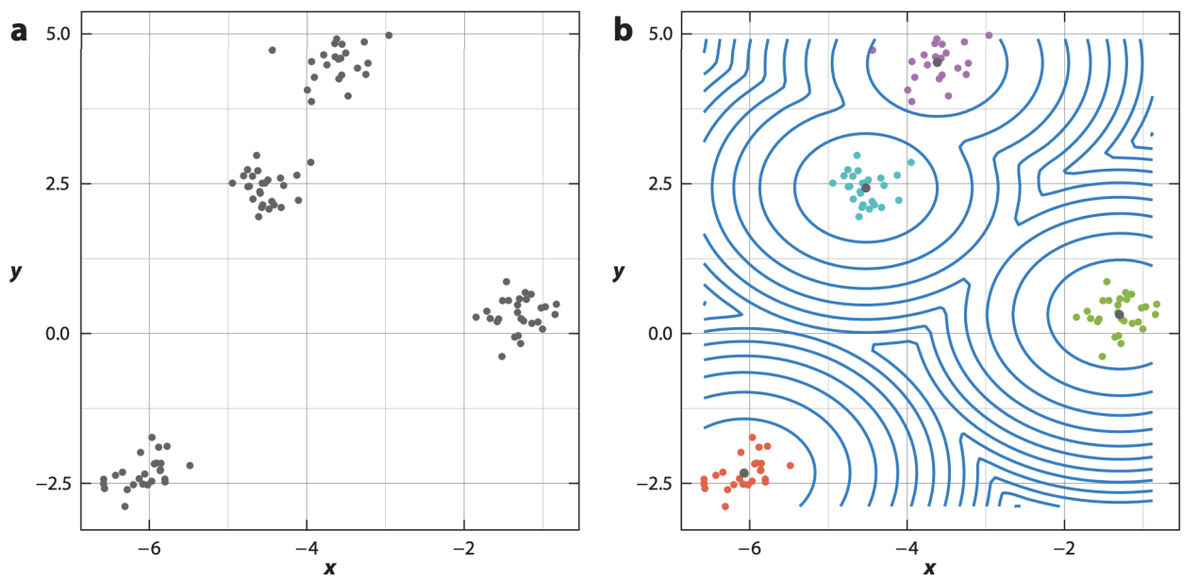
$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2) && \text{for } k = 1 \dots K \\ c_i &\sim \text{Categorical}(1/K \dots 1/K) && \text{for } i = 1 \dots n \\ x_i &\sim \mathcal{N}(c_i^\top \boldsymbol{\mu}, 1) && \text{for } i = 1 \dots n \end{aligned}$$

Note that c_i is "one-hot" encoded.

$$c_i = \begin{bmatrix} c_{i1} \\ \vdots \\ c_{iK} \end{bmatrix}$$

The shorthand $c_i = k$, means $c_{ik} = 1$ and $c_{ik'} = 0$ for all $k' \neq k$.

Figure from [Blei, 2014]
(ignore the x- and y-labels):



Bayesian mixture of Gaussians

Model evidence is intractable

$$
\begin{aligned}
p(\boldsymbol{x}) &= \sum_{\boldsymbol{c}} p(\boldsymbol{c}) \int \mathrm{d}\boldsymbol{\mu}\, p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(\boldsymbol{x} \mid \boldsymbol{\mu}, c_i) \\
&= \underbrace{\sum_{c_1=1}^{K} \cdots \sum_{c_n=1}^{K}}_{K^n \text{ summands}} p(c_{1:n}) \int \mathrm{d}\boldsymbol{\mu}\, p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(\boldsymbol{x} \mid \boldsymbol{\mu}, c_i)
\end{aligned}
$$

(Posterior is too.)

CAVI for Bayesian mixture of Gaussians

Define the mean-field family

$$q(\boldsymbol{c}, \boldsymbol{\mu}) = \left[\prod_{i=1}^{n} q(c_i)\right] \left[\prod_{k=1}^{K} q(\mu_k)\right]$$

(We will only seek to infer $\boldsymbol{c}$ and $\boldsymbol{\mu}$, and treat σ^2 as a (hyper)parameter.)

Deriving the CAVI update for c_i

We will follow the exact recipe given by Blei et al. [2017] to derive $q^*(c_i)$.

(Later we'll try a different recipe for $q^*(\mu_k)$.)

$$q^*(c_i) \propto_{c_i} \exp\{\, \mathbb{E}[\, \log \underbrace{p(c_i, \boldsymbol{c}_{\neg i}, \boldsymbol{\mu}, \boldsymbol{x})}_{\text{full joint}} \,] \,\}$$

Here the expectation is under q with respect to all latent variables except c_i.

$$\propto_{c_i} \exp\{\log p(c_i) + \mathbb{E}[\, \log p(\boldsymbol{c}_{\neg i}, \boldsymbol{\mu}, \boldsymbol{x} \mid c_i) \,] \,\}$$

Factorizing the second term

$$\propto_{c_i} \exp\{\log p(c_i) + \mathbb{E}[\, \log p(\boldsymbol{x} \mid c_i, \boldsymbol{c}_{\neg i}, \boldsymbol{\mu})] + \mathbb{E}[\, \log p(\boldsymbol{c}_{\neg i}, \boldsymbol{\mu} \mid c_i) \,] \,\}$$

Conditional independence: $p(\boldsymbol{c}_{\neg i}, \boldsymbol{\mu} \mid c_i) = p(\boldsymbol{c}_{\neg i}, \boldsymbol{\mu})$

$$\propto_{c_i} \exp\{\log p(c_i) + \mathbb{E}[\, \log p(\boldsymbol{x} \mid c_i, \boldsymbol{\mu})] \,\}$$

More conditional independence $p(x_i \mid c_i, \boldsymbol{c}_{\neg i}, \boldsymbol{\mu}) = \prod_{i=1}^n p(x_i \mid c_i, \boldsymbol{\mu})$

$$\propto_{c_i} \exp\{\log p(c_i) + \mathbb{E}[\, \log p(x_i \mid c_i, \boldsymbol{\mu})] \,\}$$

Recall that the prior is constant $p(c_i) = 1/K$ and the likelihood is normal

$$\propto_{c_i} \exp\{\mathbb{E}[\, \log p(x_i \mid c_i, \boldsymbol{\mu})] \,\}$$

Deriving the CAVI update for c_i

Recall that $c_i = [c_{i1} \dots c_{iK}]^\top$ and $c_i = k$ means $c_{ik} = 1$ and $c_{ik'} = 0$ for $k' \neq k$.

We can rewrite the log-likelihood as

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

Plugging in the form above and substituting in the normal likelihood

$$\begin{aligned}
q^*(c_i) &\propto_{c_i} \exp\left\{ \sum_{k=1}^{K} c_{ik}\, \mathbb{E}\Big[\log \mathcal{N}(x_i; \mid \mu_k,\, 1) \Big] \right\} \\
&\propto_{c_i} \exp\left\{ \sum_{k=1}^{K} c_{ik} \mathbb{E}\Big[-\tfrac{1}{2}(x_i - \mu_k)^2 \Big] \right\} \\
&\propto_{c_i} \exp\left\{ \sum_{k=1}^{K} c_{ik}\, \mathbb{E}\Big[-\tfrac{1}{2}x_i^2 + x_i\mu_k - \tfrac{1}{2}\mu_k^2 \Big] \right\} \\
&\propto_{c_i} \exp\left\{ \sum_{k=1}^{K} c_{ik} \left(x_i \mathbb{E}[\mu_k] - \tfrac{1}{2}\mathbb{E}[\mu_k^2] \right) \right\}
\end{aligned}$$

Deriving the CAVI update for c_i

From the previous slide:

$$q^*(c_i) \propto_{c_i} \exp\left\{ \sum_{k=1}^{K} c_{ik} \left(x_i \mathbb{E}[\mu_k] - \tfrac{1}{2}\mathbb{E}[\mu_k^2] \right) \right\}$$

c_i only takes K values—the probability of a given value k is

$$q^*(c_i = k) \propto_{c_{ik}} \exp\left\{ x_i \mathbb{E}[\mu_k] - \tfrac{1}{2}\mathbb{E}[\mu_k^2] \right\}$$

Since $q^*(c_i)$ is a discrete distribution characterized by K probabilities, its variational parameters *are* those probabilities. Define $\psi_i = [\psi_{i1} \dots \psi_{iK}]^\top$, where $q^*(c_i = k) = \psi_{ik}$.

$$q^*(c_i = k; \psi_i) = \psi_{ik} = \frac{\exp\left\{ x_i \mathbb{E}[\mu_k] - \tfrac{1}{2}\mathbb{E}[\mu_k^2] \right\}}{\sum_{k'=1}^{K} \exp\left\{ x_i \mathbb{E}[\mu_{k'}] - \tfrac{1}{2}\mathbb{E}[\mu_{k'}^2] \right\}}$$

This defines the optimal CAVI update. Notice it is in terms of $\mathbb{E}[\mu_k]$ and $\mathbb{E}[\mu_k^2]$.

This highlights the sense in which CAVI is a "message passing" algorithm...

...$q(c_i)$ needs $q(\mu_k)$ to send over its first two moments ("messages").

Deriving the CAVI update for μ_k

The usual recipe is

$$q^*(\mu_k) \propto_{\mu_k} \exp\{\, \mathbb{E}[\, \log \underbrace{p(\mu_k, \boldsymbol{\mu}_{\neg k}, \boldsymbol{c}, \boldsymbol{x})}_{\text{full joint}} \,] \,\}$$

But let's try something else...

Since $\log p(\boldsymbol{\mu}_{\neg k}, \boldsymbol{c}, \boldsymbol{x})$ is constant w.r.t. μ_k we can substract it... doing so gives us

$$q^*(\mu_k) \propto_{\mu_k} \exp\{\, \mathbb{E}[\, \log \underbrace{p(\mu_k \mid \boldsymbol{\mu}_{\neg k}, \boldsymbol{c}, \boldsymbol{x})}_{\text{complete conditional}} \,] \,\}$$

When we have closed-form complete conditionals (we almost always do with CAVI) this offers an alternate route, and also reveals some of the deeper structure of CAVI.

Normal-normal conjugacy

Normal prior and likelihood

$$\begin{aligned} \mu &\sim \mathcal{N}(\mu_0,\, \sigma^2) \\ x_i &\sim \mathcal{N}(\mu,\, \gamma^2) \qquad\qquad \text{for } i = 1 \dots n \end{aligned}$$

The posterior is also normal

$$P(\mu \mid x_{1:n},\, \gamma^2, \mu_0,\, \sigma^2) = \mathcal{N}\left(\mu; \tilde{m},\, \tilde{s}^2\right)$$

where the *posterior parameters* are defined as

$$\tilde{m} \triangleq \left(\frac{\gamma^2}{\gamma^2 + n\,\sigma^2}\right)\mu_0 + \left(1 - \frac{\gamma^2}{\gamma^2 + n\,\sigma^2}\right)\frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \tilde{s}^2 \triangleq \frac{\gamma^2}{\gamma^2/\sigma^2 + n}$$

Special case when $\mu_0 = 0$ and $\gamma^2 = 1$ (our case):

$$\tilde{m} = \frac{\sum_{i=1}^{n} x_i}{1/\sigma^2 + n}, \qquad \tilde{s}^2 = \frac{1}{1/\sigma^2 + n}$$

Conjugacy in the normal mixture model

Here's the model again:

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2) && \text{for } k = 1 \dots K \\ c_i &\sim \text{Categorical}(1/K \dots 1/K) && \text{for } i = 1 \dots n \\ x_i &\sim \mathcal{N}(c_i^\top \boldsymbol{\mu}, 1) && \text{for } i = 1 \dots n \end{aligned}$$

Conditional on $\boldsymbol{c} = c_{1:n}$, this is a normal-normal conjugate model for each μ_k

$$P(\mu_k \mid \boldsymbol{x}, \boldsymbol{c}, \sigma^2) = P(\mu_k \mid \boldsymbol{x}, \boldsymbol{c}, \sigma^2) = \mathcal{N}\left(\mu_k;\, \tilde{m}_k,\, \tilde{s}_k^2\right)$$

where the posterior mean and variance are defined as

$$\tilde{m}_k \triangleq \frac{\sum_{i=1}^n c_{ik}\, x_i}{1/\sigma^2 + \sum_{i=1}^n c_{ik}}, \qquad \tilde{s}_k^2 \triangleq \frac{1}{1/\sigma^2 + \sum_{i=1}^n c_{ik}}$$

(Recall that $c_i = [c_{i1} \dots c_{iK}]^\top$ where $c_i = k$ is shorthand for $c_{ik} = 1$.)

Using conjugacy to derive $q^*(\mu_k)$

Let's appeal to normal-normal conjugacy to rederive the optimal update.

The optimal variational factor for μ_k is

$$q^*(\mu_k) \propto_{\mu_k} \exp\{\, \mathbb{E}[\log \underbrace{p(\mu_k \mid \boldsymbol{\mu}_{\neg k}, \boldsymbol{c}, \boldsymbol{x})}_{\text{complete conditional}}]\,\}$$

Try plugging in the normal PDF

$$\begin{aligned}
&\propto_{\mu_k} \exp\left\{\mathbb{E}\Big[\log \mathcal{N}\left(\mu_k;\, \tilde{m}_k,\, \tilde{s}_k^2\right)\Big]\right\} \\
&= \exp\left\{\mathbb{E}\Big[\log \tfrac{1}{\tilde{s}_k^2\sqrt{2\pi}}\Big] + \mathbb{E}\Big[\tfrac{1}{2\tilde{s}_k^2}(\mu_k - \tilde{m}_k)^2\Big]\right\} \\
&= \exp\left\{\mathbb{E}\Big[\log \tfrac{1}{\tilde{s}_k^2\sqrt{2\pi}}\Big] + \mathbb{E}\Big[\tfrac{1}{2\tilde{s}_k^2}(\mu_k^2 + 2\mu_k\tilde{m}_k - \tilde{m}_k^2)\Big]\right\}
\end{aligned}$$

Both $\tilde{m}_k$ and $\tilde{s}_k$ involve variables governed by the expectation... kinda messy...

...recall that the normal is an exponential family...

Normal as an exponential family

The normal PDF can be written in exponential family form as

$$\mathcal{N}(x;\, \mu, \sigma^2) = h(x) \exp\left\{\eta(\mu, \sigma^2)^\top t(x) - A(\mu, \sigma^2)\right\}$$

The *base measure* is $h(x) = \frac{1}{\sqrt{2\pi\sigma^2}}$.

The *log normalizer* is $A(\mu, \sigma^2) = {}^{\mu^2}/_{2\sigma^2} + \log\sigma$.

The *sufficient statistic* is $t(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$.

The *natural parameter* mapping is $\eta(\mu, \sigma^2) = \begin{bmatrix} \eta_1(\mu, \sigma^2) \\ \eta_2(\mu, \sigma^2) \end{bmatrix} = \begin{bmatrix} {}^{\mu}/_{\sigma^2} \\ {}^{-1}/_{2\sigma^2} \end{bmatrix}$.

For a given value $\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$, the *inverse mapping* is $\eta^{-1}(\eta_1, \eta_2) = \begin{bmatrix} {}^{-\eta_1}/_{2\eta_2} \\ {}^{-1}/_{2\eta_2} \end{bmatrix} = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$.

Using conjugacy to derive $q^*(\mu_k)$

Let's try again using the exponential family form...

$$q^*(\mu_k) \propto_{\mu_k} \exp\left\{\mathbb{E}\Big[\log \underbrace{h(\mu_k)\exp\left\{\eta(\tilde{m}_k, \tilde{s}_k^2)^\top t(\mu_k) - A(\tilde{m}_k, \tilde{s}_k^2)\right\}}_{\text{complete conditional } P(\mu_k \mid -) = \mathcal{N}(\mu_k;\, \tilde{m}_k, \tilde{s}_k^2)}\Big]\right\}$$

(Notice that $\eta(\tilde{m}_k, \tilde{s}_k^2)$ and $A(\tilde{m}_k, \tilde{s}_k^2)$ are functions of the complete conditional's parameters, which are themselves functions of other latent variables.)

$$\propto_{\mu_k} \exp\left\{\mathbb{E}\Big[\log h(\mu_k)\Big] + \mathbb{E}\Big[\eta(\tilde{m}_k, \tilde{s}_k^2)^\top t(\mu_k)\Big] - \mathbb{E}\Big[A(\tilde{m}_k, \tilde{s}_k^2)\Big]\right\}$$

The log normalizer is constant with respect to μ_k

$$\propto_{\mu_k} \exp\left\{\mathbb{E}\Big[\log h(\mu_k)\Big] + \mathbb{E}\Big[\eta(\tilde{m}_k, \tilde{s}_k^2)^\top t(\mu_k)\Big]\right\}$$

The expectation $\mathbb{E}[\,\cdot\,]$ is with respect to all latent variables *except* μ_k.

$$\propto_{\mu_k} h(\mu_k) \exp\left\{\mathbb{E}\big[\eta(\tilde{m}_k, \tilde{s}_k^2)\big]^\top t(\mu_k)\right\}$$

We recognize this as the kernel (unnormalized density) of another normal distribution...
...with natural parameter equal to $\mathbb{E}\big[\eta(\tilde{m}_k, \tilde{s}_k^2)\big]$.

Using conjugacy to derive $q^*(\mu_k)$

We just found that $q^*(\mu_k)$ is a normal distribution:

$$q^*(\mu_k) = \mathcal{N}(\mu_k; m_k,\, s_k^2)$$

We still have not yet obtained an expression for its *variational parameters* m_k and s_k^2.

Note: These are different than the *parameters of the complete conditional* $\tilde{m}_k$ and $\tilde{s}_k^2$.

But they are similarly named for a reason...

...they are defined by the inverse mapping of the *variational natural parameter*

$$\begin{bmatrix} m_k \\ s_k^2 \end{bmatrix} = \eta^{-1}\left(\underbrace{\mathbb{E}\big[\eta(\tilde{m}_k, \tilde{s}_k^2)\big]}_{\text{variational natural parameter}} \right)$$

Let's derive the variational natural parameter...

Using conjugacy to derive $q^*(\mu_k)$

First recall the parameters of the complete conditional:

$$\tilde{m}_k = \frac{\sum_{i=1}^n c_{ik}\, x_i}{1/\sigma^2 + \sum_{i=1}^n c_{ik}}, \qquad \tilde{s}_k^2 = \frac{1}{1/\sigma^2 + \sum_{i=1}^n c_{ik}}$$

Apply the natural parameter mapping of a normal:

$$\eta(\tilde{m}_k, \tilde{s}_k^2) = \begin{bmatrix} \tilde{m}_k/\tilde{s}_k^2 \\ -1/2\tilde{s}_k^2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n c_{ik}\, x_i \\ -\frac{1}{2}\left(1/\sigma_0^2 + \sum_{i=1}^n c_{ik}\right) \end{bmatrix}$$

These are just linear functions of c_{ik}, so when we now apply $\mathbb{E}[\,\cdot\,]$, it distributes in

$$\mathbb{E}\left[\eta(\tilde{m}_k, \tilde{s}_k^2)\right] = \begin{bmatrix} \sum_{i=1}^n \mathbb{E}[c_{ik}]\, x_i \\ -\frac{1}{2}\left(1/\sigma_0^2 + \sum_{i=1}^n \mathbb{E}[c_{ik}]\right) \end{bmatrix}$$

Finally, map to the variational (canonical) parameter $\begin{bmatrix} m_k \\ s_k^2 \end{bmatrix} = \eta^{-1}\left(\mathbb{E}\left[\eta(\tilde{m}_k, \tilde{s}_k^2)\right]\right)$:

$$\tilde{m}_k = \frac{\sum_{i=1}^n \mathbb{E}[c_{ik}]\, x_i}{1/\sigma^2 + \sum_{i=1}^n \mathbb{E}[c_{ik}]}, \qquad \tilde{s}_k^2 = \frac{1}{1/\sigma^2 + \sum_{i=1}^n \mathbb{E}[c_{ik}]}$$

CAVI for Bayesian mixture of Gaussians

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 $q(c_i; \varphi_i)$ (K-categorical)

Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s_{1:K}^2$, and $\boldsymbol{\varphi} = \varphi_{1:n}$

while *the* ELBO *has not converged* **do**

  for $i \in \{1, \ldots, n\}$ **do**

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

  end

  for $k \in \{1, \ldots, K\}$ **do**

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

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

  end

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

end

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

Section 4: Variational inference with exponential families

Generic conditionally conjugate models

We introduce the *global* versus *local* distinction.

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

β are the global latent variables

z_i are the local latent variables

This distinction is relative and can start to break down in more complex models.

But it is useful for

- deriving generic rules about conditionals
- deriving stochastic updates that sub-sample the local variables

Generic conditionally conjugate models

Define the (complete data) likelihood as some exponential family:

$$p(z_i, x_i \mid \beta) = h_\ell(z_i, x_i) \exp\left\{\beta^\top t_\ell(z_i, x_i) - A_\ell(\beta)\right\}$$

All exponential families have conjugate priors. The conjugate prior on β is

$$p(\beta \mid \alpha) = h_c(\beta) \exp\left\{\alpha^\top \underbrace{[\beta,\, -A_\ell(\beta)]}_{=t_c(\beta)} - A_c(\alpha)\right\}$$

Because the conjugate prior has that form, it combines nicely with the likelihood

$$\begin{aligned}
p(\beta \mid \boldsymbol{z}, \boldsymbol{x}, \alpha) &\propto_\beta h_c(\beta) \exp\left\{\alpha^\top [\beta,\, -A_\ell(\beta)]\right\} \left[\prod_{i=1}^n \exp\left\{\beta^\top t_\ell(z_i, x_i) - A_\ell(\beta)\right\}\right] \\
&\propto_\beta h_c(\beta) \exp\left\{\beta^\top \underbrace{\left(\alpha_1 + \sum_{i=1}^n t_\ell(z_i, x_i)\right)}_{\triangleq \widehat{\alpha_1}} - \underbrace{(\alpha_2 + n)}_{\triangleq \widehat{\alpha_2}} A_\ell(\beta)\right\} \\
&\propto_\beta h_c(\beta) \exp\left\{\widehat{\alpha}^\top [\beta,\, -A_\ell(\beta)]\right\}
\end{aligned}$$

We recognize this as the kernel of the same exponential family above.

Generic conditionally conjugate models

The complete conditional $p(\beta \mid \boldsymbol{z}, \boldsymbol{x}, \alpha)$ of the global parameter β is in the same (exponential family) as the prior $P(\beta \mid \alpha)$ and has natural parameter $\widehat{\alpha}$:

$$\widehat{\alpha} = \begin{bmatrix} \alpha_1 + \sum_{i=1}^{n} t_\ell(z_i, x_i) \\ \alpha_2 + n \end{bmatrix}$$

The optimal variational update to $q(\beta)$ is then

$$q^*(\beta) \propto_\beta \exp\left\{\mathbb{E}[\log p(\beta \mid \boldsymbol{z}, \boldsymbol{x}, \alpha)]\right\}$$

We already saw this with the normal mixture model.

The optimal $q^*(\beta)$ will always be in the same family as the prior and its *variational natural parameter* λ will always be

$$\lambda = \mathbb{E}[\widehat{\alpha}] = \begin{bmatrix} \alpha_1 + \sum_{i=1}^{n} \mathbb{E}[t_\ell(z_i, x_i)] \\ \alpha_2 + n \end{bmatrix}$$

Generic conditionally conjugate models

We just found that the optimal variational family $q^*(\beta)$ for β is the same exponential family as the prior $q(\beta; \lambda)$ where the variational natural parameter is set to $\lambda = \mathbb{E}[\widehat{\alpha}]$.

We can confirm this by taking the (Euclidean) gradient of the ELBO w.r.t. λ:

$$\nabla_\lambda \mathrm{ELBO} = \nabla^2_\lambda A_c(\lambda)(\mathbb{E}[\widehat{\alpha}] - \lambda)$$

We see that the gradient is 0 when $\lambda = \mathbb{E}[\widehat{\alpha}]$.

(Note that $\nabla^2_\lambda A_c(\lambda)$ is the Hessian of the log normalizer.)

Stochastic variational inference

Gradient-based methods are amenable to sub-sampling for large data sets.

At a high level, if our gradient involved only a linear function of data points

$$\nabla_\lambda \text{ELBO} \propto_{\boldsymbol{x}} \sum_{i=1}^{n} t_\ell(x_i, z_i) \tag{19}$$

then we could sub-sample $i \sim \text{Uniform}(1, n)$ to obtain a noisy but unbiased gradient

$$\widehat{\nabla_\lambda \text{ELBO}} \propto_{\boldsymbol{x}} n\, t_\ell(x_i, z_i) \tag{20}$$

Alas, the (Euclidean) gradient does not have such structure in our setting...

...however Hoffman et al. [2013] showed that the *natural* gradient $g(\lambda)$ does!

$$g(\lambda) = \mathbb{E}[\widehat{\alpha}] - \lambda$$

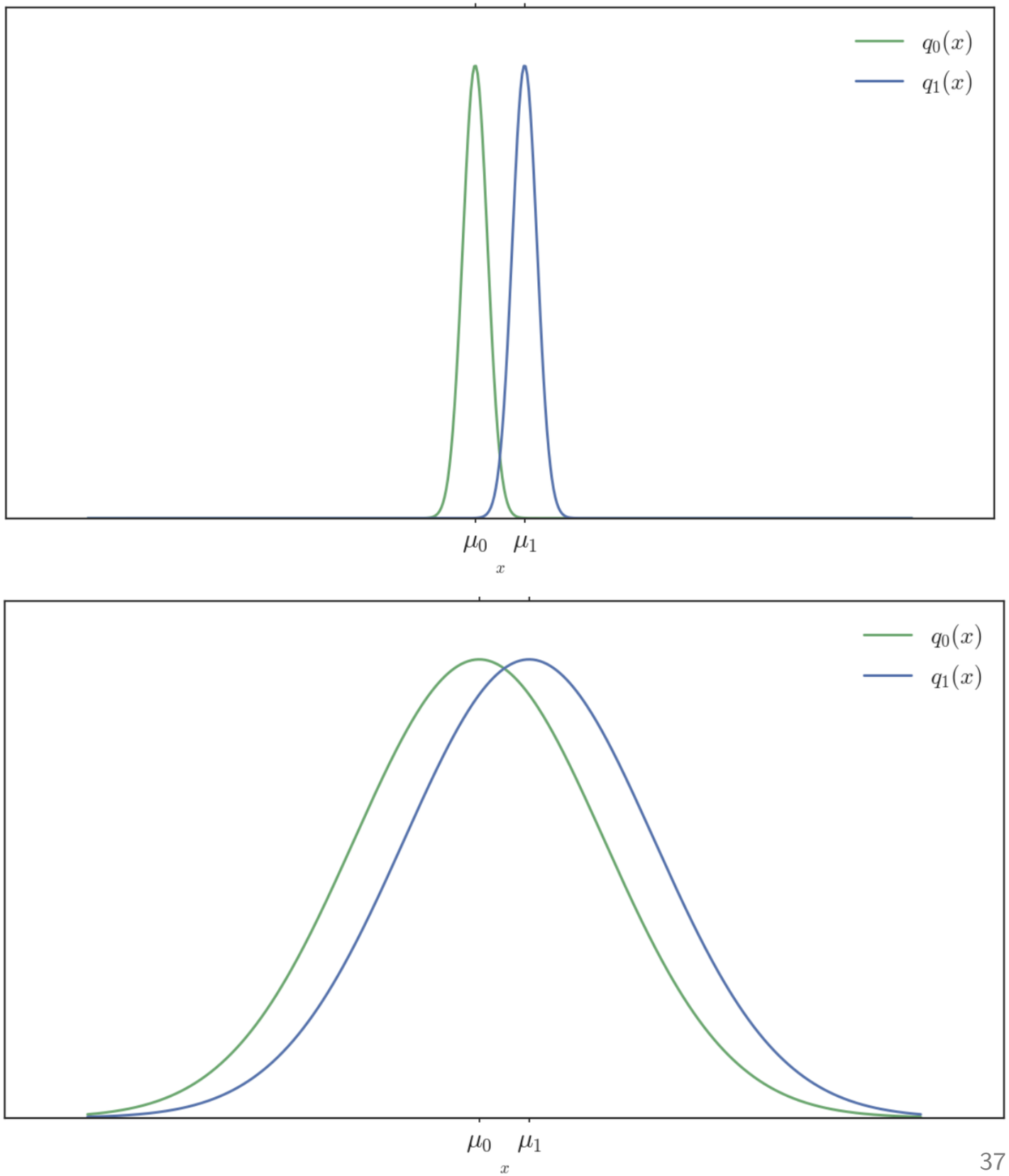
The relationship between the two gradients is

$$\nabla_\lambda \text{ELBO} = \nabla^2_\lambda A_c(\lambda)\, g(\lambda)$$

Natural gradients

Euclidean distance between parameters $\neq$ statistical distance between distributions.

(Figure from Andy Miller.)



The natural gradient is defined by a local rescaling of the Euclidean gradient which accounts for statistical distance...

Natural gradients

Amari [1998] showed that the natural gradient obtains by rescaling the Euclidean gradient by pre-multiplying with the inverse of Fisher information matrix...

$$g(\lambda) = F(\lambda)^{-1} \nabla_\lambda \mathrm{ELBO}(\lambda)$$

...where the Fisher is

$$F(\lambda) = \mathbb{E}\left[\big(\nabla_\lambda \log q(\beta;\lambda)\big) \big(\nabla_\lambda \log q(\beta;\lambda)\big)^\top \right]$$

It so happens that for exponential families, the Fisher equals the Hessian of the log normalizer... so the two matrices cancel

$$\begin{aligned} g(\lambda) &= F(\lambda)^{-1} \nabla_\lambda \mathrm{ELBO}(\lambda) \\ &= F(\lambda)^{-1} \; \nabla^2_\lambda A_c(\lambda) \; (\mathbb{E}[\widehat{\alpha}] - \lambda) \\ &= \mathbb{E}[\widehat{\alpha}] - \lambda \end{aligned}$$

Not only is the natural gradient the "right" thing to follow, it also happens to have a simple form that makes stochastic optimization easy—this is the main observation of Hoffman et al. [2013].

Stochastic variational inference

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 was 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

(It is very easy convert conditionally conjugate CAVI to SVI.)

Looking ahead

Brave new world

RIP "classical" VI (1999ish–2013)

- VI with conditionally conjugate / exponential family models is beautiful / deep.
- For a deeper dive: Wainwright and Jordan [2007].
- Natural gradients is also a beautiful idea [Hoffman et al., 2013].
- Starting next week, we leave this world in the past.

Brave new black box world (2014–present):

- Tues (Jimmy): "Black box VI" [Ranganath et al., 2014]
- Thu (Austin): "Neural VI and learning in belief networks" [Mnih and Gregor, 2014]
- **Idea:** If we can sample from $\boldsymbol{z} \sim q(\boldsymbol{z})$, we can compute "score function" gradient estimates of the ELBO.
  - complex (non-conjugate) p (e.g., parameterized by a NN)
  - complex (non-factorized) q (e.g., parameterized by a NN)
  - generic inference using automatic differentiation
  - scalable inference via stochastic optimization

References I

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10, 1998.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.

David M Blei. Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. *Annual Review of Statistics and Its Application*, 1(1):203–232, January 2014. ISSN 2326-8298, 2326-831X. doi: 10.1146/annurev-statistics-022513-115657.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1285773.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Andriy Mnih and Karol Gregor. Neural Variational Inference and Learning in Belief Networks. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Rajesh Ranganath, Sean Gerrish, and David M Blei. Black Box Variational Inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 2014.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2007. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000001.