Approximate Inference: Variational Bayes Inference (1)

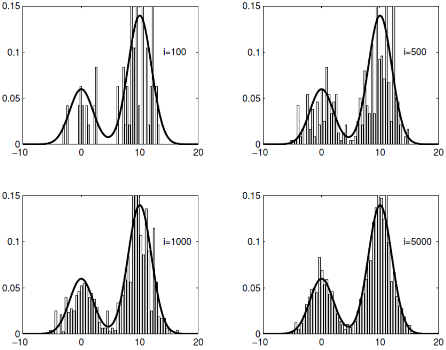
Piyush Rai

Probabilistic Machine Learning (CS772A)

Oct 10, 2017

Sampling based Methods

- Approximate a distribution by a set of randomly drawn samples

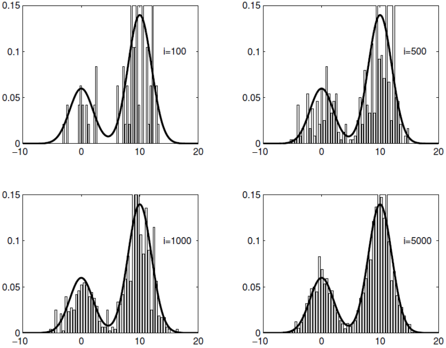


Target distribution and histogram of the MCMC samples at different iteration points.

Picture courtesy: An Introduction to MCMC for Machine Learning (Andrieu et al, 2003)

Sampling based Methods

- Approximate a distribution by a set of randomly drawn samples



Target distribution and histogram of the MCMC samples at different iteration points.

- Compute quantities that depend on this distribution by averaging using these samples

$$p(y_*|\boldsymbol{x}_*) = \mathbb{E}_{p(\boldsymbol{w}|\mathbf{X},\boldsymbol{y})}[p(y_*|\boldsymbol{x}_*, \boldsymbol{w})] = \int p(y_*|\boldsymbol{x}_*, \boldsymbol{w})p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})d\boldsymbol{w} \approx \frac{1}{L}\sum_{\ell=1}^{L} p(y_*|\boldsymbol{x}_*, \boldsymbol{w}^{(\ell)})$$

Picture courtesy: An Introduction to MCMC for Machine Learning (Andrieu et al, 2003)

Sampling based Methods

- Pros:
    - Asymptotically exact

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)
  - Often easy to implement (e.g., when doing Gibbs sampling)

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)
  - Often easy to implement (e.g., when doing Gibbs sampling)
  - Very general. Can be used to sample from a broad class of distributions

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)
  - Often easy to implement (e.g., when doing Gibbs sampling)
  - Very general. Can be used to sample from a broad class of distributions
- Cons:

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)
  - Often easy to implement (e.g., when doing Gibbs sampling)
  - Very general. Can be used to sample from a broad class of distributions
- Cons:
  - Can be expensive

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)
  - Often easy to implement (e.g., when doing Gibbs sampling)
  - Very general. Can be used to sample from a broad class of distributions
- Cons:
  - Can be expensive (but a lot of work on speeding up; online and parallel MCMC methods)

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)
  - Often easy to implement (e.g., when doing Gibbs sampling)
  - Very general. Can be used to sample from a broad class of distributions
- Cons:
  - Can be expensive (but a lot of work on speeding up; online and parallel MCMC methods)
  - Convergence can be hard to assess

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)
  - Often easy to implement (e.g., when doing Gibbs sampling)
  - Very general. Can be used to sample from a broad class of distributions
- Cons:
  - Can be expensive (but a lot of work on speeding up; online and parallel MCMC methods)
  - Convergence can be hard to assess (though there exist heuristics to diagnose convergence, and formal results on convergence rate to to the true posterior)

Sampling based Methods

- Pros:
  - Asymptotically exact (.. if you run the sampler long enough)
  - Often easy to implement (e.g., when doing Gibbs sampling)
  - Very general. Can be used to sample from a broad class of distributions
- Cons:
  - Can be expensive (but a lot of work on speeding up; online and parallel MCMC methods)
  - Convergence can be hard to assess (though there exist heuristics to diagnose convergence, and formal results on convergence rate to to the true posterior)
  - Expensive storage-wise (need to store samples) and also at prediction-time (e.g., we need to average using the collected samples)

Variational Bayes (VB) or Variational Inference (VI)

- Origin of the name: Calculus of variations (optimizing w.r.t. functions/distributions)

Picture courtesy: Modified from the original pic by Shakir Mohamed

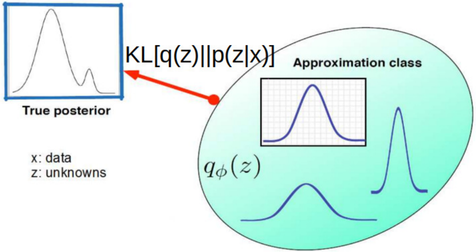
Variational Bayes (VB) or Variational Inference (VI)

- Origin of the name: Calculus of variations (optimizing w.r.t. functions/distributions)
- Assume a approximation class of distributions $\{q(\boldsymbol{z}|\phi)\}$ parameterized by free parameters ϕ

Picture courtesy: Modified from the original pic by Shakir Mohamed

Variational Bayes (VB) or Variational Inference (VI)

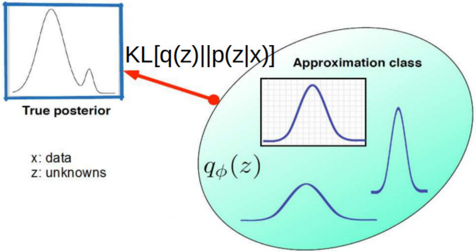
- Origin of the name: Calculus of variations (optimizing w.r.t. functions/distributions)
- Assume a approximation class of distributions $\{q(\boldsymbol{z}|\phi)\}$ parameterized by free parameters ϕ
- Approximate the true distribution $p(\boldsymbol{z}|\boldsymbol{x})$ by finding the "closest" $q(\boldsymbol{z}|\phi)$ from this class



Picture courtesy: Modified from the original pic by Shakir Mohamed

Variational Bayes (VB) or Variational Inference (VI)

- Origin of the name: Calculus of variations (optimizing w.r.t. functions/distributions)
- Assume a approximation class of distributions $\{q(\boldsymbol{z}|\phi)\}$ parameterized by free parameters ϕ
- Approximate the true distribution $p(\boldsymbol{z}|\boldsymbol{x})$ by finding the "closest" $q(\boldsymbol{z}|\phi)$ from this class

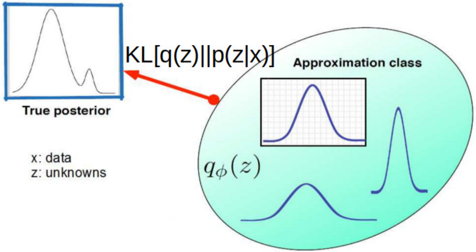


- ϕ (variational parameters) is like a "knob" that we have to tune to find the best matching $q(\boldsymbol{z}|\phi)$

Picture courtesy: Modified from the original pic by Shakir Mohamed

Variational Bayes (VB) or Variational Inference (VI)

- Origin of the name: Calculus of variations (optimizing w.r.t. functions/distributions)
- Assume a approximation class of distributions $\{q(\boldsymbol{z}|\phi)\}$ parameterized by free parameters ϕ
- Approximate the true distribution $p(\boldsymbol{z}|\boldsymbol{x})$ by finding the "closest" $q(\boldsymbol{z}|\phi)$ from this class

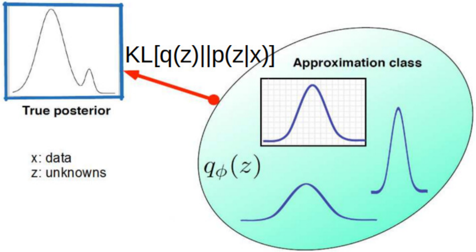


- ϕ (variational parameters) is like a "knob" that we have to tune to find the best matching $q(\boldsymbol{z}|\phi)$
- $\phi^* = \arg\min_\phi \text{KL}[q_\phi(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})]$

Picture courtesy: Modified from the original pic by Shakir Mohamed

Variational Bayes (VB) or Variational Inference (VI)

- Origin of the name: Calculus of variations (optimizing w.r.t. functions/distributions)
- Assume a approximation class of distributions $\{q(\boldsymbol{z}|\phi)\}$ parameterized by free parameters ϕ
- Approximate the true distribution $p(\boldsymbol{z}|\boldsymbol{x})$ by finding the "closest" $q(\boldsymbol{z}|\phi)$ from this class

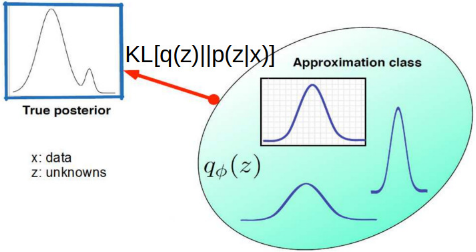


- ϕ (variational parameters) is like a "knob" that we have to tune to find the best matching $q(\boldsymbol{z}|\phi)$
- $\phi^* = \arg\min_\phi \text{KL}[q_\phi(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})]$: Approximate inference now becomes an optimization problem!

Picture courtesy: Modified from the original pic by Shakir Mohamed

Variational Bayes (VB) or Variational Inference (VI)

- Origin of the name: Calculus of variations (optimizing w.r.t. functions/distributions)
- Assume a approximation class of distributions $\{q(\boldsymbol{z}|\phi)\}$ parameterized by free parameters ϕ
- Approximate the true distribution $p(\boldsymbol{z}|\boldsymbol{x})$ by finding the "closest" $q(\boldsymbol{z}|\phi)$ from this class



- ϕ (variational parameters) is like a "knob" that we have to tune to find the best matching $q(\boldsymbol{z}|\phi)$
- $\phi^* = \arg\min_\phi \text{KL}[q_\phi(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})]$: Approximate inference now becomes an optimization problem!
- **But wait!** We don't know the true distribution $p(\boldsymbol{z}|\boldsymbol{x})$. How to solve the above problem then?

Picture courtesy: Modified from the original pic by Shakir Mohamed

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \\ \mathrm{KL}(q\|p) &= -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}\end{aligned}$$

- Note that we saw something similar in EM

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that we saw something similar in EM. But, unlike EM, there's no "parameters" Θ vs latent variables $\mathbf{Z}$ distinction. Now we'll treat everything as latent variables and collectively call it $\mathbf{Z}$

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that we saw something similar in EM. But, unlike EM, there's no "parameters" Θ vs latent variables $\mathbf{Z}$ distinction. Now we'll treat everything as latent variables and collectively call it $\mathbf{Z}$
- **An Observation:** The L.H.S. $\log p(\mathbf{X})$ is a constant w.r.t. $\mathbf{Z}$

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = - \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that we saw something similar in EM. But, unlike EM, there's no "parameters" Θ vs latent variables $\mathbf{Z}$ distinction. Now we'll treat everything as latent variables and collectively call it $\mathbf{Z}$
- **An Observation:** The L.H.S. $\log p(\mathbf{X})$ is a constant w.r.t. $\mathbf{Z}$
- Thus finding q that minimizes $KL(q||p)$ is equivalent to finding q that maximizes $\mathcal{L}(q)$

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that we saw something similar in EM. But, unlike EM, there's no "parameters" Θ vs latent variables $\mathbf{Z}$ distinction. Now we'll treat everything as latent variables and collectively call it $\mathbf{Z}$
- **An Observation:** The L.H.S. $\log p(\mathbf{X})$ is a constant w.r.t. $\mathbf{Z}$
- Thus finding q that minimizes $KL(q||p)$ is equivalent to finding q that maximizes $\mathcal{L}(q)$
  - Important: Unlike $KL(q||p)$, $\mathcal{L}(q)$ does NOT depend on the true posterior $p(\mathbf{Z}|\mathbf{X})$

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that we saw something similar in EM. But, unlike EM, there's no "parameters" Θ vs latent variables $\mathbf{Z}$ distinction. Now we'll treat everything as latent variables and collectively call it $\mathbf{Z}$
- **An Observation:** The L.H.S. $\log p(\mathbf{X})$ is a constant w.r.t. $\mathbf{Z}$
- Thus finding q that minimizes $KL(q||p)$ is equivalent to finding q that maximizes $\mathcal{L}(q)$
  - Important: Unlike $KL(q||p)$, $\mathcal{L}(q)$ does NOT depend on the true posterior $p(\mathbf{Z}|\mathbf{X})$
- Note: Since $KL \geq 0$, $\mathcal{L}(q)$ is a lower bound on the log-evidence $\log p(\mathbf{X})$ of the model m

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that we saw something similar in EM. But, unlike EM, there's no "parameters" Θ vs latent variables $\mathbf{Z}$ distinction. Now we'll treat everything as latent variables and collectively call it $\mathbf{Z}$
- **An Observation:** The L.H.S. $\log p(\mathbf{X})$ is a constant w.r.t. $\mathbf{Z}$
- Thus finding q that minimizes $KL(q||p)$ is equivalent to finding q that maximizes $\mathcal{L}(q)$
  - Important: Unlike $KL(q||p)$, $\mathcal{L}(q)$ does <u>NOT</u> depend on the true posterior $p(\mathbf{Z}|\mathbf{X})$
- Note: Since $KL \geq 0$, $\mathcal{L}(q)$ is a lower bound on the log-evidence $\log p(\mathbf{X})$ of the model m

$$\log p(\mathbf{X}|m) \geq \mathcal{L}(q) \qquad \text{(using } m \text{ to explicitly refer to the model in consideration)}$$

Variational Bayes (VB) Inference

- The key identity central to VB inference is the following (holds for any choice of q)

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that we saw something similar in EM. But, unlike EM, there's no "parameters" Θ vs latent variables $\mathbf{Z}$ distinction. Now we'll treat everything as latent variables and collectively call it $\mathbf{Z}$
- **An Observation:** The L.H.S. $\log p(\mathbf{X})$ is a constant w.r.t. $\mathbf{Z}$
- Thus finding q that minimizes $KL(q||p)$ is equivalent to finding q that maximizes $\mathcal{L}(q)$
  - Important: Unlike $KL(q||p)$, $\mathcal{L}(q)$ does NOT depend on the true posterior $p(\mathbf{Z}|\mathbf{X})$
- Note: Since $KL \geq 0$, $\mathcal{L}(q)$ is a lower bound on the log-evidence $\log p(\mathbf{X})$ of the model m

$$\log p(\mathbf{X}|m) \geq \mathcal{L}(q) \qquad \text{(using } m \text{ to explicitly refer to the model in consideration)}$$

- Therefore $\mathcal{L}(q)$ is also known as the **Evidence Lower Bound (ELBO)**

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

- Note that q depends on the variational parameters ϕ. Expanding, we get

$$\mathcal{L}(q) = \mathcal{L}(\phi) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})]$$

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that q depends on the variational parameters ϕ. Expanding, we get

$$\begin{aligned} \mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z})) \end{aligned}$$

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

- Note that q depends on the variational parameters ϕ. Expanding, we get

$$\begin{aligned} \mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \text{KL}(q(\mathbf{Z})||p(\mathbf{Z})) \end{aligned}$$

- Makes sense: Maximizing $\mathcal{L}(q)$ will give a q that explains data well and is close to the prior

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

- Note that q depends on the variational parameters ϕ. Expanding, we get

$$\begin{aligned} \mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z})) \end{aligned}$$

- Makes sense: Maximizing $\mathcal{L}(q)$ will give a q that explains data well and is close to the prior
- Maximizing $\mathcal{L}(q)$ w.r.t. q can still be hard in general (note the expectation w.r.t. q)

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

- Note that q depends on the variational parameters ϕ. Expanding, we get

$$\begin{aligned} \mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z})) \end{aligned}$$

- Makes sense: Maximizing $\mathcal{L}(q)$ will give a q that explains data well and is close to the prior
- Maximizing $\mathcal{L}(q)$ w.r.t. q can still be hard in general (note the expectation w.r.t. q)
- Some of the ways to make this problem easier

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

- Note that q depends on the variational parameters ϕ. Expanding, we get

$$\begin{aligned} \mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z})) \end{aligned}$$

- Makes sense: Maximizing $\mathcal{L}(q)$ will give a q that explains data well and is close to the prior
- Maximizing $\mathcal{L}(q)$ w.r.t. q can still be hard in general (note the expectation w.r.t. q)
- Some of the ways to make this problem easier
  1. Restricting the form of the q distribution, e.g., mean-field VB inference (today's discussion)

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) \quad = \quad \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- Note that q depends on the variational parameters ϕ. Expanding, we get

$$\begin{aligned} \mathcal{L}(q) = \mathcal{L}(\phi) \quad &= \quad \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \quad \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z})) \end{aligned}$$

- Makes sense: Maximizing $\mathcal{L}(q)$ will give a q that explains data well and is close to the prior
- Maximizing $\mathcal{L}(q)$ w.r.t. q can still be hard in general (note the expectation w.r.t. q)
- Some of the ways to make this problem easier
  1. Restricting the form of the q distribution, e.g., mean-field VB inference (today's discussion)
  2. Using Monte-Carlo approximation of the expectation/gradient of the ELBO (later)

VB by Maximizing the ELBO

- VB finds the q distribution that maximizes the ELBO

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

- Note that q depends on the variational parameters ϕ. Expanding, we get

$$\begin{aligned} \mathcal{L}(q) = \mathcal{L}(\phi) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z})) \end{aligned}$$

- Makes sense: Maximizing $\mathcal{L}(q)$ will give a q that explains data well and is close to the prior
- Maximizing $\mathcal{L}(q)$ w.r.t. q can still be hard in general (note the expectation w.r.t. q)
- Some of the ways to make this problem easier
  1. Restricting the form of the q distribution, e.g., mean-field VB inference (today's discussion)
  2. Using Monte-Carlo approximation of the expectation/gradient of the ELBO (later)
- Option (1) becomes especially easy if $p(\mathbf{X}|\mathbf{Z})$ and $p(\mathbf{Z})$ are exponential family distributions or if the model is locally conjugate

Mean-Field VB Inference

- Suppose we partition the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$

Mean-Field VB Inference

- Suppose we partition the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

$$q(\mathbf{Z}|\phi) = \prod_{i=1}^{M} q(\mathbf{Z}_i|\phi_i)$$

Mean-Field VB Inference

- Suppose we partition the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

$$q(\mathbf{Z}|\phi) = \prod_{i=1}^{M} q(\mathbf{Z}_i|\phi_i)$$

- As a short-hand, sometimes we write $q = \prod_{i=1}^{M} q_i$ where $q_i = q(\mathbf{Z}_i|\phi_i)$

Mean-Field VB Inference

- Suppose we partition the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

$$q(\mathbf{Z}|\phi) = \prod_{i=1}^{M} q(\mathbf{Z}_i|\phi_i)$$

- As a short-hand, sometimes we write $q = \prod_{i=1}^{M} q_i$ where $q_i = q(\mathbf{Z}_i|\phi_i)$
- Mean-field assumption is quite a strong assumption (can destroy structure among latent variables)

Mean-Field VB Inference

- Suppose we partition the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

$$q(\mathbf{Z}|\phi) = \prod_{i=1}^{M} q(\mathbf{Z}_i|\phi_i)$$

- As a short-hand, sometimes we write $q = \prod_{i=1}^{M} q_i$ where $q_i = q(\mathbf{Z}_i|\phi_i)$
- Mean-field assumption is quite a strong assumption (can destroy structure among latent variables)
- Many improvements have been proposed to the standard mean-field VB

Mean-Field VB Inference

- Suppose we partition the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

$$q(\mathbf{Z}|\phi) = \prod_{i=1}^{M} q(\mathbf{Z}_i|\phi_i)$$

- As a short-hand, sometimes we write $q = \prod_{i=1}^{M} q_i$ where $q_i = q(\mathbf{Z}_i|\phi_i)$
- Mean-field assumption is quite a strong assumption (can destroy structure among latent variables)
- Many improvements have been proposed to the standard mean-field VB, e.g.,
    - Structured Mean-Field (Saul and Jordan 1995). Imposes a pre-defined structure in the decomposition

Mean-Field VB Inference

- Suppose we partition the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

$$q(\mathbf{Z}|\phi) = \prod_{i=1}^{M} q(\mathbf{Z}_i|\phi_i)$$

- As a short-hand, sometimes we write $q = \prod_{i=1}^{M} q_i$ where $q_i = q(\mathbf{Z}_i|\phi_i)$
- Mean-field assumption is quite a strong assumption (can destroy structure among latent variables)
- Many improvements have been proposed to the standard mean-field VB, e.g.,
    - Structured Mean-Field (Saul and Jordan 1995). Imposes a pre-defined structure in the decomposition
    - Hierarchical Variational Models (Ranganath et al, 2016): Assumes a shared prior on $\phi_1, \ldots, \phi_M$ which imposes a coupling among them

Mean-Field VB Inference

- Suppose we partition the latent variables $\mathbf{Z}$ into M groups $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$
- Assume our approximation $q(\mathbf{Z})$ factorizes over these groups

$$q(\mathbf{Z}|\phi) = \prod_{i=1}^{M} q(\mathbf{Z}_i|\phi_i)$$

- As a short-hand, sometimes we write $q = \prod_{i=1}^{M} q_i$ where $q_i = q(\mathbf{Z}_i|\phi_i)$
- Mean-field assumption is quite a strong assumption (can destroy structure among latent variables)
- Many improvements have been proposed to the standard mean-field VB, e.g.,
  - Structured Mean-Field (Saul and Jordan 1995). Imposes a pre-defined structure in the decomposition
  - Hierarchical Variational Models (Ranganath et al, 2016): Assumes a shared prior on $\phi_1, \ldots, \phi_M$ which imposes a coupling among them
- In mean-field VB, learning the optimal q reduces to learning the optimal $q_1, \ldots, q_M$.

Mean-Field VB Inference: Method 1 (Based on Inspection)

- Under the mean-field assumption, the ELBO simplifies to

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \\
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z}
\end{aligned}
$$

Mean-Field VB Inference: Method 1 (Based on Inspection)

- Under the mean-field assumption, the ELBO simplifies to

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\begin{aligned} \mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\ &= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \end{aligned}$$

Mean-Field VB Inference: Method 1 (Based on Inspection)

- Under the mean-field assumption, the ELBO simplifies to

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} d\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, d\mathbf{Z}_i \right\} d\mathbf{Z}_j - \int q_j \ln q_j \, d\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, d\mathbf{Z}_j - \int q_j \ln q_j \, d\mathbf{Z}_j + \text{const}
\end{aligned}$$

Mean-Field VB Inference: Method 1 (Based on Inspection)

- Under the mean-field assumption, the ELBO simplifies to

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \\
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}
\end{aligned}
$$

where we have defined a new distribution $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ by the relation

$$
\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}.
$$

Mean-Field VB Inference: Method 1 (Based on Inspection)

- Under the mean-field assumption, the ELBO simplifies to

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X},\mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X},\mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X},\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}
\end{aligned}$$

where we have defined a new distribution $\widetilde{p}(\mathbf{X},\mathbf{Z}_j)$ by the relation

$$\ln \widetilde{p}(\mathbf{X},\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const}.$$

- Here $\mathbb{E}_{i \neq j}$ denotes expectation w.r.t. the q distribution except component q_j

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X},\mathbf{Z})] = \int \ln p(\mathbf{X},\mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i$$

Mean-Field VB Inference: Method 1 (Based on Inspection)

- Under the mean-field assumption, the ELBO simplifies to

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}
\end{aligned}$$

where we have defined a new distribution $\widetilde{p}(\mathbf{X}, \mathbf{Z}_j)$ by the relation

$$\ln \widetilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}.$$

- Here $\mathbb{E}_{i \neq j}$ denotes expectation w.r.t. the q distribution except component q_j

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i$$

- In the above, $\mathcal{L}(q) = -KL(q_j || \tilde{p}) + \text{const}$. Which q_j will maximize it?

Mean-Field VB Inference: Method 1 (Based on Inspection)

- Under the mean-field assumption, the ELBO simplifies to

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \\
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{X},\mathbf{Z}) - \sum_i \ln q_i \right\} \mathrm{d}\mathbf{Z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{X},\mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i \right\} \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const} \\
&= \int q_j \ln \widetilde{p}(\mathbf{X},\mathbf{Z}_j) \, \mathrm{d}\mathbf{Z}_j - \int q_j \ln q_j \, \mathrm{d}\mathbf{Z}_j + \text{const}
\end{aligned}$$

where we have defined a new distribution $\widetilde{p}(\mathbf{X},\mathbf{Z}_j)$ by the relation

$$\ln \widetilde{p}(\mathbf{X},\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const}.$$

- Here $\mathbb{E}_{i \neq j}$ denotes expectation w.r.t. the q distribution except component q_j

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X},\mathbf{Z})] = \int \ln p(\mathbf{X},\mathbf{Z}) \prod_{i \neq j} q_i \, \mathrm{d}\mathbf{Z}_i$$

- In the above, $\mathcal{L}(q) = -KL(q_j || \tilde{p}) + \text{const}$. Which q_j will maximize it? Answer: $q_j = \tilde{p}(\mathbf{X},\mathbf{Z}_j)$

Mean-Field VB Inference: Method 1 (Based on Inspection)

- The optimal $q_j^*(\mathbf{Z}_j)$ is therefore equal to $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$

Mean-Field VB Inference: Method 1 (Based on Inspection)

- The optimal $q_j^*(\mathbf{Z}_j)$ is therefore equal to $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$
- Since $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, we have

Mean-Field VB Inference: Method 1 (Based on Inspection)

- The optimal $q_j^*(\mathbf{Z}_j)$ is therefore equal to $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$
- Since $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, we have

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \, \mathrm{d}\mathbf{Z}_j}$$

- Note: Only need to compute the numerator. Denominator by usually be recognized by inspection

Mean-Field VB Inference: Method 1 (Based on Inspection)

- The optimal $q_j^*(\mathbf{Z}_j)$ is therefore equal to $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$
- Since $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, we have

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \, \mathrm{d}\mathbf{Z}_j}$$

- Note: Only need to compute the numerator. Denominator by usually be recognized by inspection
- Note: For estimating q_j, the required expectation depends on other $\{q_i\}_{i\neq j}$

Mean-Field VB Inference: Method 1 (Based on Inspection)

- The optimal $q_j^*(\mathbf{Z}_j)$ is therefore equal to $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$
- Since $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, we have

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \, \mathrm{d}\mathbf{Z}_j}$$

- Note: Only need to compute the numerator. Denominator by usually be recognized by inspection
- Note: For estimating q_j, the required expectation depends on other $\{q_i\}_{i \neq j}$
- Thus we need to cycle through updating each q_j in turn (similar to co-ordinate ascent, alternating optimization, Gibbs sampling, etc.)

Mean-Field VB Inference: Method 1 (Based on Inspection)

- The optimal $q_j^*(\mathbf{Z}_j)$ is therefore equal to $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$
- Since $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, we have

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \, \mathrm{d}\mathbf{Z}_j}$$

- Note: Only need to compute the numerator. Denominator by usually be recognized by inspection
- Note: For estimating q_j, the required expectation depends on other $\{q_i\}_{i \neq j}$
- Thus we need to cycle through updating each q_j in turn (similar to co-ordinate ascent, alternating optimization, Gibbs sampling, etc.)
- Guaranteed to converge (to a local optima). $\mathcal{L}(q)$ is concave w.r.t. each q_j (and we're maximizing)

Mean-Field VB Inference: Method 1 (Based on Inspection)

- The optimal $q_j^*(\mathbf{Z}_j)$ is therefore equal to $\tilde{p}(\mathbf{X}, \mathbf{Z}_j)$
- Since $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, we have

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)\, \mathrm{d}\mathbf{Z}_j}$$

- Note: Only need to compute the numerator. Denominator by usually be recognized by inspection
- Note: For estimating q_j, the required expectation depends on other $\{q_i\}_{i \neq j}$
- Thus we need to cycle through updating each q_j in turn (similar to co-ordinate ascent, alternating optimization, Gibbs sampling, etc.)
- Guaranteed to converge (to a local optima). $\mathcal{L}(q)$ is concave w.r.t. each q_j (and we're maximizing)
- Note: We didn't specify any particular form of q_j to begin with

An Example of Mean-Field VB via Inspection Method

- Suppose we have N obs. $\mathcal{D} = \{x_1, \ldots, x_N\}$ from a 1-D Gaussian $\mathcal{N}(\mu, \tau)$

$$p(\mathcal{D}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- Assume the following priors on the mean μ and precision τ

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}\left(\mu|\mu_0, (\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}(\tau|a_0, b_0) \end{aligned}$$

- Goal: Infer the posterior distribution $p(\mu, \tau|\mathcal{D})$

An Example of Mean-Field VB via Inspection Method

- Suppose we have N obs. $\mathcal{D} = \{x_1, \dots, x_N\}$ from a 1-D Gaussian $\mathcal{N}(\mu, \tau)$

$$p(\mathcal{D}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- Assume the following priors on the mean μ and precision τ

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}\left(\mu|\mu_0,(\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}(\tau|a_0,b_0) \end{aligned}$$

- Goal: Infer the posterior distribution $p(\mu, \tau|\mathcal{D})$
- We know how to do it using other simpler ways but let's try VB on it

An Example of Mean-Field VB via Inspection Method

- Suppose we have N obs. $\mathcal{D} = \{x_1, \ldots, x_N\}$ from a 1-D Gaussian $\mathcal{N}(\mu, \tau)$

$$p(\mathcal{D}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

- Assume the following priors on the mean μ and precision τ

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}\left(\mu|\mu_0, (\lambda_0\tau)^{-1}\right) \\ p(\tau) &= \mathrm{Gam}(\tau|a_0, b_0) \end{aligned}$$

- Goal: Infer the posterior distribution $p(\mu, \tau|\mathcal{D})$
- We know how to do it using other simpler ways but let's try VB on it
- Assume the following factorized distribution

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$$

An Example of Mean-Field VB Inference (via Inspection Method)

- Using the inspection method, we have $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, and

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \, \mathrm{d}\mathbf{Z}_j}$$

An Example of Mean-Field VB Inference (via Inspection Method)

- Using the inspection method, we have $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, and

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \, \mathrm{d}\mathbf{Z}_j}$$

- Variational distribution for the Gaussian's mean:

$$\begin{aligned} \ln q_\mu^\star(\mu) &= \mathbb{E}_\tau\left[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)\right] + \text{const} \\ &= -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu - \mu_0)^2 + \sum_{n=1}^N (x_n - \mu)^2\right\} + \text{const.} \end{aligned}$$

Completing the square over μ we see that $q_\mu(\mu)$ is a Gaussian $\mathcal{N}\left(\mu|\mu_N, \lambda_N^{-1}\right)$ with mean and precision given by

$$\begin{aligned} \mu_N &= \frac{\lambda_0\mu_0 + N\overline{x}}{\lambda_0 + N} \\ \lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau]. \end{aligned}$$

An Example of Mean-Field VB Inference (via Inspection Method)

- Using the inspection method, we have $\ln q_j^*(\mathbf{Z}_j) = \ln \tilde{p}(\mathbf{X}, \mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + \text{const}$, and

$$q_j^\star(\mathbf{Z}_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]\right) \mathrm{d}\mathbf{Z}_j}$$

- Variational distribution for the Gaussian's mean:

$$\begin{aligned}\ln q_\mu^\star(\mu) &= \mathbb{E}_\tau\left[\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)\right] + \text{const} \\ &= -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu - \mu_0)^2 + \sum_{n=1}^N (x_n - \mu)^2\right\} + \text{const}.\end{aligned}$$

Completing the square over μ we see that $q_\mu(\mu)$ is a Gaussian $\mathcal{N}\left(\mu|\mu_N, \lambda_N^{-1}\right)$ with mean and precision given by

$$\begin{aligned}\mu_N &= \frac{\lambda_0\mu_0 + N\overline{x}}{\lambda_0 + N} \\ \lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau].\end{aligned}$$

- Note that we didn't assume q_μ to be a Gaussian (we just found out by inspection)

An Example of Mean-Field VB Inference (via Inspection Method)

- Assuming shape-rate parameterization of gamma prior on the precision τ, the variational distribution for the Gaussian's precision (verify):

$$
\begin{aligned}
\ln q_\tau^\star(\tau) &= \mathbb{E}_\mu \left[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)\right] + \ln p(\tau) + \text{const} \\
&= (a_0 - 1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau + \frac{1}{2}\ln\tau \\
&\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const}
\end{aligned}
$$

and hence $q_\tau(\tau)$ is a gamma distribution $\text{Gam}(\tau|a_N, b_N)$ with parameters

$$
\begin{aligned}
a_N &= a_0 + \frac{N+1}{2} \\
b_N &= b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]
\end{aligned}
$$

An Example of Mean-Field VB Inference (via Inspection Method)

- Assuming shape-rate parameterization of gamma prior on the precision τ, the variational distribution for the Gaussian's precision (verify):

$$\begin{aligned}\ln q_\tau^\star(\tau) &= \mathbb{E}_\mu\left[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)\right] + \ln p(\tau) + \text{const} \\ &= (a_0 - 1)\ln\tau - b_0\tau + \frac{N}{2}\ln\tau + \frac{1}{2}\ln\tau \\ &\quad -\frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^N (x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + \text{const}\end{aligned}$$

and hence $q_\tau(\tau)$ is a gamma distribution $\text{Gam}(\tau|a_N, b_N)$ with parameters

$$\begin{aligned}a_N &= a_0 + \frac{N+1}{2} \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^N (x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]\end{aligned}$$

- Note that we didn't assume q_μ to be a Gamma (we just found out by inspection)

The Full Algorithm

1. Initialize the variational distribution parameters $\mu_N, \lambda_N, a_N, b_N$ randomly

The Full Algorithm

1. Initialize the variational distribution parameters $\mu_N, \lambda_N, a_N, b_N$ randomly
2. Update μ_N, λ_N as follows (depends on a_N, b_N)

$$\begin{aligned}
\mu_N &= \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N} \\
\lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau]
\end{aligned}$$

The Full Algorithm

1. Initialize the variational distribution parameters $\mu_N, \lambda_N, a_N, b_N$ randomly
2. Update μ_N, λ_N as follows (depends on a_N, b_N)

$$\begin{aligned} \mu_N &= \frac{\lambda_0 \mu_0 + N\bar{x}}{\lambda_0 + N} \\ \lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau] \end{aligned}$$

3. Update a_N, b_N as follows (depends on μ_N, λ_N)

$$\begin{aligned} a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_\mu \left[\sum_{n=1}^{N} (x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2 \right] \end{aligned}$$

The Full Algorithm

1. Initialize the variational distribution parameters $\mu_N, \lambda_N, a_N, b_N$ randomly
2. Update μ_N, λ_N as follows (depends on a_N, b_N)

$$\begin{aligned}\mu_N &= \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \\ \lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau]\end{aligned}$$

3. Update a_N, b_N as follows (depends on μ_N, λ_N)

$$\begin{aligned}a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]\end{aligned}$$

4. Go to step 2 if not converged

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) \quad = \quad \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})]$$

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$\begin{aligned}\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}\end{aligned}$$

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$\begin{aligned}
\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}
\end{aligned}$$

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$
\begin{aligned}
\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}
\end{aligned}
$$

- Note: The above may get further simplified due to independence structures, e.g.,

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$\begin{aligned}\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}\end{aligned}$$

- Note: The above may get further simplified due to independence structures, e.g.,
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$;

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$\begin{aligned}\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}\end{aligned}$$

- Note: The above may get further simplified due to independence structures, e.g.,
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$;conditionally independent priors simplify $\log p(\mathbf{Z})$

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$\begin{aligned}
\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}
\end{aligned}$$

- Note: The above may get further simplified due to independence structures, e.g.,
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$;conditionally independent priors simplify $\log p(\mathbf{Z})$
  - The mean-field assumption simplifies $q(\mathbf{Z})$ as $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$\begin{aligned}\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}\end{aligned}$$

- Note: The above may get further simplified due to independence structures, e.g.,
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$;conditionally independent priors simplify $\log p(\mathbf{Z})$
  - The mean-field assumption simplifies $q(\mathbf{Z})$ as $q(\mathbf{Z}) = \prod_{i=1}^M q_i(\mathbf{Z}_i)$
  - Note that the last term reduces to sum of entropies of q_i's (which usually has known forms)

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$
\begin{aligned}
\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}
\end{aligned}
$$

- Note: The above may get further simplified due to independence structures, e.g.,
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$;conditionally independent priors simplify $\log p(\mathbf{Z})$
  - The mean-field assumption simplifies $q(\mathbf{Z})$ as $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$
  - Note that the last term reduces to sum of entropies of q_i's (which usually has known forms)
- Now take partial derivatives of $\mathcal{L}(\phi_1, \ldots, \phi_M)$ w.r.t. $\phi_1, \ldots, \phi_M$ to find their optimal values

Mean-Field VB Inference: Method 2 (via ELBO Derivatives)

- Assume some form for each $q_i(\mathbf{Z}_i)$ (e.g., the same distribution as the prior $p(\mathbf{Z}_i)$)
  - Suppose the free variational parameters of $q_i(\mathbf{Z}_i)$ are ϕ_i (say mean and variance of the Gaussian)
- Now write down the **full ELBO** expression (Imp: $\mathbf{Z}_i$'s will go aways since they are integrated out)

$$\begin{aligned}\mathcal{L}(q) = \mathcal{L}(\phi_1, \ldots, \phi_M) &= \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z})] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}\end{aligned}$$

- Note: The above may get further simplified due to independence structures, e.g.,
  - i.i.d. observations simplify $\log p(\mathbf{X}|\mathbf{Z})$;conditionally independent priors simplify $\log p(\mathbf{Z})$
  - The mean-field assumption simplifies $q(\mathbf{Z})$ as $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$
  - Note that the last term reduces to sum of entropies of q_i's (which usually has known forms)
- Now take partial derivatives of $\mathcal{L}(\phi_1, \ldots, \phi_M)$ w.r.t. $\phi_1, \ldots, \phi_M$ to find their optimal values
  - Note: Each ϕ_j usually depends on the other ϕ_i's ($i \neq j$). Co-ordinate ascent/descent like procedure

An Example of Mean-Field VB via ELBO Derivatives

- Let's revisit the Gaussian parameter estimation via explicitly taking derivatives of ELBO
- Suppose $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N)$ and $q_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$
- Note: our variational distribution $q(\mathbf{Z}) = q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$

An Example of Mean-Field VB via ELBO Derivatives

- Let's revisit the Gaussian parameter estimation via explicitly taking derivatives of ELBO
- Suppose $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N)$ and $q_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$
- Note: our variational distribution $q(\mathbf{Z}) = q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$
- Can write the ELBO as (by definition)

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$$

An Example of Mean-Field VB via ELBO Derivatives

- Let's revisit the Gaussian parameter estimation via explicitly taking derivatives of ELBO
- Suppose $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N)$ and $q_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$
- Note: our variational distribution $q(\mathbf{Z}) = q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$
- Can write the ELBO as (by definition)

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$$

- Log joint probability $\log p(\mathbf{X}, \mathbf{Z})$ for this model

$$\log p(\mathcal{D}, \mathbf{Z}) = \log p(\mathcal{D}, \mu, \tau) = \log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$$

An Example of Mean-Field VB via ELBO Derivatives

- Let's revisit the Gaussian parameter estimation via explicitly taking derivatives of ELBO
- Suppose $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N)$ and $q_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$
- Note: our variational distribution $q(\mathbf{Z}) = q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$
- Can write the ELBO as (by definition)

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$$

- Log joint probability $\log p(\mathbf{X}, \mathbf{Z})$ for this model

$$\log p(\mathcal{D}, \mathbf{Z}) = \log p(\mathcal{D}, \mu, \tau) = \log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$$

- Likewise, $\log q(\mathbf{Z}) = \log q_\mu(\mu) + \log q_\tau(\tau)$

An Example of Mean-Field VB via ELBO Derivatives

- Let's revisit the Gaussian parameter estimation via explicitly taking derivatives of ELBO
- Suppose $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N)$ and $q_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$
- Note: our variational distribution $q(\mathbf{Z}) = q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$
- Can write the ELBO as (by definition)

$$\mathcal{L}(q) \quad = \quad \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$$

- Log joint probability $\log p(\mathbf{X}, \mathbf{Z})$ for this model

$$\log p(\mathcal{D}, \mathbf{Z}) = \log p(\mathcal{D}, \mu, \tau) = \log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$$

- Likewise, $\log q(\mathbf{Z}) = \log q_\mu(\mu) + \log q_\tau(\tau)$
- Can plug these into Eq 1 and take expectations w.r.t. $q(\mathbf{Z}) = q_\mu(\mu)q_\tau(\tau)$

An Example of Mean-Field VB via ELBO Derivatives

- Let's revisit the Gaussian parameter estimation via explicitly taking derivatives of ELBO
- Suppose $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N)$ and $q_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$
- Note: our variational distribution $q(\mathbf{Z}) = q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$
- Can write the ELBO as (by definition)

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$$

- Log joint probability $\log p(\mathbf{X}, \mathbf{Z})$ for this model

$$\log p(\mathcal{D}, \mathbf{Z}) = \log p(\mathcal{D}, \mu, \tau) = \log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$$

- Likewise, $\log q(\mathbf{Z}) = \log q_\mu(\mu) + \log q_\tau(\tau)$
- Can plug these into Eq 1 and take expectations w.r.t. $q(\mathbf{Z}) = q_\mu(\mu)q_\tau(\tau)$
- Expectations simplify due to the factored form of q (do it as an exercise)

An Example of Mean-Field VB via ELBO Derivatives

- Let's revisit the Gaussian parameter estimation via explicitly taking derivatives of ELBO
- Suppose $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N)$ and $q_\tau(\tau) = \text{Gamma}(\tau|a_N, b_N)$
- Note: our variational distribution $q(\mathbf{Z}) = q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$
- Can write the ELBO as (by definition)

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$$

- Log joint probability $\log p(\mathbf{X}, \mathbf{Z})$ for this model

$$\log p(\mathcal{D}, \mathbf{Z}) = \log p(\mathcal{D}, \mu, \tau) = \log p(\mathcal{D}|\mu, \tau) + \log p(\mu|\tau) + \log p(\tau)$$

- Likewise, $\log q(\mathbf{Z}) = \log q_\mu(\mu) + \log q_\tau(\tau)$
- Can plug these into Eq 1 and take expectations w.r.t. $q(\mathbf{Z}) = q_\mu(\mu)q_\tau(\tau)$
- Expectations simplify due to the factored form of q (do it as an exercise)
- Finally take derivatives w.r.t. each variational parameter $\mu_N, \lambda_N, a_N, b_N$

Mean-Field VB Inference: Method 1 or Method 2?

- Both are iterative (work in a cyclic fashion)

Mean-Field VB Inference: Method 1 or Method 2?

- Both are iterative (work in a cyclic fashion)
- Method 1 preferred if you are comfortable computing expectations and "recognizing" the forms of distributions

Mean-Field VB Inference: Method 1 or Method 2?

- Both are iterative (work in a cyclic fashion)
- Method 1 preferred if you are comfortable computing expectations and "recognizing" the forms of distributions
- Method 2 preferred if you are comfortable taking derivatives (and also computing expectations)

Mean-Field VB Inference: Method 1 or Method 2?

- Both are iterative (work in a cyclic fashion)
- Method 1 preferred if you are comfortable computing expectations and "recognizing" the forms of distributions
- Method 2 preferred if you are comfortable taking derivatives (and also computing expectations)
  - Also more generally applicable in a wide variety of situations
  - Usually considered to be the more direct method

Mean-Field VB Inference: Method 1 or Method 2?

- Both are iterative (work in a cyclic fashion)
- Method 1 preferred if you are comfortable computing expectations and "recognizing" the forms of distributions
- Method 2 preferred if you are comfortable taking derivatives (and also computing expectations)
  - Also more generally applicable in a wide variety of situations
  - Usually considered to be the more direct method
  - The value of $\mathcal{L}$ can be readily calculated (since we have written its full expression in terms of ϕ_i's)

Mean-Field VB Inference: Method 1 or Method 2?

- Both are iterative (work in a cyclic fashion)
- Method 1 preferred if you are comfortable computing expectations and "recognizing" the forms of distributions
- Method 2 preferred if you are comfortable taking derivatives (and also computing expectations)
  - Also more generally applicable in a wide variety of situations
  - Usually considered to be the more direct method
  - The value of $\mathcal{L}$ can be readily calculated (since we have written its full expression in terms of ϕ_i's)
  - Can use the current value of $\mathcal{L}$ (ELBO) to easily check for convergence

Another Example: VB for Linear Regression via ELBO Derivatives

- Consider a Bayesian linear regression model with unknown noise variance α^{-1} (assume λ known)

$$y_i \sim \text{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \text{Normal}(0, \lambda^{-1} I), \quad \alpha \sim \text{Gamma}(a, b)$$

$$p(y, w, \alpha | x) = p(\alpha) p(w) \prod_{i=1}^{N} p(y_i | x_i, w, \alpha)$$

$$q(w, \alpha) = q(\alpha) q(w) = \text{Gamma}(\alpha | a', b') \text{Normal}(w | \mu', \Sigma')$$

Another Example: VB for Linear Regression via ELBO Derivatives

- Consider a Bayesian linear regression model with unknown noise variance α^{-1} (assume λ known)

$$y_i \sim \text{Normal}(x_i^T w, \alpha^{-1}), \quad w \sim \text{Normal}(0, \lambda^{-1} I), \quad \alpha \sim \text{Gamma}(a, b)$$

$$p(y, w, \alpha | x) = p(\alpha) p(w) \prod_{i=1}^{N} p(y_i | x_i, w, \alpha)$$

$$q(w, \alpha) = q(\alpha) q(w) = \text{Gamma}(\alpha | a', b') \text{Normal}(w | \mu', \Sigma')$$

- The ELBO $\mathcal{L}(q) = \int q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z} - \int q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}$ is

$$\begin{aligned}
\mathcal{L}(a', b', \mu', \Sigma') \quad = \quad & \int q(\alpha) \ln p(\alpha) d\alpha + \int q(w) \ln p(w) dw \\
& + \sum_{i=1}^{N} \int \int q(\alpha) q(w) \ln p(y_i | x_i, w, \alpha) dw d\alpha \\
& - \int q(\alpha) \ln q(\alpha) d\alpha - \int q(w) \ln q(w) dw
\end{aligned}$$

Another Example: VB for Linear Regression via ELBO Derivatives

- ELBO is now a function of the variational parameters a', b', μ', Σ'

$$
\begin{aligned}
\mathcal{L}(a', b', \mu', \Sigma') &= (a-1)(\psi(a') - \ln b') - b\frac{a'}{b'} + \text{constant} \\
& -\frac{\lambda}{2}(\mu'^T\mu' + \text{tr}(\Sigma')) + \text{constant} \\
& +\frac{N}{2}(\psi(a') - \ln b') - \sum_{i=1}^{N} \frac{1}{2}\frac{a'}{b'}\left((y_i - x_i^T\mu')^2 + x_i^T\Sigma' x_i\right) + \text{constant} \\
& +a' - \ln b' + \ln\Gamma(a') + (1-a')\psi(a') \\
& +\frac{1}{2}\ln|\Sigma'| + \text{constant}
\end{aligned}
$$

- ψ is the digamma function (derivative of log of gamma function)

Another Example: VB for Linear Regression via ELBO Derivatives

- ELBO is now a function of the variational parameters a', b', μ', Σ'

$$
\begin{aligned}
\mathcal{L}(a', b', \mu', \Sigma') &= (a-1)(\psi(a') - \ln b') - b\frac{a'}{b'} + \text{constant} \\
&\quad -\frac{\lambda}{2}(\mu'^T\mu' + \text{tr}(\Sigma')) + \text{constant} \\
&\quad +\frac{N}{2}(\psi(a') - \ln b') - \sum_{i=1}^{N}\frac{1}{2}\frac{a'}{b'}\left((y_i - x_i^T\mu')^2 + x_i^T\Sigma' x_i\right) + \text{constant} \\
&\quad +a' - \ln b' + \ln\Gamma(a') + (1-a')\psi(a') \\
&\quad +\frac{1}{2}\ln|\Sigma'| + \text{constant}
\end{aligned}
$$

- ψ is the digamma function (derivative of log of gamma function)
- Can now take gradients w.r.t. each of a', b', μ', Σ' and estimate these in an alternating fashion

Another Example: VB for Linear Regression via ELBO Derivatives

- ELBO is now a function of the variational parameters a', b', μ', Σ'

$$
\begin{aligned}
\mathcal{L}(a', b', \mu', \Sigma') \quad = \quad & (a-1)(\psi(a') - \ln b') - b\frac{a'}{b'} + \text{constant} \\
& -\frac{\lambda}{2}(\mu'^T\mu' + \text{tr}(\Sigma')) + \text{constant} \\
& +\frac{N}{2}(\psi(a') - \ln b') - \sum_{i=1}^{N} \frac{1}{2}\frac{a'}{b'}\left((y_i - x_i^T\mu')^2 + x_i^T\Sigma' x_i\right) + \text{constant} \\
& +a' - \ln b' + \ln\Gamma(a') + (1-a')\psi(a') \\
& +\frac{1}{2}\ln|\Sigma'| + \text{constant}
\end{aligned}
$$

- ψ is the digamma function (derivative of log of gamma function)
- Can now take gradients w.r.t. each of a', b', μ', Σ' and estimate these in an alternating fashion
- This will give us $q(\boldsymbol{w}, \alpha) = \text{Normal}(\boldsymbol{w}|\mu', \Sigma')\text{Gamma}(\alpha|a', b')$

Another Example: VB for Linear Regression via ELBO Derivatives

Inputs: Data and definitions $q(\alpha) = \text{Gamma}(\alpha|a', b')$ and $q(w) = \text{Normal}(w|\mu', \Sigma')$

Output: Values for a', b', μ' and Σ'

1. Initialize a'_0, b'_0, μ'_0 and Σ'_0 in some way
2. For iteration $t = 1, \ldots, T$
   - Update $q(\alpha)$ by setting

$$
\begin{aligned}
a'_t &= a + \frac{N}{2} \\
b'_t &= b + \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i^T\mu'_{t-1})^2 + x_i^T\Sigma'_{t-1}x_i
\end{aligned}
$$

   - Update $q(w)$ by setting

$$
\begin{aligned}
\Sigma'_t &= \left(\lambda I + \frac{a'_t}{b'_t}\sum_{i=1}^{N} x_i x_i^T\right)^{-1} \\
\mu'_t &= \Sigma'_t\left(\frac{a'_t}{b'_t}\sum_{i=1}^{N} y_i x_i\right)
\end{aligned}
$$

   - Evaluate $\mathcal{L}(a'_t, b'_t, \mu'_t, \Sigma'_t)$ to assess convergence (i.e., decide T).

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$
- For many models, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is an exponential family distribution

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$
- For many models, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is an exponential family distribution

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\mathbf{x}_n, \mathbf{z}_n) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\right\}$$

  where $\boldsymbol{\eta}$ denotes the natural parameters and $\boldsymbol{u}$ is the sufficient statistics

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$
- For many models, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is an exponential family distribution

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\mathbf{x}_n, \mathbf{z}_n) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\right\}$$

  where $\boldsymbol{\eta}$ denotes the natural parameters and $\boldsymbol{u}$ is the sufficient statistics

- Suppose the conjugate prior for the parameters $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}|\nu_0, \mathbf{v}_0) = f(\nu_0, \boldsymbol{\chi}_0) g(\boldsymbol{\eta})^{\nu_0} \exp\left\{\nu_o \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}_0\right\}$$

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$
- For many models, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is an exponential family distribution

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\mathbf{x}_n, \mathbf{z}_n) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\right\}$$

  where $\boldsymbol{\eta}$ denotes the natural parameters and $\boldsymbol{u}$ is the sufficient statistics
- Suppose the conjugate prior for the parameters $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}|\nu_0, \mathbf{v}_0) = f(\nu_0, \boldsymbol{\chi}_0) g(\boldsymbol{\eta})^{\nu_0} \exp\left\{\nu_o \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}_0\right\}$$

- Note: This prior is equivalent to having ν_0 pseudo-observations, with sufficient statistics $\boldsymbol{u} = \chi_0$

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$
- For many models, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is an exponential family distribution

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\mathbf{x}_n, \mathbf{z}_n) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\right\}$$

  where $\boldsymbol{\eta}$ denotes the natural parameters and $\boldsymbol{u}$ is the sufficient statistics
- Suppose the conjugate prior for the parameters $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}|\nu_0, \mathbf{v}_0) = f(\nu_0, \boldsymbol{\chi}_0) g(\boldsymbol{\eta})^{\nu_0} \exp\left\{\nu_o \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}_0\right\}$$

- Note: This prior is equivalent to having ν_0 pseudo-observations, with sufficient statistics $\boldsymbol{u} = \chi_0$
- We are interested in the posterior over both $\mathbf{Z}$ and $\boldsymbol{\eta}$.

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$
- For many models, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is an exponential family distribution

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\mathbf{x}_n, \mathbf{z}_n) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\right\}$$

  where $\boldsymbol{\eta}$ denotes the natural parameters and $\boldsymbol{u}$ is the sufficient statistics
- Suppose the conjugate prior for the parameters $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}|\nu_0, \mathbf{v}_0) = f(\nu_0, \boldsymbol{\chi}_0) g(\boldsymbol{\eta})^{\nu_0} \exp\left\{\nu_o \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}_0\right\}$$

- Note: This prior is equivalent to having ν_0 pseudo-observations, with sufficient statistics $\boldsymbol{u} = \chi_0$
- We are interested in the posterior over both $\mathbf{Z}$ and $\boldsymbol{\eta}$. Usually intractable.

Mean-Field VB for Exponential Family

Mean-Field VB updates are easy to identify/derive if likelihoods/priors are in exponential family

- Let's assume some model with data $\mathbf{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, local latent variables $\mathbf{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$
- For many models, the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$ is an exponential family distribution

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^{N} h(\mathbf{x}_n, \mathbf{z}_n) g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\right\}$$

  where $\boldsymbol{\eta}$ denotes the natural parameters and $\boldsymbol{u}$ is the sufficient statistics
- Suppose the conjugate prior for the parameters $\boldsymbol{\eta}$ is

$$p(\boldsymbol{\eta}|\nu_0, \mathbf{v}_0) = f(\nu_0, \boldsymbol{\chi}_0) g(\boldsymbol{\eta})^{\nu_0} \exp\left\{\nu_o \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}_0\right\}$$

- Note: This prior is equivalent to having ν_0 pseudo-observations, with sufficient statistics $\boldsymbol{u} = \chi_0$
- We are interested in the posterior over both $\mathbf{Z}$ and $\boldsymbol{\eta}$. Usually intractable. Let's use Mean-Field VB.

Mean-Field VB for Exponential Family

- Using method 1, variational post'r for **Z** can be written (only keeping terms that depend on **Z**)

$$\begin{aligned}\ln q^\star(\mathbf{Z}) &= \mathbb{E}_\eta[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\ &= \sum_{n=1}^{N}\left\{\ln h(\mathbf{x}_n,\mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)\right\} + \text{const}\end{aligned}$$

Mean-Field VB for Exponential Family

- Using method 1, variational post'r for **Z** can be written (only keeping terms that depend on **Z**)

$$\begin{aligned}\ln q^\star(\mathbf{Z}) &= \mathbb{E}_\eta[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\ &= \sum_{n=1}^{N} \{\ln h(\mathbf{x}_n, \mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\} + \text{const}\end{aligned}$$

- Exponentiating again, we get the following form

$$q^\star(\mathbf{z}_n) = h(\mathbf{x}_n, \mathbf{z}_n) g\left(\mathbb{E}[\boldsymbol{\eta}]\right) \exp\left\{\mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)\right\}$$

Mean-Field VB for Exponential Family

- Using method 1, variational post'r for **Z** can be written (only keeping terms that depend on **Z**)

$$\begin{aligned}\ln q^\star(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\eta}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\ &= \sum_{n=1}^{N} \left\{\ln h(\mathbf{x}_n,\mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)\right\} + \text{const}\end{aligned}$$

- Exponentiating again, we get the following form

$$q^\star(\mathbf{z}_n) = h(\mathbf{x}_n,\mathbf{z}_n) g\left(\mathbb{E}[\boldsymbol{\eta}]\right) \exp\left\{\mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)\right\}$$

- Likewise, the variational posterior for the parameters $\boldsymbol{\eta}$ (only keeping terms that depend on $\boldsymbol{\eta}$)

$$\ln q^\star(\boldsymbol{\eta}) = \ln p(\boldsymbol{\eta}|\nu_0,\boldsymbol{\chi}_0) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \text{const}$$

$$= \nu_0 \ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}_0 + \sum_{n=1}^{N}\left\{\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^{\mathrm{T}}\mathbb{E}_{\mathbf{z}_n}[\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)]\right\} + \text{const}$$

Mean-Field VB for Exponential Family

- Using method 1, variational post'r for $\mathbf{Z}$ can be written (only keeping terms that depend on $\mathbf{Z}$)

$$\begin{aligned}\ln q^\star(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\eta}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\ &= \sum_{n=1}^{N}\left\{\ln h(\mathbf{x}_n,\mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)\right\} + \text{const}\end{aligned}$$

- Exponentiating again, we get the following form

$$q^\star(\mathbf{z}_n) = h(\mathbf{x}_n,\mathbf{z}_n)g\left(\mathbb{E}[\boldsymbol{\eta}]\right)\exp\left\{\mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)\right\}$$

- Likewise, the variational posterior for the parameters $\boldsymbol{\eta}$ (only keeping terms that depend on $\boldsymbol{\eta}$)

$$\begin{aligned}\ln q^\star(\boldsymbol{\eta}) &= \ln p(\boldsymbol{\eta}|\nu_0,\boldsymbol{\chi}_0) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\ &= \nu_0\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}_0 + \sum_{n=1}^{N}\left\{\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^{\mathrm{T}}\mathbb{E}_{\mathbf{z}_n}[\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)]\right\} + \text{const}\end{aligned}$$

- Again, exponentiating gives the following variational distribution

$$q^\star(\boldsymbol{\eta}) = f(\nu_N,\boldsymbol{\chi}_N)g(\boldsymbol{\eta})^{\nu_N}\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}_N\right\}$$

$$\begin{aligned}\nu_N &= \nu_0 + N \\ \boldsymbol{\chi}_N &= \boldsymbol{\chi}_0 + \sum_{n=1}^{N}\mathbb{E}_{\mathbf{z}_n}[\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)]\end{aligned}$$

Mean-Field VB for Exponential Family

- Using method 1, variational post'r for **Z** can be written (only keeping terms that depend on **Z**)

$$\begin{aligned}\ln q^\star(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\eta}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\ &= \sum_{n=1}^{N}\left\{\ln h(\mathbf{x}_n,\mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)\right\} + \text{const}\end{aligned}$$

- Exponentiating again, we get the following form

$$q^\star(\mathbf{z}_n) = h(\mathbf{x}_n,\mathbf{z}_n)g\left(\mathbb{E}[\boldsymbol{\eta}]\right)\exp\left\{\mathbb{E}[\boldsymbol{\eta}^{\mathrm{T}}]\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)\right\}$$

- Likewise, the variational posterior for the parameters $\boldsymbol{\eta}$ (only keeping terms that depend on $\boldsymbol{\eta}$)

$$\ln q^\star(\boldsymbol{\eta}) = \ln p(\boldsymbol{\eta}|\nu_0,\boldsymbol{\chi}_0) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\eta})] + \text{const}$$

$$= \nu_0\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}_0 + \sum_{n=1}^{N}\left\{\ln g(\boldsymbol{\eta}) + \boldsymbol{\eta}^{\mathrm{T}}\mathbb{E}_{\mathbf{z}_n}[\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)]\right\} + \text{const}$$

- Again, exponentiating gives the following variational distribution

$$q^\star(\boldsymbol{\eta}) = f(\nu_N,\boldsymbol{\chi}_N)g(\boldsymbol{\eta})^{\nu_N}\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}_N\right\}$$

$$\begin{aligned}\nu_N &= \nu_0 + N \\ \boldsymbol{\chi}_N &= \boldsymbol{\chi}_0 + \sum_{n=1}^{N}\mathbb{E}_{\mathbf{z}_n}[\mathbf{u}(\mathbf{x}_n,\mathbf{z}_n)]\end{aligned}$$

- Again note that updates of $q(\mathbf{Z})$ and $q(\boldsymbol{\eta})$ are coupled

VB and Expectation Maximization (EM)

- VB can be seen as a generalization of the EM algorithm
- Unlike EM, in VB there is no distinction between parameters Θ and latent variables $\mathbf{Z}$

VB and Expectation Maximization (EM)

- VB can be seen as a generalization of the EM algorithm
- Unlike EM, in VB there is no distinction between parameters Θ and latent variables $\mathbf{Z}$
- VB treats all unknowns of the model as latent variables and calls them $\mathbf{Z}$

VB and Expectation Maximization (EM)

- VB can be seen as a generalization of the EM algorithm
- Unlike EM, in VB there is no distinction between parameters Θ and latent variables $\mathbf{Z}$
- VB treats all unknowns of the model as latent variables and calls them $\mathbf{Z}$
- Since there is no notion of "parameters", VB is like EM without the "M step"

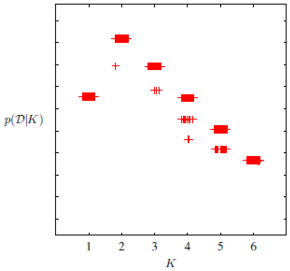
VB and Expectation Maximization (EM)

- VB can be seen as a generalization of the EM algorithm
- Unlike EM, in VB there is no distinction between parameters Θ and latent variables $\mathbf{Z}$
- VB treats all unknowns of the model as latent variables and calls them $\mathbf{Z}$
- Since there is no notion of "parameters", VB is like EM without the "M step"
- VB can be used within an EM algorithm if the E step is intractable
    - This is known as Variational EM algorithm

ELBO for Model Selection

- ELBO can also be used for model selection
- We can compute ELBO for each model and then choose the one with largest value of ELBO
- An Example: The ELBO plot for a Gaussian Mixture Model with different K values

Plot of the variational lower bound $\mathcal{L}$ versus the number K of components in the Gaussian mixture model, for the Old Faithful data, showing a distinct peak at $K = 2$ components. For each value of K, the model is trained from 100 different random starts, and the results shown as '+' symbols plotted with small random horizontal perturbations so that they can be distinguished. Note that some solutions find suboptimal local maxima, but that this happens infrequently.



- Note that unlike likelihood, ELBO doesn't monotonically increase with K (penalizes large K)

Figure courtesy: PRML (Bishop, 2006)

Some Properties of VB

Recall that VB is equivalent to finding q by minimizing $\text{KL}(q||p)$

$$\text{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

If the true posterior p is very small in some region then, to minimize $\text{KL}(q||p)$, the approx. dist. q will also have to be very small (otherwise KL will be very large)

Some Properties of VB

Recall that VB is equivalent to finding q by minimizing $\mathrm{KL}(q||p)$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

If the true posterior p is very small in some region then, to minimize $\mathrm{KL}(q||p)$, the approx. dist. q will also have to be very small (otherwise KL will be very large)

This has two key consequences for VB

- Underestimates the variances of the true posterior
- For multimodal posteriors, VB locks onto one of the modes

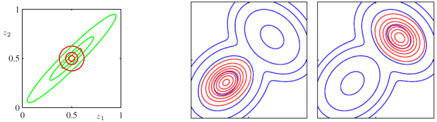


Figure: (Left) Zero-Forcing Property of VB, (Right) For multi-modal posterior, VB locks onto one of the models

Some Properties of VB

Recall that VB is equivalent to finding q by minimizing $\mathrm{KL}(q||p)$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

If the true posterior p is very small in some region then, to minimize $\mathrm{KL}(q||p)$, the approx. dist. q will also have to be very small (otherwise KL will be very large)

This has two key consequences for VB

- Underestimates the variances of the true posterior
- For multimodal posteriors, VB locks onto one of the modes

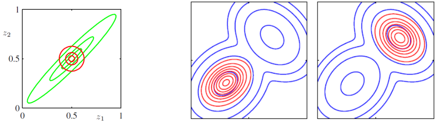


Figure: (Left) Zero-Forcing Property of VB, (Right) For multi-modal posterior, VB locks onto one of the models

Note: Some other inference methods, e.g., Expectation Propagation (EP) can avoid this behavior

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)
- Finds the best q by maximizing the ELBO (or minimizing $\mathrm{KL}(q||p)$)

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)
- Finds the best q by maximizing the ELBO (or minimizing $\mathrm{KL}(q||p)$)
- VB is guaranteed to converge to a local optima (in finite time)

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)
- Finds the best q by maximizing the ELBO (or minimizing $\text{KL}(q||p)$)
- VB is guaranteed to converge to a local optima (in finite time)
- Simplifying assumption on q (e.g., mean field VB) can make VB update very easy to derive

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)
- Finds the best q by maximizing the ELBO (or minimizing $\text{KL}(q||p)$)
- VB is guaranteed to converge to a local optima (in finite time)
- Simplifying assumption on q (e.g., mean field VB) can make VB update very easy to derive

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)
- Finds the best q by maximizing the ELBO (or minimizing $\text{KL}(q||p)$)
- VB is guaranteed to converge to a local optima (in finite time)
- Simplifying assumption on q (e.g., mean field VB) can make VB update very easy to derive
- More advanced VB methods can handle richer forms of $q(\mathbf{Z})$, likelihood and priors

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)
- Finds the best q by maximizing the ELBO (or minimizing $\text{KL}(q||p)$)
- VB is guaranteed to converge to a local optima (in finite time)
- Simplifying assumption on q (e.g., mean field VB) can make VB update very easy to derive
- More advanced VB methods can handle richer forms of $q(\mathbf{Z})$, likelihood and priors
  - E.g., Monte-Carlo approximations of ELBO and its derivatives

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)
- Finds the best q by maximizing the ELBO (or minimizing $\text{KL}(q||p)$)
- VB is guaranteed to converge to a local optima (in finite time)
- Simplifying assumption on q (e.g., mean field VB) can make VB update very easy to derive
- More advanced VB methods can handle richer forms of $q(\mathbf{Z})$, likelihood and priors
  - E.g., Monte-Carlo approximations of ELBO and its derivatives
- Implementations of many classic/advanced VB methods available in Stan, Edward, etc.

Summary

- VB is a deterministic approximate inference method (unlike sampling methods)
- Finds the best q by maximizing the ELBO (or minimizing $\text{KL}(q||p)$)
- VB is guaranteed to converge to a local optima (in finite time)
- Simplifying assumption on q (e.g., mean field VB) can make VB update very easy to derive
- More advanced VB methods can handle richer forms of $q(\mathbf{Z})$, likelihood and priors
    - E.g., Monte-Carlo approximations of ELBO and its derivatives
- Implementations of many classic/advanced VB methods available in Stan, Edward, etc.
- VB can be a very useful inference method to apply Bayesian models for large-scale data. A lot of recent work on stochastic (i.e., online) variational inference algorithms that work with small randomly chosen minibatches of data and can easily scale to massive-scale data sets