Approximate Inference: Variational Inference

CMSC 691
UMBC
Goal: Posterior Inference

Hyperparameters $\alpha$

Unknown “parameters” $\Theta$

Data: $p(\Theta)$

Likelihood model: $p_{\alpha}(\Theta | \text{Data})$
(Some) Learning Techniques

MAP/MLE: Point estimation, basic EM

Variational Inference: Functional Optimization

Sampling/Monte Carlo

what we’ve already covered

today

next class
Outline

Variational Inference

Basic Technique

Example: Topic Models
Variational Inference: Core Idea

- Observed $x$, latent r.v.s $\theta$
- We have some joint model $p(\theta, x)$
- We want to compute $p(\theta|x)$ but this is computationally difficult
Variational Inference: Core Idea

• Observed $x$, latent r.v.s $\theta$
• We have some joint model $p(\theta, x)$
• We want to compute $p(\theta|x)$ but this is computationally difficult

• Solution: approximate $p(\theta|x)$ with a different distribution $q_\lambda(\theta)$ and make $q_\lambda(\theta)$ “close” to $p(\theta|x)$
Variational Inference

$p(\theta \mid x)$

Difficult to compute
Variational Inference

Minimize the "difference" by changing $\lambda$

$p(\theta | x)$: Difficult to compute

$q(\theta)$: controlled by parameters $\lambda$

$q(\theta)$: Easy (ier) to compute
Variational Inference

$p(\theta | x)$

Difficult to compute

$q(\theta)$

Easy(ier) to compute

Minimize the "difference" by changing $\lambda$
Variational Inference: A Gradient-Based Optimization Technique

Set $t = 0$

**Pick** a starting value $\lambda_t$

Until **converged**:  
1. Get value $y_t = F(q(\bullet; \lambda_t))$
2. Get gradient $g_t = F'(q(\bullet; \lambda_t))$
3. Get **scaling factor** $\rho_t$
4. Set $\lambda_{t+1} = \lambda_t + \rho_t * g_t$
5. Set $t += 1$
Set $t = 0$

Pick a starting value $\lambda_t$

Until converged:
1. Get value $y_t = F(q(\bullet; \lambda_t))$
2. Get gradient $g_t = F'(q(\bullet; \lambda_t))$
3. Get scaling factor $\rho_t$
4. Set $\lambda_{t+1} = \lambda_t + \rho_t * g_t$
5. Set $t += 1$
Variational Inference: The Function to Optimize

Any easy-to-compute distribution

\[ \mathcal{D}_{KL} (q(\theta) \parallel p(\theta | x)) \]

Posterior of desired model
Variational Inference:
The Function to Optimize

Any easy-to-compute distribution

\[
\min_{q} \mathbb{D}_{KL} (q(\theta) \ || \ p(\theta \mid x))
\]

- Find the best distribution (calculus of variations)
- Posterior of desired model
Variational Inference:
The Function to Optimize

\[
\min_{q} D_{KL}(q(\theta) \parallel p(\theta \mid x))
\]

Find the best distribution

Parameters for desired model

\[
q(\theta)
\]
Variational Inference: The Function to Optimize

\[
\min_q D_{KL} \left( q(\theta) \ || \ p(\theta \mid x) \right)
\]

Find the best distribution

Parameters for desired model

\[
q \left( \theta \mid \lambda \right)
\]

Variational parameters for \( \theta \)
Variational Inference: The Function to Optimize

KL-Divergence (expectation)

\[
\min_q D_{KL}(q(\theta) \mid\mid p(\theta \mid x))
\]

Find the best distribution

\[
D_{KL}(q(\theta) \mid\mid p(\theta \mid x)) = \\
\mathbb{E}_{q(\theta)} \left[ \log \frac{q(\theta)}{p(\theta \mid x)} \right]
\]

Parameters for desired model

\[
q(\theta \mid \lambda)
\]

Variational parameters for \(\theta\)
Variational Inference

\[
\min_q \mathbb{E}_{q(\theta)} \log \left( \frac{q(\theta)}{p(\theta | x)} \right)
\]

Find the \textit{best} distribution

Parameters for \textit{desired} model

\[
q(\theta | \lambda)
\]

Variational parameters for \(\theta\)
Exponential Family Recap: "Easy" Posterior Inference

\[ p(\theta \mid x) \propto \pi(x \mid \theta) p(\theta) \]

\( p \) is the conjugate prior for \( \pi \)

Exponential Family Recap: "Easy" Expectations

\[ \mathbb{E}_{p_\theta} [f(x)] = \nabla_\theta A(\theta) \]
Variational Inference

\[
\min_{q} \text{KL}(q(\theta) \parallel p(\theta | x))
\]

Find the best distribution

When \( p \) and \( q \) are the same exponential family form, the variational update \( q(\theta) \) is (often) computable (in closed form)
Variational Inference: A Gradient-Based Optimization Technique

Set $t = 0$

**Pick** a starting value $\lambda_t$

Let

$$F(q(\cdot; \lambda_t)) = KL[q(\cdot; \lambda_t) || p(\cdot)]$$

Until **converged**:

1. Get value $y_t = F(q(\cdot; \lambda_t))$
2. Get gradient $g_t = F'(q(\cdot; \lambda_t))$
3. Get **scaling factor** $\rho_t$
4. Set $\lambda_{t+1} = \lambda_t + \rho_t * g_t$
5. Set $t += 1
Variational Inference: Maximization or Minimization?

\[ \mathcal{L}(q, p) = \mathbb{E}_q [\log p(x, \theta)] - \mathbb{E}_q [q(\theta)] \]

\[ = -D_{KL}(q(\theta) \| p(\theta | x)) + \text{constant} \]

\[ = \text{Evidence Lower Bound (ELBO)} \]
Evidence Lower Bound (ELBO)

$$\log p(x) = \log \int p(x, \theta) d\theta$$
Evidence Lower Bound (ELBO)

\[ \log p(x) = \log \int p(x, \theta) d\theta \]

\[ = \log \int p(x, \theta) \frac{q(\theta)}{q(\theta)} d\theta \]
Evidence Lower Bound (ELBO)

\[
\log p(x) = \log \int p(x, \theta) d\theta \\
= \log \int p(x, \theta) \frac{q(\theta)}{q(\theta)} d\theta \\
= \log \mathbb{E}_{q(\theta)} \left[ \frac{p(x, \theta)}{q(\theta)} \right]
\]
Evidence Lower Bound (ELBO)

\[
\log p(x) = \log \int p(x, \theta) d\theta
\]

\[
= \log \int p(x, \theta) \frac{q(\theta)}{q(\theta)} d\theta
\]

\[
= \log \mathbb{E}_{q(\theta)} \left[ \frac{p(x, \theta)}{q(\theta)} \right]
\]

\[
\geq \mathbb{E}_{q(\theta)} [p(x, \theta)] - \mathbb{E}_{q(\theta)} [q(\theta)]
\]

\[
= \mathcal{L}(q)
\]
Outline

Variational Inference

Basic Technique

Example: Topic Models
Mixture Model vs. Admixture Model

- Both consider K generating distributions
- Mixture model
- Admixture model
Mixture Model vs. Admixture Model

• Both consider K generating distributions

• Mixture model
  – Each of the N datapoints is generated from one of those K distributions

• Admixture model
Mixture Model vs. Admixture Model

- Both consider K generating distributions
- Mixture model
  - Each of the N datapoints is generated from one of those K distributions
- Admixture model
  - Each of the N datapoints is generated from a mixture of those K distributions
Bag-of-Items Models: Admixture Models

\[ p(\text{Three: 1, people: 2, attack: 2, ...}) = p(\text{Unigram counts}) \]

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Bag-of-Items Models: Admixture Models

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

\[ p(\cdot) = p_{\phi, \omega}(\cdot) \]

Unigram counts

- Three: 1
- people: 2
- attack: 2
- ...

Global (corpus-level) parameters interact with local (document-level) parameters
Latent Dirichlet Allocation (Blei et al., 2003)

Per-document (unigram) word counts
Latent Dirichlet Allocation (Blei et al., 2003)

Count of word $j$ in document $i$
Latent Dirichlet Allocation
(Blei et al., 2003)

Count of word $j$
in document $i$

Core assumptions:
1. K “topics”: distributions over possible vocab words
Latent Dirichlet Allocation (Blei et al., 2003)

Count of word $j$ in document $i$

Core assumptions:
1. $K$ “topics”: distributions over possible vocab words
2. Each document $i$ has general “preferences” for which topics to use
Latent Dirichlet Allocation (Blei et al., 2003)

Count of word j in document i

Core assumptions:
1. K “topics”: distributions over possible vocab words
2. Each document i has general “preferences” for which topics to use
3. Each observed word j in a document i can come from a different topic
Latent Dirichlet Allocation (Blei et al., 2003)

Count of word j in document i

\[ \text{Per-document (unigram) word counts} \]

K “topics”: distribution over vocabulary

\[ \text{Per-topic word usage} \]

\[ \text{Per-document (latent) topic usage} \]
Latent Dirichlet Allocation
(Blei et al., 2003)

Per-document (unigram) word counts

Per-document (latent) topic usage

Per-topic word usage
Latent Dirichlet Allocation (Blei et al., 2003)

Core assumptions:
1. $K$ “topics”: distributions over possible vocab words
2. Each document $i$ has general “preferences” for which topics to use
3. Each observed word $j$ in a document $i$ can come from a different topic
Latent Dirichlet Allocation (Blei et al., 2003)

Core assumptions:
1. K “topics”: distributions over possible vocab words
2. Each document i has general “preferences” for which topics to use
3. Each observed word j in a document i can come from a different topic
Latent Dirichlet Allocation
(Blei et al., 2003)

Core assumptions:
1. K “topics”: distributions over possible vocab words
2. Each document i has general “preferences” for which topics to use
3. Each observed word j in a document i can come from a different topic
Latent Dirichlet Allocation (Blei et al., 2003)

Core assumptions:
1. K “topics”: distributions over possible vocab words
2. Each document $i$ has general “preferences” for which topics to use
3. Each observed word $j$ in a document $i$ can come from a different topic
Latent Dirichlet Allocation  
(Blei et al., 2003)

Core assumptions:
1. K “topics”: distributions over possible vocab words  
2. Each document i has general “preferences” for which topics to use  
3. Each observed word j in a document i can come from a different topic
Variational Inference: LDirA

\[ p: \text{True model} \]

\[ w^{(d,n)} \sim \text{Discrete}(\phi_{z^{(d,n)}}) \]
\[ \phi_k \sim \text{Dirichlet}(\beta) \]
\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

Explicit conditioning left off (for space)
Variational Inference: LDirA

\[ w^{(d,n)} \sim \text{Discrete}(\phi_{z^{(d,n)}}) \]
\[ \phi_k \sim \text{Dirichlet}(\beta) \]
\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

\[ \phi_k \sim \text{Dirichlet}(\lambda_k) \]
\[ z^{(d,n)} \sim \text{Discrete}(\psi^{(d,n)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\gamma_d) \]

Explicit conditioning left off (for space)
Variational Inference: LDirA

\[ w^{(d,n)} \sim \text{Discrete}(\phi_{z^{(d,n)}}) \]
\[ \phi_k \sim \text{Dirichlet}(\beta) \]
\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

\[ \phi_k \sim \text{Dirichlet}(\lambda_k) \]
\[ z^{(d,n)} \sim \text{Discrete}(\psi^{(d,n)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\gamma_d) \]

Notice: no shared parameters!!!
Variational Inference: A Gradient-Based Optimization Technique

Set $t = 0$

Pick a starting value $\lambda_t$

Let

$$F(q(\bullet;\lambda_t)) = \text{KL}[q(\bullet;\lambda_t) \mid \mid p(\bullet)]$$

Until converged:

1. Get value $y_t = F(q(\bullet;\lambda_t))$
2. Get gradient $g_t = F'(q(\bullet;\lambda_t))$
3. Get scaling factor $\rho_t$
4. Set $\lambda_{t+1} = \lambda_t + \rho_t * g_t$
5. Set $t += 1$
Variational Inference: LDirA Topic Proportions

\[ p: \text{True model} \]
\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

\[ q: \text{Mean-field approximation} \]
\[ z^{(d,n)} \sim \text{Discrete}(\psi^{(d,n)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\gamma_d) \]

\[ \mathbb{E}_{q(\theta^{(d)})}\left[ \log p(\theta^{(d)} \mid \alpha) \right] \]
Variational Inference: LDirA Topic Proportions

\textbf{p: True model}

\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

\textbf{q: Mean-field approximation}

\[ z^{(d,n)} \sim \text{Discrete}(\psi^{(d,n)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\gamma_d) \]

\[ \mathbb{E}_{q(\theta^{(d)})} \left[ \log p(\theta^{(d)} \mid \alpha) \right] = \]

\[ \mathbb{E}_{q(\theta^{(d)})} \left[ (\alpha - 1)^T \log \theta^{(d)} + C \right] = \]

\[ \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \]

\[ \text{params} = (\alpha_k - 1)_k \]

\[ \text{suff. stats.} = (\log \theta_k)_k \]
**Variational Inference: LDirA Topic Proportions**

\[ p: \text{True model} \]

\[
\begin{align*}
    z^{(d,n)} &\sim \text{Discrete}(\theta^{(d)}) \\
    \theta^{(d)} &\sim \text{Dirichlet}(\alpha)
\end{align*}
\]

\[ q: \text{Mean-field approximation} \]

\[
\begin{align*}
    z^{(d,n)} &\sim \text{Discrete}(\psi^{(d,n)}) \\
    \theta^{(d)} &\sim \text{Dirichlet}(\nu_d)
\end{align*}
\]

\[
\mathbb{E}_{q(\theta^{(d)})} \left[ \log p(\theta^{(d)} | \alpha) \right] = \\
\mathbb{E}_{q(\theta^{(d)})} \left[ (\alpha - 1)^T \log \theta^{(d)} + C \right]
\]

**params** = \( (\psi_k - 1)_k \)

**suff. stats.** = \( (\log \theta_k)_k \)

*expectation of sufficient statistics of q distribution*
Variational Inference: LDirA Topic Proportions

\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

**p: True model**

\[ z^{(d,n)} \sim \text{Discrete}(\psi^{(d,n)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\nu_d) \]

**q: Mean-field approximation**

\[ \mathbb{E}_{q(\theta^{(d)})} \left[ \log p(\theta^{(d)} \mid \alpha) \right] = \]
\[ (\alpha - 1)^T \mathbb{E}_{q(\theta^{(d)})} \left[ \log \theta^{(d)} \right] + C \]

The expectation of the sufficient statistics is the gradient of the log normalizer.
Variational Inference: LDirA Topic Proportions

\( p: \) True model
\[
\begin{align*}
z^{(d,n)} &\sim \text{Discrete}(\theta^{(d)}) \\
\theta^{(d)} &\sim \text{Dirichlet}(\alpha)
\end{align*}
\]

\( q: \) Mean-field approximation
\[
\begin{align*}
z^{(d,n)} &\sim \text{Discrete}(\psi^{(d,n)}) \\
\theta^{(d)} &\sim \text{Dirichlet}(\gamma_d)
\end{align*}
\]

\[
\mathbb{E}_{q(\theta^{(d)})} \left[ \log p(\theta^{(d)} | \alpha) \right] = \\
\mathbb{E}_{q(\theta^{(d)})} \left[ (\alpha - 1)^T \log \theta^{(d)} + C \right] = \\
(\alpha - 1)^T \nabla_{\gamma_d} A(\gamma_d - 1) + C
\]

expectation of the sufficient statistics is the gradient of the log normalizer
Variational Inference: LDirA Topic Proportions

\[ p: \text{True model} \]
\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

\[ q: \text{Mean-field approximation} \]
\[ z^{(d,n)} \sim \text{Discrete}(\psi^{(d,n)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\gamma_d) \]

\[ \mathbb{E}_{q(\theta^{(d)})}[\log p(\theta^{(d)} | \alpha)] = (\alpha - 1)^T \nabla_{\gamma_d} A(\gamma_d - 1) + C \]

\[ \mathcal{L} \bigg|_{\gamma_d} = (\alpha - 1)^T \nabla_{\gamma_d} A(\gamma_d - 1) + M(\gamma_d) \]

there’s more math to do!
Variational Inference: A Gradient-Based Optimization Technique

Set $t = 0$
Pick a starting value $\lambda_t$
Let
$$F(q(\bullet; \lambda_t)) = KL[q(\bullet; \lambda_t) || p(\bullet)]$$
Until converged:
1. Get value $y_t = F(q(\bullet; \lambda_t))$
2. Get gradient $g_t = F'(q(\bullet; \lambda_t))$
3. Get scaling factor $\rho_t$
4. Set $\lambda_{t+1} = \lambda_t + \rho_t * g_t$
5. Set $t += 1$
Variational Inference: LDirA Topic Proportions

\[ p: \text{True model} \]
\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

\[ q: \text{Mean-field approximation} \]
\[ z^{(d,n)} \sim \text{Discrete}(\psi^{(d,n)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\gamma_d) \]

\[ \mathcal{L} \Bigg|_{\gamma_d} = (\alpha - 1)^T \nabla_{\gamma_d} A(\gamma_d - 1) + M(\gamma_d) \]

\[ \nabla_{\gamma_d} \mathcal{L} \Bigg|_{\gamma_d} = (\alpha - 1)^T \nabla_{\gamma_d}^2 A(\gamma_d - 1) + \nabla_{\gamma_d} M(\gamma_d) \]
Variational Inference: LDirA Topic Proportions

\[ p: \text{True model} \]
\[ z^{(d,n)} \sim \text{Discrete}(\theta^{(d)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\alpha) \]

\[ q: \text{Mean-field approximation} \]
\[ z^{(d,n)} \sim \text{Discrete}(\psi^{(d,n)}) \]
\[ \theta^{(d)} \sim \text{Dirichlet}(\gamma_d) \]

\[ \mathcal{L} \bigg|_{\gamma_d} = (\alpha - 1)^T \nabla_{\gamma_d} A(\gamma_d - 1) + M(\gamma_d) \]

\[ \nabla_{\gamma_d} \mathcal{L} \bigg|_{\gamma_d} = (\alpha - 1)^T \nabla_{\gamma_d}^2 A(\gamma_d - 1) + \nabla_{\gamma_d} M(\gamma_d) \]

analytically solve this for faster convergence (Blei et al., 2003)
Variational Inference: Core Idea

Basic Technique

- Observed $x$, latent r.v.s $\theta$
- We have some joint model $p(\theta, x)$
- We want to compute $p(\theta|x)$ but this is computationally difficult

Example: Topic Models

- Solution: approximate $p(\theta|x)$ with a different distribution $q_\lambda(\theta)$ and make $q_\lambda(\theta)$ “close” to $p(\theta|x)$