Coordinate Ascent
Variational Inference
Reminders

• Lecture on Friday, Recitation on Monday
• Exam Rubrics and Exam Viewings
• Homework 4: MCMC
  – Out: Mon, Oct 24
  – Due: Fri, Nov 3 at 11:59pm
• Homework 5: Variational Inference
  – Out: Fri, Nov 3
  – Due: Wed, Nov 16 at 11:59pm
MEAN FIELD WITH GRADIENT ASCENT
Mean Field V.I. Overview

1. **Goal**: estimate $p_\alpha(z \mid x)$
   we assume this is intractable to compute exactly

2. **Idea**: approximate with another distribution $q_\theta(z) \approx p_\alpha(z \mid x)$ for each $x$

3. **Mean Field**: assume $q_\theta(z) = \prod_t q_t(z_t; \theta)$
   i.e., we decompose over variables
   other choices for the decomposition of $q_\theta(z)$ give rise to “structured mean field”

4. **Optimization Problem**: pick the $q$ that minimizes $\text{KL}(q \parallel p)$
   \[
   \hat{q}(z) = \arg\min_{q(z) \in \mathcal{Q}} \text{KL}(q(z) \parallel p(z \mid x))
   \]
   \[
   \hat{\theta} = \arg\min_{\theta \in \Theta} \text{KL}(q_\theta(z) \parallel p_\alpha(z \mid x))
   \]

5. **Optimization Algorithm**: pick your favorite {coordinate descent, gradient descent, etc.}
Mean Field V.I. Overview

1. **Goal**: estimate $p_{\alpha}(z \mid x)$
   we assume this is intractable to compute exactly

2. **Idea**: approximate with another distribution $q_{\theta}(z) \approx p_{\alpha}(z \mid x)$
   for each $x$

3. **Mean Field**: assume $q_{\theta}(z) = \prod_t q_t(z_t; \theta)$
   i.e., we decompose over variables
   other choices for the decomposition of $q_{\theta}(z)$ give rise to
   “structured mean field”

4. **Optimization Problem**: pick the $q$ that minimizes $KL(q || p)$
   \[ \hat{\theta} = \arg\min_{\theta} KL(q_{\theta}(z) \mid \mid p_{\alpha}(z \mid x)) = \arg\max_{\theta} \text{ELBO}(q_{\theta}) \]
   \[ \text{ELBO}(q_{\theta}) = E_{q_{\theta}(z)} \left[ \log p_{\alpha}(x, z) \right] - E_{q_{\theta}(z)} \left[ \log q_{\theta}(z) \right] \]
   \[ \text{ELBO}(q_{\theta}) = E_{q_{\theta}(z)} \left[ \log \tilde{p}_{\alpha}(z \mid x) \right] - E_{q_{\theta}(z)} \left[ \log q_{\theta}(z) \right] \]

5. **Optimization Algorithm**: pick your favorite {coordinate ascent, gradient ascent, etc.}
**Mean Field V.I. Overview**

1. "**Goal:** estimate \( p_\alpha(z \mid x) \)
   
   we assume this is intractable to compute exactly

2. "**Idea:** approximate with another distribution \( q_\theta(z) \approx p_\alpha(z \mid x) \)
   
   for each \( x \)

3. "**Mean Field:** assume \( q_\theta(z) = \prod_t q_t(z_t; \theta) \)
   
   i.e., we decompose over variables

   other choices for the decomposition of \( q_\theta(z) \) give rise to “structured mean field”

4. "**Optimization Problem:** pick the \( q \) that minimizes \( KL(q \| p) \)
   
   \[
   \hat{\theta} = \arg\min_{\theta} KL(q_\theta(z) \| p_\alpha(z \mid x)) = \arg\max_{\theta} ELBO(q_\theta)
   \]

   \[
   ELBO(q_\theta) = E_{q_\theta(z)} [\log p_\alpha(x, z)] - E_{q_\theta(z)} [\log q_\theta(z)]
   \]

5. "**Optimization Algorithm:** gradient ascent
Mean Field w/Gradient Ascent

- **Note**: GA does local maximization, but ELBO is generally non-convex

- **Algorithm**:
  - Initialize \( \theta \)
  - while not converged:
    \[
    \theta \leftarrow \theta + \gamma \nabla_{\theta} \text{ELBO}(q_{\theta})
    \]

- **Gradient of ELBO**:
  \[
  \nabla_{\theta} \text{ELBO}(q_{\theta}) = \nabla_{\theta} \mathbb{E}_{q_{\theta}}[\log p_{\alpha}(x, z)] - \nabla_{\theta} \mathbb{E}_{q_{\theta}}[\log q_{\theta}(z)]
  \]
  \[
  = \cdots
  \]
  \[
  = \cdots
  \]
  \[
  = \text{easy b/c of a simple } q_{\theta}
  \]
BACKGROUND: BLOCK COORDINATE DESCENT
Coordinate Descent

• Goal: minimize some objective

\[ \hat{\theta}^* = \arg\min_{\hat{\theta}} J(\hat{\theta}) \]

• Idea: iteratively pick one variable and minimize the objective w.r.t. just that one variable, \textit{keeping all the others fixed}. 

\[ \theta_2 \]

\[ \theta_1 \]
Block Coordinate Descent

• Goal: minimize some objective \textbf{(with 2 blocks)}

\[
\tilde{\alpha}^*, \tilde{\beta}^* = \arg\min_{\tilde{\alpha}, \tilde{\beta}} J(\tilde{\alpha}, \tilde{\beta})
\]

• Idea: iteratively pick one \textit{block} of variables (\(\tilde{\alpha}\) or \(\tilde{\beta}\)) and minimize the objective w.r.t. that block, keeping the other(s) fixed.

\textbf{Init}\quad \alpha, \beta

\textbf{while} not converged:

\[
\tilde{\alpha} = \arg\min_{\tilde{\alpha}} J(\tilde{\alpha}, \tilde{\beta})
\]

\[
\tilde{\beta} = \arg\min_{\tilde{\beta}} J(\tilde{\alpha}, \tilde{\beta})
\]
Block Coordinate Descent

• Goal: minimize some objective (with $T$ blocks)

$$\alpha_1, \ldots, \alpha_T = \arg\min_{\alpha_1} \cdots \arg\min_{\alpha_T} J(\alpha_1, \ldots, \alpha_T)$$

• Idea: iteratively pick one block of variables (e.g. the vector $\alpha_t$) and minimize the objective w.r.t. that block, keeping the other(s) fixed.

Init. $\alpha_1, \ldots, \alpha_T$

while not converged:

for $t = 1, \ldots, T$ :

$$\alpha_t = \arg\min_{\alpha_t} J(\alpha_1, \ldots, \alpha_T)$$
COORDINATE ASCENT VARIATIONAL INFERENCE (CAVI)
1. **Goal**: estimate $p_\alpha(z \mid x)$  
   we assume this is intractable to compute exactly

2. **Idea**: approximate with another distribution $q_\theta(z) \approx p_\alpha(z \mid x)$ for each $x$

3. **Mean Field**: assume $q_\theta(z) = \prod_t q_t(z_t; \theta)$  
i.e., we decompose over variables  
other choices for the decomposition of $q_\theta(z)$ give rise to “structured mean field”

4. **Optimization Problem**: pick the $q$ that minimizes $\text{KL}(q \mid \mid p)$  
   $$\hat{\theta} = \arg\min_\theta \text{KL}(q_\theta(z) \mid \mid p_\alpha(z \mid x)) = \arg\max_\theta \text{ELBO}(q_\theta)$$
   $$\text{ELBO}(q_\theta) = E_{q_\theta(z)} \left[ \log p_\alpha(x, z) \right] - E_{q_\theta(z)} \left[ \log q_\theta(z) \right]$$
   $$\text{ELBO}(q_\theta) = E_{q_\theta(z)} \left[ \log \tilde{p}_\alpha(z \mid x) \right] - E_{q_\theta(z)} \left[ \log q_\theta(z) \right]$$

5. **Optimization Algorithm**: coordinate ascent  
i.e. pick the best $q_t(z_t)$ based on the other $\{q_s(z_s)\}_{s \neq t}$ being fixed

Choosing coordinate descent here yields the Coordinate Ascent Variational Inference (CAVI) algorithm
CAVI Algorithm

Coordinate Ascent Variational Inference (CAVI)

- here we assume a mean field approximation
- application of coordinate ascent to maximization of ELBO
- converges to a local optimum of the nonconvex ELBO objective

```
1: procedure CAVI(pα)
2: Let qθ(z) = \prod_{t=1}^{T} q_t(z_t)
3: while ELBO(qθ) has not converged do
4:   for t ∈ {1, ..., T} do
5:     Set qt(z_t) ∝ exp(E_{q_{¬t}}[log pα(z_t | z_{¬t}, x)])
6:     while keeping all \{qs(·)\}_{s \neq t} fixed
7:   Compute ELBO(qθ) = E_{qθ(z)} [log pα(x, z)] - E_{qθ}(z) [log qθ(z)]
8: return qθ
```
CAVI Algorithm

Similar to Belief Propagation: can be viewed as **message passing** where we update our **variable beliefs** based on what **neighbors** think it should be

1: **procedure** CAVI($p_{\alpha}$)
2: Let $q_{\theta}(z) = \prod_{t=1}^{T} q_{t}(z_{t})$
3: while ELBO($q_{\theta}$) has not converged do
4: for $t \in \{1, \ldots, T\}$ do
5: Set $q_{t}(z_{t}) \propto \exp(E_{q_{\neg t}}[\log p_{\alpha}(z_{t} | z_{\neg t}, x)])$
6: while keeping all $\{q_{s}(\cdot)\}_{s \neq t}$ fixed
7: Compute ELBO($q_{\theta}$) = $E_{q_{\theta}(z)} [\log p_{\alpha}(x, z)] - E_{q_{\theta}(z)} [\log q_{\theta}(z)]$
8: return $q_{\theta}$

Unlike Gibbs Sampling:
• we compute an entire distribution (instead of sampling a value)
• we condition on variable marginals (instead of on variable assignment)
Variational Inference

Whiteboard

– Computing marginals from a trained mean field approximation
EXAMPLE: CAVI FOR DISCRETE FACTOR GRAPH
CAVI for a Discrete Factor Graph

\[ p_\alpha(z \mid x) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \psi_c(z_c, x) \]
\[ q_\theta(z) = \prod_{t=1}^{T} q_t(z_t) \]

1: \textbf{procedure} CAVI(p_\alpha) 
2: \hspace{1em} Let \( q_\theta(z) = \prod_{t=1}^{T} q_t(z_t) \) \hspace{1em} \( \triangleright \) Mean field approx. 
3: \hspace{1em} \textbf{while} \ ELBO(q_\theta) \hspace{1em} \textbf{has not converged} \hspace{1em} \textbf{do} 
4: \hspace{2em} \textbf{for} \( t \in \{1, \ldots, T\} \) \hspace{1em} \textbf{do} \hspace{1em} \( \triangleright \) For each variable 
5: \hspace{3em} Set \( q_t(z_t) \propto \exp(\mathcal{E}_{q_{\neg t}}[\log p_\alpha(z_t \mid z_{\neg t}, x)]) \)
6: \hspace{3em} while keeping all \( \{q_s(\cdot)\}_{s \neq t} \) fixed 
7: \hspace{1em} Compute \( \text{ELBO}(q_\theta) = E_{q_\theta(z)}[\log p_\alpha(x, z)] - E_{q_\theta(z)}[\log q_\theta(z)] \)
8: \hspace{1em} \textbf{return} \( q_\theta \)

\[ \Rightarrow q_t(z_t) \propto \exp \left( \sum_{z_{MB(z_t)} \in MB(z_t)} \prod_{s \in MB(z_t)} q_s(z_s) \log \prod_{c \in N(z_t)} \psi_c(z_c) \right) \]

efficiently computed assuming number of neighbors \( N(z_t) \) is not too large
CAVI as Message Passing

Case 1: One Neighbor

\[
q_1(z_1) = \frac{\exp(0.08 + 0.16)}{Z} \\
p = \frac{\exp(2.4 + 0)}{Z} \\
n = \frac{\exp(0.8 + 0.2)}{Z}
\]

CAVI message passing differs from BP in several ways:
- the beliefs are normalized (i.e. beliefs = marginals)
- no messages to factors (i.e. all messages are directly to a variable)
- matrix-vector product is exponentiated and normalized

\[
q_t(z_t) \propto \exp \left( \sum_{z_{MB}(z_t)} \prod_{s \in MB(z_t)} q_s(z_s) \log \prod_{c \in N(z_t)} \psi_c(z_c) \right)
\]
**Sum-Product Belief Propagation**

**Factor Message**

\[ \mu_{\alpha \rightarrow i}(x_i) = \sum_{x_\alpha : x_\alpha[i] = x_i} \psi_\alpha(x_\alpha) \prod_{j \in \mathcal{N}(\alpha) \setminus i} \mu_{j \rightarrow \alpha}(x_\alpha[i]) \]
Case 2: Two Neighbors

\[
q_t(z_t) \propto \exp \left( \sum_{z_{MB(z_t)}} \prod_{s \in MB(z_t)} q_s(z_s) \log \prod_{c \in N(z_t)} \psi_c(z_c) \right)
\]
For a pairwise MRF, we have the following simplified the update rules:

\[
\mu_{s \to t}(z_t) = \sum_{z_s} q_s(z_s) \psi_{s,t}(z_s, z_t)
\]

\[
q_t(z_t) \propto \exp \left( \prod_{s \in \text{MB}(z_t)} \mu_{s \to t}(z_t) \right)
\]
Variational Inference

**Whiteboard**

– Computing the CAVI update
  • Multinomial full conditionals
– Example: two variable factor graph
  • Joint distribution
  • Mean Field Variational Inference
  • Gibbs Sampling
Q2: What qeeshins do you have?

Q&A