Mean Field Variational Inference
Q:
The parameters of a K-dimensional Dirichlet(\(\alpha\)) are a vector \(\alpha\) of length K, so why are Dirichlet parameters sometimes given as a scalar? For example...

“We use a Dirichlet prior with parameter \(\alpha = 0.1\).”

A:
Great question!

A K-dimensional Dirichlet prior is said to be symmetric if all the values in the vector \(\alpha\) are the same, i.e. for all \(k\), \(\alpha_k = c\) where \(c\) is a scalar constant.

We sometimes call this restricted version the symmetric Dirichlet distribution.
Reminders

• Exam Rubrics and Exam Viewings

• Homework 4: MCMC
  – Out: Mon, Oct 24
  – Due: Fri, Nov 3 at 11:59pm

• Homework 5: MCMC
  – Out: Mon, Oct 24
  – Due: Fri, Nov 3 at 11:59pm
Reminders

Happy Halloween!
SEMANTIC SEGMENTATION
Case Study: Image Segmentation

- Image segmentation (FG/BG) by modeling of interactions btw RVs
  - Images are noisy.
  - Objects occupy continuous regions in an image.

\[ Y^* = \arg \max_{y \in \{0,1\}^n} \left( \sum_{i \in S} V_i(y_i, X) + \sum_{i \in S} \sum_{j \in N_i} V_{i,j}(y_i, y_j) \right). \]

**Input image**  **Pixel-wise separate optimal labeling**  **Locally-consistent joint optimal labeling**

\[ Y: \text{labels} \quad X: \text{data (features)} \quad S: \text{pixels} \quad N_i: \text{neighbors of pixel } i \]
Grid CRF

- Suppose we want to image segmentation using a grid model
Grid CRF

• Suppose we want to image segmentation using a grid model

Assuming our label set is only \{foreground, background\} each factor is a table with $2^2$ entries.
Grid CRF

• Suppose we want to image segmentation using a grid model
• What happens when we run variable elimination?
Grid CRF

• Suppose we want to image segmentation using a grid model
• What happens when we run variable elimination?
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?
Grid CRF

• Suppose we want to image segmentation using a grid model
• What happens when we run variable elimination?
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?
Grid CRF

• Suppose we want to image segmentation using a grid model
• What happens when we run variable elimination?

This new factor has $2^5$ entries
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?

For an $M \times M$ grid the new factor has $2^M$ entries
Grid CRF

- Suppose we want to image segmentation using a grid model
- What happens when we run variable elimination?

For an $M \times M$ grid the new factor has $2^M$ entries

In general, for high treewidth graphs like this, we turn to approximate inference (which we’ll cover soon!)
Grid CRF

• Suppose we want to image segmentation using a grid model
• What happens when we run variable elimination?
• Can we instead run belief propagation to do exact inference?
HIGH-LEVEL INTRO TO VARIATIONAL INFERENC
Variational Inference

**Problem:**
- For observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$, estimating the posterior $p(\mathbf{z} \mid \mathbf{x})$ is intractable

Narrative adapted from Jason Eisner’s High-Level Explanation of VI:
[https://www.cs.jhu.edu/~jason/tutorials/variational.html](https://www.cs.jhu.edu/~jason/tutorials/variational.html)
Problem:

– For observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$, estimating the posterior $p(\mathbf{z} \mid \mathbf{x})$ is intractable.
Variational Inference

**Problem:**
- For observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$, estimating the posterior $p(\mathbf{z} | \mathbf{x})$ is intractable.
- For training data $\mathbf{x}$ and parameters $\mathbf{z}$, estimating the posterior $p(\mathbf{z} | \mathbf{x})$ is intractable.

Narrative adapted from Jason Eisner’s High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html
Variational Inference

Problem:
- For observed variables $\mathbf{x}$ and latent variables $\mathbf{z}$, estimating the posterior $p(\mathbf{z} | \mathbf{x})$ is intractable
- For training data $\mathbf{x}$ and parameters $\mathbf{z}$, estimating the posterior $p(\mathbf{z} | \mathbf{x})$ is intractable

Solution:
- Approximate $p(\mathbf{z} | \mathbf{x})$ with a simpler $q(\mathbf{z})$
- Typically $q(\mathbf{z})$ has more independence assumptions than $p(\mathbf{z} | \mathbf{x})$ – fine b/c $q(\mathbf{z})$ is tuned for a specific $\mathbf{x}$
- **Key idea**: pick a single $q(\mathbf{z})$ from some family $Q$ that best approximates $p(\mathbf{z} | \mathbf{x})$

Narrative adapted from Jason Eisner’s High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html
Variational Inference

**Terminology:**

- $q(z)$: the **variational approximation**
- $Q$: the **variational family**
- Usually $q_\theta(z)$ is parameterized by some $\theta$ called **variational parameters**
- Usually $p_\alpha(z \mid x)$ is parameterized by some fixed $\alpha$ – we’ll call them the parameters

**Example Algorithms:**

- mean-field variational inference
- loopy belief propagation
- tree-reweighted belief propagation
- expectation propagation

Narrative adapted from Jason Eisner’s High-Level Explanation of VI: [https://www.cs.jhu.edu/~jason/tutorials/variational.html](https://www.cs.jhu.edu/~jason/tutorials/variational.html)
Variational Inference

Is this trivial?

– Note: We are not defining a new distribution simple $q_\theta(z \mid x)$, there is one simple $q_\theta(z)$ for each $p_\alpha(z \mid x)$
– Consider the MCMC equivalent of this:
  • you could draw samples $z^{(i)} \sim p(z \mid x)$
  • then train some simple $q_\theta(z)$ on $z^{(1)}$, $z^{(2)}$, ..., $z^{(N)}$
  • hope that the sample adequately represents the posterior for the given $x$
– How is VI different from this?
  • VI doesn’t require sampling
  • VI is fast and deterministic
  • Why? b/c we choose an objective function (KL divergence) that defines which $q_\theta$ best approximates $p_\alpha$, and exploit the special structure of $q_\theta$ to optimize it

Narrative adapted from Jason Eisner’s High-Level Explanation of VI: https://www.cs.jhu.edu/~jason/tutorials/variational.html
Variational Inference

V.I. offers a new design decision

– Choose the distribution $p_{\alpha}(z \mid x)$ that you really want, i.e. don’t just simplify it to make it computationally convenient

– Then design a the structure of another distribution $q_{\theta}(z)$ such that V.I. is efficient

Narrative adapted from Jason Eisner’s High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html
TYPES OF VARIATIONAL APPROXIMATIONS
Mean Field Approximation

The mean field approximation assumes our variational approximation $q_\theta(z)$ treats each variable as independent.

$$p_\alpha(z \mid x) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \psi_c(z_c, x)$$

$$q_\theta(z) = \prod_{t=1}^T q_t(z_t)$$
Mean Field Approximation

The **mean field approximation** assumes our variational approximation $q_{\theta}(z)$ treats each variable as independent.

**Ising Model**

\[
p_{\alpha}(z \mid x) = \frac{1}{Z(x)} \prod_{c \in C} \psi_c(z_c, x)
\]

\[
q_{\theta}(z) = \prod_{t=1}^{T} q_t(z_t)
\]
Structured Mean Field

- If q is not a mean-field approximation, but decomposes over “blocks” of variables, then we have the **Structured Mean Field algorithm**

- Connection to related algorithms:
  - This is analogous to **Blocked** Gibbs Sampling
  - This is analogous to **Generalized** Belief Propagation
  - The names here (**Structured**, **Blocked**, **Generalized**) are different b/c they were invented by different people and no-one thought to rename them all “Blocked”
Structured Mean Field

• We can also apply more general forms of mean field approximations (involving clusters) to the Ising model:

• Instead of making all latent variables independent (i.e. naïve mean field, previous example), clusters of (disjoint) latent variables are independent.
Structured Mean Field

- For a factorial HMM, we could decompose into chains
Collapsed vs. Uncollapsed V.I.

Just as we had collapsed and uncollapsed Gibbs samplers for LDA...

...we can have collapsed and uncollapsed variational inference for LDA
**Collapsed vs. Uncollapsed V.I.**

**Latent Dirichlet Allocation (LDA)**

- Uncollapsed Variational Inference, aka. Explicit V.I. (original distribution)

![Diagram of LDA]

- Dirichlet
- Document-specific topic distribution
- Topic assignment
- Observed word
- Approximate with $q$
- Topic
- Dirichlet
- $\alpha$
- $\theta_m$
- $z_{mn}$
- $x_{mn}$
- $\phi_k$
- $\beta$
- $N_m$
- $M$
- $K$
Collapsed vs. Uncollapsed V.I.

Latent Dirichlet Allocation (LDA)
• Uncollapsed Variational Inference, aka. Explicit V.I. (mean field variational approximation)

$$\theta_m$$
Document-specific topic distribution

$$z_{mn}$$
Topic assignment

$$\phi_k$$
Topic

$$N_m$$
$$M$$

$$K$$
Collapsed vs. Uncollapsed V.I.

Latent Dirichlet Allocation (LDA)
• Collapsed Variational Inference (original distribution)

![Diagram showing the graphical model for the SCTM](Image)
Collapsed vs. Uncollapsed V.I.

Latent Dirichlet Allocation (LDA)
• Collapsed Variational Inference (mean field variational approximation)
MEAN FIELD VARIATIONAL INFERENCE
Side Note

Contrast of three variational inference techniques:

1. Mean field variational inference minimizes $KL(q \parallel p)$
2. Expectation propagation minimizes $KL(p \parallel q)$
3. Loopy Belief Propagation minimizes the Bethe Free Energy

We are focused here on $KL(q \parallel p)$
KL Divergence

- **Definition:** for two distributions $q(x)$ and $p(x)$ over $x \in \mathcal{X}$, the KL Divergence is:

  $$KL(q\|p) = E_{q(x)} \left[ \log \frac{q(x)}{p(x)} \right] = \left\{ \begin{array}{ll} \sum_x q(x) \log \frac{q(x)}{p(x)} \\ \int_x q(x) \log \frac{q(x)}{p(x)} \, dx \end{array} \right.$$  

- **Properties:**
  - $KL(q\|p)$ measures the **proximity** of two distributions $q$ and $p$
  - KL is **not** symmetric: $KL(q\|p) \neq KL(p\|q)$
  - KL is minimized when $q(x) = p(x)$ for all $x \in \mathcal{X}$
KL divergence

**Understanding the Behavior of KL as an objective function**

**Example 1:** Keeping all else constant, consider the effect of a particular $x'$ on $KL(q \parallel p)$

<table>
<thead>
<tr>
<th></th>
<th>$q(x')$</th>
<th>$p(x')$</th>
<th>$q(x') \log(q(x')/p(x'))$</th>
<th>effect on $KL(q \parallel p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9</td>
<td>0.9</td>
<td>0</td>
<td>no increase</td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>0.1</td>
<td>1.97</td>
<td>big increase</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
<td>0.9</td>
<td>-0.21</td>
<td>little decrease</td>
</tr>
<tr>
<td>4</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>little decrease</td>
</tr>
</tbody>
</table>

**Example 2:** Which $q$ distribution minimizes $KL(q \parallel p)$?

\[
p = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \quad q^{(1)} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad q^{(2)} = \begin{bmatrix} 0.7 \\ 0.2 \\ 0.1 \end{bmatrix} \quad q^{(3)} = \begin{bmatrix} 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}
\]

Q: If we’re minimizing $KL$, why not return $q^{(3)}$?
A: Because it’s not a distribution!

$q^{(2)}$ minimizes KL
KL Divergence

Understanding the Behavior of KL as an objective function

Example 3: Which q distribution minimizes KL(q || p)?

\[ p(x) = \mathcal{N}(\mu = [0, 0]^T, \Sigma) \]
\[ q(x_1, x_2) = \mathcal{N}_1(x_1 \mid \mu_1, \sigma_1^2)\mathcal{N}_2(x_2 \mid \mu_2, \sigma_2^2) \]
Two Cases for Intractability

• **Case 1:**

given a **joint distribution** \( p(x, z) \)

\[ p(z \mid x) = \frac{p(x, z)}{p(x)} \]

we assume \( p(x) \) is intractable

• **Case 2:**

give **factor graph** and potentials

\[ p(z \mid x) = \frac{\tilde{p}(x, z)}{Z(x)} \]

we assume \( Z(x) \) is intractable
Mean Field Approximation

The mean field approximation assumes our variational approximation \( q_\theta(z) \) treats each variable as independent.

\[
p_\alpha(z \mid x) = \frac{1}{Z(x)} \prod_{c \in C} \psi_c(z_c, x)
\]

\[
q_\theta(z) = \prod_{t=1}^{T} q_t(z_t)
\]
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 $KL(q \parallel p)$
   $$\hat{q}(z) = \arg\min_{q(z) \in \mathcal{Q}} KL(q(z) \parallel p(z \mid x))$$
   $$\hat{\theta} = \arg\min_{\theta \in \Theta} KL(q_\theta(z) \parallel p_\alpha(z \mid x))$$

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

- **Question**: How do we minimize KL?

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} KL(q_{\theta}(z) \parallel p_\alpha(z \mid x))
\]

- **Answer #1**: Oh no! We can’t even compute this KL.

Why we can’t compute KL...

\[
KL(q(z) \parallel p(z \mid x)) = E_{q(z)} \left[ \log \left( \frac{q(z)}{p(z \mid x)} \right) \right]
\]

\[
= E_{q(z)} \left[ \log q(z) \right] - E_{q(z)} \left[ \log p(z \mid x) \right]
\]

\[
= E_{q(z)} \left[ \log q(z) \right] - E_{q(z)} \left[ \log p(x, z) \right] + E_{q(z)} \left[ \log p(x) \right]
\]

we have the same problem with an intractable data likelihood p(x) or an intractable partition function Z(x)

we assumed this is intractable to compute!
Optimizing KL Divergence

• **Question**: How do we minimize KL?

\[
\hat{\theta} = \arg\min_{\theta \in \Theta} \text{KL}(q_{\theta}(z) \mid\mid p_\alpha(z \mid x))
\]

• **Answer #1**: Oh no! We can’t even compute this KL.

Why we can’t compute KL...

\[
\text{KL}(q(z) \mid\mid p(z \mid x)) = E_{q(z)} \left[ \log \left( \frac{q(z)}{p(z \mid x)} \right) \right] = E_{q(z)} [\log q(z)] - E_{q(z)} [\log p(z \mid x)] = E_{q(z)} [\log q(z)] - E_{q(z)} [\log \tilde{p}(z \mid x)] + E_{q(z)} [\log Z(x)] = E_{q(z)} [\log q(z)] - E_{q(z)} [\log \tilde{p}(z \mid x)] + \log Z(x)
\]

we have the same problem with an intractable data likelihood \( p(x) \) or an intractable partition function \( Z(x) \)

we assumed this is intractable to compute!
Optimizing KL Divergence

• **Question:** How do we minimize KL?

\[ \hat{\theta} = \arg\min_{\theta \in \Theta} \text{KL}(q_\theta(z) \| p_\alpha(z \mid x)) \]

• **Answer #2:** We don’t need to compute this KL. We can instead maximize the ELBO (i.e. Evidence Lower BOund)

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

Here is why...

\[
\theta = \arg\min_{\theta} \text{KL}(q_\theta(z) \| p_\alpha(z \mid x))
\]

\[
= \arg\min_{\theta} E_{q_\theta(z)} \left[ \log q_\theta(z) \right] - E_{q_\theta(z)} \left[ \log p_\alpha(x, z) \right] + \log p_\alpha(x)
\]

\[
= \arg\min_{\theta} E_{q_\theta(z)} \left[ \log q_\theta(z) \right] - E_{q_\theta(z)} \left[ \log p_\alpha(x, z) \right]
\]

\[
= \arg\max_{\theta} \text{ELBO}(q_\theta)
\]

dropping the intractable term gives the ELBO
Optimizing KL Divergence

- **Question**: How do we minimize KL?

  \[ \hat{\theta} = \arg \min_{\theta \in \Theta} \text{KL}(q_{\theta}(z) \mid\mid p_{\alpha}(z \mid x)) \]

- **Answer #2**: We don’t need to compute this KL. We can instead maximize the ELBO (i.e., Evidence Lower BOund).

  \[ \text{ELBO}(q_{\theta}) = E_{q_{\theta}(z)} \left[ \log \tilde{p}_{\alpha}(z \mid x) \right] - E_{q_{\theta}(z)} \left[ \log q_{\theta}(z) \right] \]

  The ELBO for a UGM

Here is why...

\[ \theta = \arg \min_{\theta} \text{KL}(q_{\theta}(z) \mid\mid p_{\alpha}(z \mid x)) \]

\[ = \arg \min_{\theta} E_{q_{\theta}(z)} \left[ \log q_{\theta}(z) \right] - E_{q_{\theta}(z)} \left[ \log \tilde{p}_{\alpha}(z \mid x) \right] + \log Z_{\alpha}(x) \]

\[ = \arg \min_{\theta} E_{q_{\theta}(z)} \left[ \log q_{\theta}(z) \right] - E_{q_{\theta}(z)} \left[ \log \tilde{p}_{\alpha}(z \mid x) \right] \]

\[ = \arg \max_{\theta} \text{ELBO}(q_{\theta}) \]

dropping the intractable term gives the ELBO
ELBO as Objective Function

What does maximizing ELBO($q_\theta$) accomplish?

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

1. The first expectation is high if $q_\theta$ puts probability mass on the same values of $z$ that $p_\alpha$ puts probability mass

2. The second term is the entropy of $q_\theta$ and the entropy will be high if $q_\theta$ spreads its probability mass evenly
ELBO as lower bound

• *For a DGM:*  
  – ELBO(q) is a lower bound for log $p(x)$

• *For a UGM:*  
  – ELBO(q) is a lower bound for log $Z(x)$

**Takeaway:** in variational inference, we find the $q$ that gives the **tightest bound** on the normalization constant for $p(z \mid x)$
Variational Inference

Whiteboard

– Evidence Lower Bound (ELBO)
– ELBO’s relation to log p(x)