Topic Modeling
+
Variational Inference
Reminders

• Project Proposal
  – Due: Wed, Mar. 31 at 11:59pm

• Homework 4: MCMC
  – Out: Wed, Mar. 24
  – Due: Wed, Apr. 7 at 11:59pm
Outline

• Applications of Topic Modeling
• Review: Latent Dirichlet Allocation (LDA)
  1. Beta-Bernoulli
  2. Dirichlet-Multinomial
  3. Dirichlet-Multinomial Mixture Model
  4. LDA
• Bayesian Inference for Parameter Estimation
  – Exact inference
  – EM
  – Monte Carlo EM
  – Gibbs sampler
  – Collapsed Gibbs sampler
• Extensions of LDA
  – Correlated topic models
  – Dynamic topic models
  – Polylingual topic models
  – Supervised LDA
EXTENSIONS OF LDA
Extensions to the LDA Model

- Correlated topic models
  - Logistic normal prior over topic assignments
- Dynamic topic models
  - Learns topic changes over time
- Polylingual topic models
  - Learns topics aligned across multiple languages
  
\[
\begin{align*}
\sum_{\beta_k} & \quad \eta_d \quad Z_{d,n} \quad W_{d,n} \\
\mu & \quad \alpha \quad \theta_d \quad \beta_{k,1} \quad \beta_{k,2} \quad \ldots \quad \beta_{k,T} \\
\phi^1 \quad \phi^L & \quad \beta^1 \quad \beta^L
\end{align*}
\]
Correlated Topic Models

- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*. 
Correlated Topic Models

- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics.*
• The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).

• The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

\[
\mathbf{X} \sim \mathcal{N}_K(\mu, \Sigma) \\
\theta_i \propto \exp\{x_i\}.
\]
Correlated Topic Models

- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a “map” of topics and how they are related
- Provides a better fit to text data, but computation is more complex
Correlated Topic Models

(Blei & Lafferty, 2004)

Slide from David Blei, MLSS 2012
Dynamic Topic Models

High-level idea:
• Divide the documents up by year
• Start with a separate topic model for each year
• Then add a dependence of each year on the previous one

(Blei & Lafferty, 2006)
Dynamic Topic Models

(Blei & Lafferty, 2006)

1789

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

2009

AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

• LDA assumes that the order of documents does not matter.
• Not appropriate for sequential corpora (e.g., that span hundreds of years)
• Further, we may want to track how language changes over time.
• Dynamic topic models let the topics *drift* in a sequence.

Slide from David Blei, MLSS 2012
Dynamic Topic Models

Generative Story

1. Draw topics $\beta_t | \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   a. Draw $\eta \sim \mathcal{N}(\alpha_t, \alpha^2 I)$
   b. For each word:
      i. Draw $Z \sim \text{Mult}(\pi(\eta))$.
      ii. Draw $W_{t,d,n} \sim \text{Mult}(\pi(\beta_{t,z}))$.

\begin{align*}
\pi(\beta_{k,t})_w &= \frac{\exp(\beta_{k,t}^w)}{\sum_w \exp(\beta_{k,t}^w)}.
\end{align*}
Dynamic Topic Models

Top ten most likely words in a “drifting” topic shown at 10-year increments

(Blei & Lafferty, 2006)
Dynamic Topic Models

Posterior estimate of word frequency as a function of year for three words each in two separate topics:

"Theoretical Physics"

FORCERELATIVITYLASER

"Neuroscience"

NERVENEROXYGENNEURON

(Blei & Lafferty, 2006)
Polylingual Topic Models

• **Data Setting:** Comparable versions of each document exist in multiple languages (e.g. the Wikipedia article for “Barak Obama” in twelve languages)

• **Model:** Very similar to LDA, except that the topic assignments, \( z \), and words, \( w \), are sampled separately for each language.
Polylingual Topic Models

**Topic 1 (twelve languages)**

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CY</td>
<td>sadwrn blaned gallair at lloeren mytholeg</td>
</tr>
<tr>
<td>DE</td>
<td>space nasa sojus flug mission</td>
</tr>
<tr>
<td>EL</td>
<td>διαστημικό sts nasa αγγλ small</td>
</tr>
<tr>
<td>EN</td>
<td><strong>space mission launch satellite nasa spacecraft</strong></td>
</tr>
<tr>
<td>FA</td>
<td>فضائيي ماموريت ناسا مدار فضانورد ماهوره</td>
</tr>
<tr>
<td>FI</td>
<td>sojuz nasa apollo ensimmäinen space lento</td>
</tr>
<tr>
<td>FR</td>
<td>spatiale mission orbite mars satellite spatial</td>
</tr>
<tr>
<td>HE</td>
<td>התכלת האורח תכלל צוור א תוכנית</td>
</tr>
<tr>
<td>IT</td>
<td>spaziale missione programma space sojuz stazione</td>
</tr>
<tr>
<td>PL</td>
<td>misja kosmicznej stacji misji space nasa</td>
</tr>
<tr>
<td>RU</td>
<td>космический союз космического спутник станции</td>
</tr>
<tr>
<td>TR</td>
<td>uzay soyuz ay uzaya salyut sovyetler</td>
</tr>
</tbody>
</table>

(Mimno et al., 2009)
Polylingual Topic Models

Topic 2 (twelve languages)

| CY | sbaen madrid el la josé sbaeneg |
| DE | de spanischer spanischen spanien madrid la |
| EL | ισπανίας ισπανία de ισπανός ντε μαδρίτη |
| EN | de spanish spain la madrid y |
| FA | ترین اسپانیا اسپانیایی کویا مادرید |
| FI | espanja de espanjan madrid la real |
| FR | espagnol espagne madrid espagnole juan y |
| HE | ספרד ספרדיות de מדריד הספרדים קובה |
| IT | de spagna spagnolo spagnola madrid el |
| PL | de hiszpański hiszpanii la juan y |
| RU | де мадрид испания испания испанский de |
| TR | ispanya ispanyol madrid la küba real |
### Polylingual Topic Models

#### Topic 3 (twelve languages)

<table>
<thead>
<tr>
<th>Language</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>CY</td>
<td>bardd gerddi iaiith beirdd fardd gymraeg</td>
</tr>
<tr>
<td>DE</td>
<td>dichter schriftsteller literatur gedichte gedicht werk</td>
</tr>
<tr>
<td>EL</td>
<td>ποιητής ποίηση ποιητή έργο ποιητές ποιήματα</td>
</tr>
<tr>
<td>EN</td>
<td>poet poetry literature literary poems poem</td>
</tr>
<tr>
<td>FA</td>
<td>شاعر شعر أدبيات فارسي أدبي آثار</td>
</tr>
<tr>
<td>FI</td>
<td>runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi</td>
</tr>
<tr>
<td>FR</td>
<td>poète écrivain littérature poésie littéraire ses</td>
</tr>
<tr>
<td>HE</td>
<td>מشور סופר ספרי ספרות המשורר</td>
</tr>
<tr>
<td>IT</td>
<td>poeta letteratura poesia opere versi poema</td>
</tr>
<tr>
<td>PL</td>
<td>poeta literatury poezji pisarz in jego</td>
</tr>
<tr>
<td>RU</td>
<td>поэт его писатель литературы поэзии драматург</td>
</tr>
<tr>
<td>TR</td>
<td>şair edebiyat şiir yazar edebiyatı adlı</td>
</tr>
</tbody>
</table>

(Mimno et al., 2009)
Polylingual Topic Models

Size of each square represents proportion of tokens assigned to the specified topic.

**Figure 8:** Squares represent the proportion of tokens in each language assigned to a topic. The left topic, *world ski km won*, centers around Nordic countries. The center topic, *actor role television actress*, is relatively uniform. The right topic, *ottoman empire khan byzantine*, is popular in all languages but especially in regions near Istanbul.

**Table 5:** Percent of English query documents for which the translation was in the top \( n \in \{1, 5, 10, 20\} \) documents by JS divergence between topic distributions. To reduce the effect of short documents we consider only document pairs where the query and target documents are longer than 100 words.

<table>
<thead>
<tr>
<th>Language</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>78.0</td>
<td>90.7</td>
<td>93.8</td>
</tr>
<tr>
<td>DE</td>
<td>76.6</td>
<td>90.0</td>
<td>93.4</td>
</tr>
<tr>
<td>EL</td>
<td>77.1</td>
<td>90.4</td>
<td>93.3</td>
</tr>
<tr>
<td>ES</td>
<td>81.2</td>
<td>92.3</td>
<td>94.8</td>
</tr>
<tr>
<td>FI</td>
<td>76.7</td>
<td>91.0</td>
<td>94.0</td>
</tr>
<tr>
<td>FR</td>
<td>80.1</td>
<td>91.7</td>
<td>94.3</td>
</tr>
<tr>
<td>IT</td>
<td>79.1</td>
<td>91.2</td>
<td>94.1</td>
</tr>
<tr>
<td>NL</td>
<td>76.6</td>
<td>90.1</td>
<td>93.4</td>
</tr>
<tr>
<td>PL</td>
<td>80.8</td>
<td>92.0</td>
<td>94.7</td>
</tr>
<tr>
<td>PT</td>
<td>80.4</td>
<td>92.1</td>
<td>94.9</td>
</tr>
<tr>
<td>SV</td>
<td>80.4</td>
<td>92.1</td>
<td>94.9</td>
</tr>
</tbody>
</table>

In this section, we explore two questions relating to comparable text corpora and polylingual topic modeling. First, we explore whether comparable document tuples support the alignment of fine-grained topics, as demonstrated earlier using parallel documents. This property is useful for building machine translation systems as well as for human readers who are either learning new languages or analyzing texts in languages they do not know. Second, because comparable texts may not use exactly the same topics, it becomes crucially important to be able to characterize differences in topic prevalence at the document level (do different languages have different perspectives on the same article?) and at the language-wide level (which topics do particular languages focus on?).

**5.1 Data Set**
We downloaded XML copies of all Wikipedia articles in twelve different languages: Welsh, German, Greek, English, Farsi, Finnish, French, Hebrew, Italian, Polish, Russian and Turkish. These versions of Wikipedia were selected to provide a diverse range of language families, geographic areas, and quantities of text. We preprocessed the data by removing tables, references, images and info-boxes. We dropped all articles in non-English languages that did not link to an English article. In the English version of Wikipedia we dropped all articles that were not linked to by any other language in our set. For efficiency, we truncated each article to the nearest word after 1000 characters and dropped the 50 most common word types in each language. Even with these restrictions, the size of the corpus is 148.5 million words.

We present results for a PLTM with 400 topics. 1000 Gibbs sampling iterations took roughly four days on one CPU with current hardware.

**5.2 Which Languages Have High Topic Divergence?**
As with EuroParl, we can calculate the Jensen-Shannon divergence between pairs of documents within a comparable document tuple. We can then average over all such document-document divergences for each pair of languages to get an overall “disagreement” score between languages. Interestingly, we find that almost all languages in our corpus, including several pairs that have historically been in conflict, show average JS divergences of between approximately 0.08 and 0.12 for \( T = 400 \), consistent with our findings for EuroParl translations. Subtle differences of sentiment may be below the granularity of the model.
Supervised LDA

LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?

Many data are paired with response variables.
- User reviews paired with a number of stars
- Web pages paired with a number of “likes”
- Documents paired with links to other documents
- Images paired with a category

Supervised LDA are topic models of documents and responses. They are fit to find topics predictive of the response.
Supervised LDA

1. Draw topic proportions $\theta | \alpha \sim \text{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n | \theta \sim \text{Mult}(\theta)$.
   - Draw word $w_n | z_n, \beta_{1:K} \sim \text{Mult}(\beta_{z_n})$.
3. Draw response variable $y | z_{1:N}, \eta, \sigma^2 \sim \mathcal{N}(\eta^T \bar{z}, \sigma^2)$, where
   $$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Slide from David Blei, MLSS 2012
Summary: Topic Modeling

• The Task of Topic Modeling
  – Topic modeling enables the analysis of large (possibly unannotated) corpora
  – Applicable to more than just bags of words
  – Extrinsic evaluations are often appropriate for these unsupervised methods

• Constructing Models
  – LDA is comprised of simple building blocks (Dirichlet, Multinomial)
  – LDA itself can act as a building block for other models

• Approximate Inference
  – Many different approaches to inference (and learning) can be applied to the same model
What if we don’t know the number of topics, $K$, ahead of time?

**Solution:** Bayesian Nonparametrics

- New modeling constructs:
  - Chinese Restaurant Process (Dirichlet Process)
  - Indian Buffet Process
- e.g. an **infinite number of topics** in a finite amount of space
Summary: Approximate Inference

- Markov Chain Monte Carlo (MCMC)
  - Metropolis-Hastings, Gibbs sampling, Hamiltonian MCMC, slice sampling, etc.
- Variational inference
  - Minimizes $KL(q||p)$ where $q$ is a simpler graphical model than the original $p$
- Loopy Belief Propagation
  - Belief propagation applied to general (loopy) graphs
- Expectation propagation
  - Approximates belief states with moments of simpler distributions
- Spectral methods
  - Uses tensor decompositions (e.g. SVD)
Slice Sampling, Hamiltonian Monte Carlo

MCMC (AUXILIARY VARIABLE METHODS)
Auxiliary variables

The point of MCMC is to marginalize out variables, but one can introduce more variables:

\[
\int f(x)P(x) \, dx = \int f(x)P(x, v) \, dx \, dv
\]

\[
\approx \frac{1}{S} \sum_{s=1}^{S} f(x^{(s)}), \quad x, v \sim P(x, v)
\]

We might want to do this if

- \(P(x|v)\) and \(P(v|x)\) are simple

- \(P(x, v)\) is otherwise easier to navigate
Slice Sampling

• Motivation:
  – Want **samples** from \( p(x) \) and don’t know the normalizer \( Z \)
  – Choosing a proposal at the correct **scale** is difficult

• Properties:
  – *Similar to Gibbs Sampling*: one-dimensional transitions in the state space
  – *Similar to Rejection Sampling*: (asymptotically) draws samples from the **region under the curve**

\[ \tilde{p}(x) \]

– An MCMC method with an **adaptive proposal**
Slice sampling idea

Sample point uniformly under curve $\tilde{P}(x) \propto P(x)$

This is just an auxiliary-variable Gibbs Sampler!

Problem: Sampling from the conditional $p(x | u)$ might be infeasible.

$p(u|x) = \text{Uniform}[0, \tilde{P}(x)]$

$p(x|u) \propto \begin{cases} 1 & \tilde{P}(x) \geq u \\ 0 & \text{otherwise} \end{cases} = \text{"Uniform on the slice"}$
Slice Sampling

Figure 29.16

Each panel is labelled by the steps of the algorithm that are executed in it. At step 1, $P^\ast(x)$ is evaluated at the current point $x$. At step 2, a vertical coordinate is selected giving the point $(x, u')$ shown by the box; at steps 3a-c, an interval of size $w$ containing $(x, u')$ is created at random. At step 3d, $P^\ast$ is evaluated at the left end of the interval and is found to be larger than $u'$, so a step to the left of size $w$ is made. At step 3e, $P^\ast$ is evaluated at the right end of the interval and is found to be smaller than $u'$, so no stepping out to the right is needed. When step 3d is repeated, $P^\ast$ is found to be smaller than $u'$, so the stepping out halts. At step 5a a point is drawn from the interval, shown by a ◦. Step 6 establishes that this point is above $P^\ast$ and step 8 shrinks the interval to the rejected point in such a way that the original point $x$ is still in the interval. When step 5 is repeated, the new coordinate $x'$ (which is to the right-hand side of the interval) gives a value of $P^\ast$ greater than $u'$, so this point $x'$ is the outcome at step 7.
Slice Sampling

Figure 29.16: Slice sampling. Each panel is labelled by the steps of the algorithm that are executed in it. At step 1, \( P^* \) is evaluated at the current point \( x \). At step 2, a vertical coordinate is selected giving the point \((x, u')\) shown by the box; At steps 3a-c, an interval of size \( w \) containing \((x, u')\) is created at random. At step 3d, \( P^* \) is evaluated at the left end of the interval and is found to be larger than \( u' \), so a step to the left of size \( w \) is made. At step 3e, \( P^* \) is evaluated at the right end of the interval and is found to be smaller than \( u' \), so no stepping out to the right is needed. When step 3d is repeated, \( P^* \) is found to be smaller than \( u' \), so the stepping out halts. At step 5a, a point is drawn from the interval, shown by a ◯. Step 6 establishes that this point is above \( P^* \) and step 8 shrinks the interval to the rejected point in such a way that the original point \( x \) is still in the interval. When step 5 is repeated, the new coordinate \( x' \) (which is to the right-hand side of the interval) gives a value of \( P^* \) greater than \( u' \), so this point \( x' \) is the outcome at step 7.
Figure 29.16. Slice sampling. Each panel is labelled by the steps of the algorithm that are executed in it. At step 1, $P^\ast(x)$ is evaluated at the current point $x$. At step 2, a vertical coordinate is selected giving the point $(x, u')$ shown by the box; At steps 3a-c, an interval of size $w$ containing $(x, u')$ is created at random. At step 3d, $P^\ast$ is evaluated at the left end of the interval and is found to be larger than $u'$, so a step to the left of size $w$ is made. At step 3e, $P^\ast$ is evaluated at the right end of the interval and is found to be smaller than $u'$, so no stepping out to the right is needed. When step 3d is repeated, $P^\ast$ is found to be smaller than $u'$, so the stepping out halts. At step 5a, a point is drawn from the interval, shown by a ◦. Step 6 establishes that this point is above $P^\ast$ and step 8 shrinks the interval to the rejected point in such a way that the original point $x$ is still in the interval. When step 5 is repeated, the new coordinate $x'$ (which is to the right-hand side of the interval) gives a value of $P^\ast$ greater than $u'$, so this point $x'$ is the outcome at step 7.
Slice Sampling

**Goal:** sample \((x, u)\) given \((u^{(t)}, x^{(t)})\).

**Part 1:** Stepping Out

Sample interval \((x_l, x_r)\) enclosing \(x^{(t)}\).

Expand until endpoints are "outside" region under curve.

**Part 2:** Sample \(x\) (Shrinking)

Draw \(x\) from within the interval \((x_l, x_r)\), then accept or shrink.
Slice Sampling

**Goal:** sample \((x, u)\) given \((u^{(t)}, x^{(t)})\).

\[u \sim \text{Uniform}(0, p(x^{(t)})\]

**Part 1: Stepping Out**

Sample interval \((x_l, x_r)\) enclosing \(x^{(t)}\).

\[r \sim \text{Uniform}(u, w)\]

\[(x_l, x_r) = (x^{(t)} - r, x^{(t)} + w - r)\]

Expand until endpoints are ”outside” region under curve.

\[\text{while}(\tilde{p}(x_l) > u)\{x_l = x_l - w\}\]

\[\text{while}(\tilde{p}(x_r) > u)\{x_r = x_r + w\}\]

**Part 2: Sample \(x\) (Shrinking)**

Draw \(x\) from within the interval \((x_l, x_r)\), then accept or shrink.
Slice Sampling

**Goal:** sample \((x, u)\) given \((u^{(t)}, x^{(t)})\).

\[ u \sim \text{Uniform}(0, p(x^{(t)})) \]

**Part 1:** Stepping Out

Sample interval \((x_l, x_r)\) enclosing \(x^{(t)}\).

\[ r \sim \text{Uniform}(u, w) \]

\[ (x_l, x_r) = (x^{(t)} - r, x^{(t)} + w - r) \]

Expand until endpoints are "outside" region under curve.

while(\(\tilde{p}(x_l) > u\))\{\(x_l = x_l - w\}\}

while(\(\tilde{p}(x_r) > u\))\{\(x_r = x_r + w\}\}

**Part 2:** Sample \(x\) (Shrinking)

while(true) {

Draw \(x\) from within the interval \((x_l, x_r)\), then accept or shrink.

\[ x \sim \text{Uniform}(x_l, x_r) \]

if(\(\tilde{p}(x) > u\))\{break\}

else if(\(x > x^{(t)}\))\{\(x_r = x\)\}

else\{\(x_l = x\)\}

\}

\[ x^{(t+1)} = x, \quad u^{(t+1)} = u \]
Slice Sampling

Multivariate Distributions

– Resample each variable $x_i$ one-at-a-time (just like Gibbs Sampling)

– Does not require sampling from

$$p(x_i | \{x_j\}_{j \neq i})$$

– Only need to evaluate a quantity proportional to the conditional

$$p(x_i | \{x_j\}_{j \neq i}) \propto \tilde{p}(x_i | \{x_j\}_{j \neq i})$$
Hamiltonian Monte Carlo

- Suppose we have a distribution of the form:
  \[ p(x) = \frac{\exp\{-E(x)\}}{Z} \]
  where \( x \in \mathcal{R}^N \)

- We could use random-walk M-H to draw samples, but it seems a shame to discard gradient information \( \nabla_x E(x) \)

- If we can evaluate it, the gradient tells us where to look for high-probability regions!
Background: Hamiltonian Dynamics

Applications:

– Following the motion of atoms in a fluid through time
– Integrating the motion of a solar system over time
– Considering the evolution of a galaxy (i.e. the motion of its stars)
– “molecular dynamics”
– “N-body simulations”

Properties:

– Total energy of the system $H(x,p)$ stays constant
– Dynamics are reversible

Important for detailed balance
Background: Hamiltonian Dynamics

Let $\mathbf{x} \in \mathbb{R}^N$ be a position

$\mathbf{p} \in \mathbb{R}^N$ be a momentum

Potential energy: $E(\mathbf{x})$

Kinetic energy: $K(\mathbf{p}) = \mathbf{p}^T \mathbf{p} / 2$

Total energy: $H(\mathbf{x}, \mathbf{p}) = E(\mathbf{x}) + K(\mathbf{p})$

Given a starting position $x^{(1)}$ and a starting momentum $p^{(1)}$ we can simulate the Hamiltonian dynamics of the system via:

1. Euler’s method
2. Leapfrog method
3. etc.
Background: Hamiltonian Dynamics

Parameters to tune:
1. Step size, $\epsilon$
2. Number of iterations, $L$

Leapfrog Algorithm:

\[
\text{for } \tau \text{ in } 1 \ldots L:
\]
\[
p = p - \frac{\epsilon}{2} \nabla_x E(x)
\]
\[
x = x + \epsilon p
\]
\[
p = p - \frac{\epsilon}{2} \nabla_x E(x)
\]
Background: Hamiltonian Dynamics

Figure 5.1 shows the result of using Euler's method to approximate the dynamics defined by the Hamiltonian of Equation 5.8, starting from $q(0) = 0$ and $p(0) = 1$, and using a stepsize of $\varepsilon = 0.3$ for 20 steps (i.e. to $\tau = 0.3 \times 20 = 6$). The results are not good—Euler's method produces a trajectory that diverges to infinity, but the true trajectory is a circle. Using a smaller value of $\varepsilon$, and correspondingly more steps, produces a more accurate result at $\tau = 6$, but although the divergence to infinity is slower, it is not eliminated.

(a) Momentum ($p$) Euler's method, stepsize 0.3
(b) Modified Euler's method, stepsize 0.3
(c) Leapfrog method, stepsize 0.3
(d) Leapfrog method, stepsize 1.2

FIGURE 5.1 Results using three methods for approximating Hamiltonian dynamics, when $H(q, p) = \frac{q^2}{2} + \frac{p^2}{2}$. The initial state was $q = 0$, $p = 1$. The stepsize was $\varepsilon = 0.3$ for (a), (b), and (c), and $\varepsilon = 1.2$ for (d). Twenty steps of the simulated trajectory are shown for each method, along with the true trajectory (in gray). Figure from Neal (2011)
Hamiltonian Monte Carlo

\textbf{Preliminaries}

Goal:
\[ p(x) = \frac{\exp\{-E(x)\}}{Z} \quad \text{where} \quad x \in \mathbb{R}^N \]

Define:
\[ K(p) = \frac{p^T p}{2} \]
\[ H(x, p) = E(x) + K(p) \]
\[ p(x, p) = \frac{\exp\{-H(x, p)\}}{Z_H} \]
\[ = \frac{\exp\{-E(x)\} \exp\{-K(p)\}}{Z_H} \]

\textbf{Note:}
Since \( p(x, p) \) is separable...

\[ \Rightarrow \sum_p p(x, p) = \frac{\exp\{-E(x)\}}{Z} \quad \text{Target dist.} \]
\[ \Rightarrow \sum_x p(x, p) = \frac{\exp\{-K(x)\}}{Z_K} \quad \text{Gaussian} \]
Whiteboard

• Hamiltonian Monte Carlo algorithm (aka. Hybrid Monte Carlo)
Hamiltonian Monte Carlo

Figure 5.3

A trajectory for a two-dimensional Gaussian distribution, simulated using 25 leapfrog steps with a stepsize of $\epsilon = 0.25$. The ellipses plotted are one standard deviation from the means. The initial state had $q = [-1.50, -1.55]^T$ and $p = [-1, 1]^T$.

Figure 5.3 shows a trajectory based on this Hamiltonian, such as might be used to propose a new state in the HMC method, computed using $L = 25$ leapfrog steps, with a stepsize of $\epsilon = 0.25$. Since the full state space is four-dimensional, Figure 5.3 shows the two position coordinates and the two momentum coordinates in separate plots, while the third plot shows the value of the Hamiltonian after each leapfrog step.

Notice that this trajectory does not resemble a random walk. Instead, starting from the lower left-hand corner, the position variables systematically move upward and to the right, until they reach the upper right-hand corner, at which point the direction of motion is reversed. The consistency of this motion results from the role of the momentum variables. The projection of $p$ in the diagonal direction will change only slowly, since the gradient in that direction is small, and hence the direction of diagonal motion stays the same for many leapfrog steps. While this large-scale diagonal motion is happening, smaller-scale oscillations occur, moving back and forth across the “valley” created by the high correlation between the variables.

The need to keep these smaller oscillations under control limits the stepsize that can be used. As can be seen in the rightmost plot in Figure 5.3, there are also oscillations in the value of the Hamiltonian (which would be constant if the trajectory were simulated exactly). If a larger stepsize were used, these oscillations would be larger. At a critical stepsize ($\epsilon = 0.45$ in this example), the trajectory becomes unstable, and the value of the Hamiltonian grows without bound. As long as the stepsize is less than this, however, the error in the Hamiltonian stays bounded regardless of the number of leapfrog steps done. This lack of growth in the error is not guaranteed for all Hamiltonians, but it does hold for many distributions more complex than Gaussians. As can be seen, however, the error in the Hamiltonian along the trajectory does tend to be positive more often than negative. In this example, the error is $+0.41$ at the end of the trajectory, so if this trajectory were used for an HMC proposal, the probability of accepting the endpoint as the next state would be $\exp(-0.41) = 0.66$.

5.3.3.2 Sampling from a Two-Dimensional Distribution

Figures 5.4 and 5.5 show the results of using HMC and a simple random-walk Metropolis method to sample from a bivariate Gaussian similar to the one just discussed, but with stronger correlation of 0.98.
M-H vs. HMC

Random-walk Metropolis

Hamiltonian Monte Carlo

FIGURE 5.4
Twenty iterations of the random-walk Metropolis method (with 20 updates per iteration) and of the Hamiltonian Monte Carlo method (with 20 leapfrog steps per trajectory) for a two-dimensional Gaussian distribution with marginal standard deviations of one and correlation 0.98. Only the two position coordinates are plotted, with ellipses drawn one standard deviation away from the mean.

In this example, as in the previous one, HMC used a kinetic energy (defining the momentum distribution) of $K(p) = p^T p / 2$. The results of 20 HMC iterations, using trajectories of $L = 20$ leapfrog steps with stepsize $\epsilon = 0.18$, are shown in the right plot of Figure 5.4. These values were chosen so that the trajectory length, $\epsilon L$, is sufficient to move to a distant point in the distribution, without being so large that the trajectory will often waste computation time by doubling back on itself. The rejection rate for these trajectories was 0.09.

Figure 5.4 also shows every 20th state from 400 iterations of random-walk Metropolis, with a bivariate Gaussian proposal distribution with the current state as mean, zero correlation, and the same standard deviation for the two coordinates. The standard deviation of the proposals for this example was 0.18, which is the same as the stepsize used for HMC proposals, so that the change in state in these random-walk proposals was comparable to that for a single leapfrog step for HMC. The rejection rate for these random-walk proposals was 0.37.

Figure from Neal (2011)
HIGH-LEVEL INTRO TO VARIATIONAL INFERENCE
Variational Inference

Problem:
– For observed variables $x$ and latent variables $z$, estimating the posterior $p(z \mid x)$ 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 $x$ and latent variables $z$, estimating the posterior $p(z | x)$ is intractable.
– For training data $x$ and parameters $z$, estimating the posterior $p(z | x)$ 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 \( x \) and latent variables \( z \), estimating the posterior \( p(z \mid x) \) is intractable
– For training data \( x \) and parameters \( z \), estimating the posterior \( p(z \mid x) \) is intractable

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

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
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 \( z^{(1)}, z^{(2)}, \ldots, z^{(N)} \)
  • 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
EXAMPLES OF VARIATIONAL APPROXIMATIONS
Mean Field for MRFs

• Mean field approximation for Markov random field (such as the Ising model):

\[ q(x) = \prod_{s \in V} q(x_s) \]
Variational Inference for MRFs

• 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 figure), clusters of (disjoint) latent variables are independent.
V.I. for Factorial HMM

• For a factorial HMM, we could decompose into chains
LDA Inference

• Explicit Variational Inference (original distribution)
LDA Inference

- Explicit Variational Inference (variational approximation)
LDA Inference

- Collapsed Variational Inference
MEAN FIELD VARIATIONAL INFERENCE
KL Divergence

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

\[
KL(q \parallel p) = E_{q(x)}[\log q(x)/p(x)]
\]

• **Properties:**
  
  – KL(q \parallel p) measures the **proximity** of two distributions q and p
  
  – KL is **not** symmetric: \( KL(q \parallel p) \neq KL(p \parallel q) \)
  
  – KL is minimized when \( q(x) = p(x) \) for all \( x \in \mathcal{X} \)
Variational Inference

Whiteboard

- Background: KL Divergence
- Mean Field Variational Inference (overview)
- Evidence Lower Bound (ELBO)
- ELBO’s relation to log p(x)