Beyond EM: Bayesian Techniques for HLT

Hal Daumé III
me@hal3.name

http://bayes.hal3.name/

Acknowledgments: David Blei, Yee-Whye Teh, Aaron D'Souza
Horse Racing

Lucky

\( \theta_{\text{Lucky}} \)

Pokey

\( \theta_{\text{Pokey}} \)

Steady

\( \theta_{\text{Steady}} \)
Who Should Be Here?

“My EM converges to garbage!”

“I want to integrate domain knowledge.”

“My independence assumptions don't factor nicely!”

“Bayesian techniques are worthless...
 too hard...
 too slow...”
Tutorial Goals

Understand when to be Bayesian

Know the natural prior distributions

Draw complex graphical models

Implement a Gibbs sampler for LDA

Read NIPS/UAI/etc. papers
Empirical Motivation

Mean Average Precision

- Random
- Position
- IR (1 hour)
- EM (2 hours)
- Bayesian (2.5 hours)

See also: DM06
Model for Q-F Summarization

- Suppose a document D is relevant to two queries, Q1 and Q2
- Mark each sentence with the degree to which it is about:
  - Q1
  - Q2
  - D, but not Q1 nor Q2
  - General English
- Now, mark each word in that sentence with an absolute judgment about where it came from
  - Sentences which are more like Q1 are more likely to have words from Q1
  - General English words are likely to be consistent across the whole corpus
  - Document-specific words are likely to be consistent across the whole document
  - Query-specific words are likely to be consistent across all documents relevant to a given query
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
### A Brief Refresher

#### Distributions

<table>
<thead>
<tr>
<th>Binary</th>
<th>$\text{Bin}(x \mid N, \theta) \propto \theta^n (1 - \theta)^{N-n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>K classes</td>
<td>$\text{Mult}(\bar{x} \mid \bar{\theta}) \propto \prod \theta_k^{x_k}$</td>
</tr>
</tbody>
</table>

#### Expectations:

$$E_{x \sim p}[f(x)] = \begin{cases} 
\sum_{x \in X} p(x) f(x) & X \text{ is discrete} \\
\int_{X} dx \ p(x) f(x) & X \text{ is continuous}
\end{cases}$$

#### Probability Calculus:

$$p(x_1 : N) = \prod_{n} p(x_n \mid x_{1:n-1}) \quad p(a \mid b) = \frac{p(a) p(b \mid a)}{p(b)}$$
The Bayesian Paradigm

- Every statistical problem has *data* and *parameters*
- Find a probability *distribution* of the *parameters* given the data using Bayes' Rule:

\[
P(params \mid data) = \frac{P(params)P(data \mid params)}{P(data)}
\]

- Use the posterior to:
  - Predict unseen data (machine learning)
  - Reach scientific conclusions (statistics)
  - Make optimal decisions (Bayesian decision theory)
Models, Parameters and Data

- Model = Our explanation of the world (data)
  - Examples: maximum entropy models, IBM model 1, trigram LM

- Parameters = All unknown aspects of the model
  - Examples: “lambda” parameters, T-table, p(ate | the man)

- Data = All observed variables

Inference problems:
- Estimate parameters (or their distribution)
- Estimate missing data (prediction)
- Find a good model
What is a *Good* Model?

- We can consider models by looking at the probability that they generate our data set (the marginal likelihood of the data):

  \[
P(\text{data} \mid \text{model})
  \]

  all possible data sets

  Model 1

  Model 2

  Model 3

  Current data set
Tutorial Outline

- Introduction to the Bayesian Paradigm
- **Background Material**
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Graphical Models

- Convenient notation for representing probability distributions and conditional independence assumptions

- A observed random variable

- A unobserved/hidden random variable

- A observed/known parameter

- A unobserved/unknown parameter

- A submodel replicated $N$ times

- An indication of conditional dependence

See also: Murphy
Example 1: Naïve Bayes

- For each example \( n \):
  - Choose a class \( Y \) by:
    \[
    p(Y = y | \pi) \propto \pi_y
    \]
  - For each feature \( f \):
    - Choose \( X \) by:
      \[
      p(X_f | \theta^Y) \propto \theta^Y_f
      \]

\[
X \mid \theta, Y \sim \text{Binomial}(X \mid \theta^Y)
\]
\[
Y \mid \pi \sim \text{Multinomial}(\pi)
\]
Example 1: Naïve Bayes

\[ p(D \mid \theta, \pi) = \prod_n p(y_n \mid \pi) \prod_f p(x_{nf} \mid y_n, \theta) \]

\[ = \prod_n \pi^y_n (1 - \pi)^{1 - y_n} \prod_f \prod_v \theta_{yfv}^{x_{nfv}} \]

\[ \pi \quad \text{if } y_n = 1 \]
\[ 1 - \pi \quad \text{if } y_n = 0 \]
\[ \theta_{yfv} \quad \text{if } x_{nfv} = 1 \]

\[ \theta_{yfv} = \text{probability that feature } f \text{ takes value } v \text{ if the class is } y \]
Example 2: Maximum Entropy

For each example \( n \):  
- Choose a class \( Y \) by:

\[
p(Y = y \mid X, \theta) \propto \exp\left[ \sum_f X_f \theta_f \right]
\]
Example 3: Hidden Markov Models

Task: Write out corresponding probability distribution.

See also: Bilmes
Example for Summarization

Consider a stupid summarization model:

- Each word in a document is drawn independently
- Each word is drawn either from a *general English* model, or a document specific model
- We don't know which words are drawn from which

\[
p(w | \pi, \beta^G, \beta^D) = \prod_m \prod_n \sum_{\zeta_{mn}} p(z_{mn} | \pi)
\]

\[
p(w | \beta^G)^{z_{mn}} p(w | \beta^D_m)^{1-z_{mn}}
\]
Fun with Graphical Models

- Easy to propose extensions to the model: add sentences!
Fun with Graphical Models

- Add queries!

Task: Write out corresponding probability distribution.
Tutorial Outline

- Introduction to the Bayesian Paradigm
- **Background Material**
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
- Survey of Popular Models
- Pointers to Literature
- Conclusions

- Laplace Approximation
- Variational Approximation
- Message Passing...
Maximum Likelihood Estimators (MLE)

- Take a parameterized model and some data
- Find the parameters that maximize the likelihood of that data (i.e., the 'probability' of the parameters given the data):

\[
L(\theta, \pi \mid X_{1:N}, Y_{1:N}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{Y_{nk}} (1 - \pi_k)^{1 - Y_{nk}} \prod_{f=1}^{F} \left( \theta_f^n \right)^{Y_{nf}} (1 - \theta_f^n)^{1 - X_{nf}}
\]

\[
l(\theta, \pi) = \sum_{n} \sum_{k} (Y_{nk} \log \pi_k + (1 - Y_{nk}) \log (1 - \pi_k)) \\
+ \sum_{n} \sum_{f} (X_{nf} \log \theta_f^n + (1 - X_{nf}) \log (1 - \theta_f^n))
\]

\[
\frac{\partial l}{\partial \pi} = \sum_{n} \sum_{k} \left[ \frac{Y_{nk} - 1 - Y_{nk}}{\pi_k} \right] \\
\frac{\partial l}{\partial \theta^k} = \sum_{n : Y_{n} = k} \sum_{f} \left[ \frac{X_{nf}}{\theta_f^k} - \frac{1 - X_{nf}}{1 - \theta_f^k} \right]
\]
Tutorial Outline

- Introduction to the Bayesian Paradigm
- **Background Material**
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
MLE with hidden variables

- Consider a stupid summarization model:
  - Each word in a document is drawn independently
  - Each word is drawn either from a general English model, or a document specific model
  - We don't know which words are drawn from which

\[
p(w \mid \pi, \beta^G, \beta^D) = \prod_m \prod_n \sum_{z_{mn}} p(z_{mn} \mid \pi) p(w \mid \beta^G_{z_{mn}}) p(w \mid \beta^D_{1-z_{mn}})
\]

- Compute log likelihood:

\[
l(\pi, \beta \mid w) = \sum_m \sum_n \log \sum_{z_{mn}} ...
\]

- Uh oh! Logs can't go inside sums!
Expectation Maximization

- We would like to move the log inside the sum, but can we?
- Jensen's Inequality to the rescue:

\[
\log p(x \mid \theta) = \log \int_z dz \ p(x, z \mid \theta) \\
= \log \int_z dz \ q(z) \frac{p(X, z \mid \theta)}{q(z)} \\
\geq \int_z dz q(z) \log \frac{p(X, z \mid \theta)}{q(z)} \\
= \int_z q(z) \log p(x, z \mid \theta) - \int_z q(z) \log q(z) \\
= \mathbb{E}_{z \sim q} \{\log p(x, z \mid \theta)\} - \mathbb{E}_{z \sim q} \{\log q(z)\}
\]

- For any distribution \( Q \) (with the same support)
- How should we choose \( Q \)?
Expectation Maximization

- If we set \( q(z) = p(z \mid x, \theta) \) then the lower bound becomes an equality:

\[
\int_z dz \ q(z) \log \frac{p(x, z \mid \theta)}{q(z)} = \int_z dz \ p(x \mid z, \theta) \log \frac{p(x, z \mid \theta)}{p(x \mid z, \theta)} \\
= \int_z dz \ p(x \mid z, \theta) \log \frac{p(z \mid x, \theta) p(x \mid \theta)}{p(x \mid z, \theta)} \\
= \int_z dz \ p(x \mid z, \theta) \log p(x \mid \theta) \\
= \log p(x \mid \theta) \int_z dz \ p(x \mid z, \theta) \\
= \log p(x \mid \theta)
\]

- So, when computing \( E_{z \sim q} \{ \log p(x, z \mid \theta) \} \), the expectation should be taken with respect to the true posterior
**EM in Practice**

- Recall, we wanted to estimate parameters for:
  \[
  p(w \mid \pi, \beta^G, \beta^D) = \prod_{m} \prod_{n} \sum_{z_{mn}} p(z_{mn} \mid \pi) p(w \mid \beta^G)^{z_{mn}} p(w \mid \beta_m^D)^{1-z_{mn}}
  \]

  \[
  = \prod_{m} \prod_{n} E_{z_{mn} \sim \pi} \{ p(w \mid \beta^G)^{z_{mn}} p(w \mid \beta_m^D)^{1-z_{mn}} \}
  \]

- So we replace the hidden variables with their expectations:
  \[
  I(\beta \mid w) \geq \sum_{m} \sum_{n} E[z_{mn}] \log p(w \mid \beta^G) + (1 - E[z_{mn}]) \log p(w \mid \beta_m^D)
  \]

- All we need to do is calculate the expectations:
  \[
  E[z_{mn}] \propto p(z_{mn}=1 \mid \pi) p(w \mid \beta^G)
  \]

- And now the computation proceeds as in the no-hidden-variable setting
EM Summed Up

- Initialize parameters however you desire
- Repeat:
  - \textit{E-STEP}:
    Compute expectations of hidden variables under the current parameter settings
  - \textit{M-STEP}:
    Optimize parameters given those expectation

- This procedure is guaranteed to:
  - Converge to a (local) maximum
  - Monotonically increase the incomplete log-likelihood
EM Graphically
EM on our simple model

- Suppose we have three words: \{A, B, C\}
- Document 1 = [A B], Document 2 = [A C]
- Initialized uniformly

E-step: \[ E_{\{z_{mn}\}} \propto p(z_{mn} = 1 \mid \pi)p(w \mid \beta^G) \]

\[
E\{z_{11}\} = \frac{\pi \beta_A^G}{\pi \beta_A^G + (1-\pi)\beta_{1A}^D} = \frac{0.5 \times 1/3}{0.5 \times 1/3 + 0.5 \times 1/3} = 0.5
\]

\[
E\{z_{12}\} = E\{z_{21}\} = E\{z_{22}\} = 0.5
\]

M-step:

\[
\beta_A^G = \frac{1}{Z}[E\{z_{11}\} + E\{z_{21}\}] = \frac{1}{2}
\]

\[
\beta_B^G = \frac{1}{Z}[E\{z_{12}\}] = \frac{1}{4}
\]

\[
\beta_C^G = \frac{1}{Z}[E\{z_{22}\}] = \frac{1}{4}
\]

\[
\beta_{1A}^D = \frac{1}{Z}[1 - E\{z_{11}\}] = \frac{1}{2}
\]

\[
\beta_{1B}^D = \frac{1}{Z}[1 - E\{z_{12}\}] = \frac{1}{2}
\]

\[
\beta_{1C}^G = 0
\]

\[
\beta_{2A}^D = \frac{1}{Z}[1 - E\{z_{21}\}] = \frac{1}{2}
\]

\[
\beta_{2B}^D = 0
\]

\[
\pi = \frac{E\{z_{11}\} + E\{z_{21}\}}{E\{z_{11}\} + E\{z_{21}\} + E\{z_{12}\} + E\{z_{22}\}} = \frac{1}{2}
\]
EM on our simple model

- Suppose we have three words: \{A, B, C\}
- Document 1 = \[A \ B\], Document 2 = \[A \ C\]
- Initialized uniformly

Task: Implement EM for this model + data

Incomplete log likelihood

Complete log likelihood
Problems with Maximum Likelihood

**Powerful model ⇒ Worthless results**
(due to overfitting...)

**Theoretically unjustified**
(some would argue...)

**Computationally Expensive**
(all that cross-validation...)

**Background knowledge is 0/1**
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
- Survey of Popular Models
- Pointers to Literature
- Conclusions

Laplace Approximation
Variational Approximation
Message Passing...
What is a Prior?

- Recall Bayes' Rule:

\[
P(\theta \mid D) = \frac{P(\theta) P(D \mid \theta)}{\int_\theta d\theta P(\theta) P(D \mid \theta)}
\]

- A prior is a specification of our beliefs about the values parameters can take, before seeing any data.
How Does the Posterior Behave?

Take sequence of data \( x_1, \ldots, x_N \) ...

\[
p(\theta) = \text{just the prior } p(\theta) p(x_1 | \theta)
\]

\[
p(\theta | x_1) = \frac{p(\theta) p(x_1 | \theta)}{\int d\theta \, p(\theta) p(x_1 | \theta)}
\]

\[
p(\theta | x_1, x_2) = \frac{p(\theta | x_1) p(x_2 | \theta)}{\int d\theta \, p(\theta | x_1) p(x_2 | \theta)}
\]

\[
\vdots
\]

\[
p(\theta | x_1 : N) = \frac{p(\theta | x_1 : N-1) p(x_N | \theta)}{\int d\theta \, p(\theta | x_1 : N-1) p(x_N | \theta)}
\]

\[
\text{or } p(\theta) \prod_{n}^{N} p(x_n | \theta) = \frac{p(\theta) \prod_{n}^{N} p(x_n | \theta)}{\int d\theta \, p(\theta) \prod_{n}^{N} p(x_n | \theta)}
\]
Binomial Example
Specifying Priors

- A prior is a map $\pi$ that:
  - Assigns to every setting of parameters a real value
  - Integrates to 1 over the parameter space

- Such a beast can be difficult to describe! Tools:
  - When the parameters are discrete, we can set them by hand
  - Otherwise, we will often choose a parametric prior $\pi(\theta) = \pi(\theta | \alpha)$ and deal with the hyper-parameters
  - Or choose a set of priors and integrate over them (robust Bayes)
  - ...

See also: Ber[3.1-6, 4.7], MK[22,48], Was[11.1]
Empirical Bayes

- Specify a class of priors (typically a functional form):

\[ \Gamma = \{ \pi : \pi(\theta) = g(\theta | \alpha) \} \]

- Estimate the prior by maximizing the marginal likelihood:

\[
\hat{\pi} = \max_{\pi \in \Gamma} p(x | \pi) \\
= \max_{\alpha \in A} \int d\theta \pi(\theta | \alpha) p(x | \theta)
\]
Conjugate (convenient) Priors

Recall: \[ p(\theta \mid x_{1:N}) = \frac{p(\theta) \prod_{n} p(x_n \mid \theta)}{\int d\theta \ p(\theta) \prod_{n} p(x_n \mid \theta)} \]

Given a distribution \( p(x \mid \theta) \)

And a prior \( \pi(\theta \mid \alpha) \)

The prior is *conjugate* if:

\[ p(\theta \mid \alpha, x) = \frac{\pi(\theta \mid \alpha) p(x \mid \theta)}{\int_{\theta} F^{\pi(\alpha)}(\theta) p(x \mid \theta)} = \pi(\theta \mid \hat{\alpha}) \]
## Summary of Distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Domain</th>
<th>Picture</th>
<th>Parametric Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>Binary</td>
<td><img src="image" alt="Binomial" /></td>
<td>$Bin(x \mid N, \theta) \propto \theta^n (1-\theta)^{N-n}$</td>
</tr>
<tr>
<td>Multinomial</td>
<td>K classes</td>
<td><img src="image" alt="Multinomial" /></td>
<td>$Mult(\bar{x} \mid \bar{\theta}) \propto \prod \theta^x_k$</td>
</tr>
<tr>
<td>Beta</td>
<td>[0,1]</td>
<td><img src="image" alt="Beta" /></td>
<td>$Beta(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$</td>
</tr>
<tr>
<td>Gamma</td>
<td>[0,\infty)</td>
<td><img src="image" alt="Gamma" /></td>
<td>$Gam(x \mid a, b) \propto x^{-a-1} \exp(-bx)$</td>
</tr>
<tr>
<td>Dirichlet</td>
<td>Simplex</td>
<td><img src="image" alt="Dirichlet" /></td>
<td>$Dir(\bar{\theta} \mid \bar{\alpha}) \propto \prod \theta^{\alpha_k-1}$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>Reals</td>
<td><img src="image" alt="Gaussian" /></td>
<td>$Nor(x \mid \mu, \sigma^2) \propto \exp\left((-\frac{(x-\mu)^2}{2\sigma^2})\right)$</td>
</tr>
</tbody>
</table>
Binomial and Beta Distributions

- Binomial distribution models flips of coins (domain={0,1}):
  - Probability that a coin, bias \( \theta \), flipped \( N \) times will come up \( x \) heads
  - Parameters: \( N \in \mathbb{N}^+, \theta \in [0,1] \)
  - Distribution: \( \text{Bin}(x \mid N, \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x} \)
  - Moments: \( \mu = N\theta, \var = N\theta(1-\theta+N\theta) \)

- Beta distribution models nothing (we care about) (domain=[0,1]):
  - Parameters: \( \alpha \in \mathbb{R}^+, \beta \in \mathbb{R}^+ \)
  - Distribution: \( \text{Beta}(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \)
  - Moments: \( \mu = \frac{\alpha}{\alpha+\beta}, \var = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \)

- Beta is conjugate to binomial:
  - Posterior parameters: \( \hat{\alpha} = \alpha + x, \hat{\beta} = \beta + N-x \)
  - Marginal distribution:
    \[
p(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{N}{x} \frac{\Gamma(\alpha+x)\Gamma(\beta+N-x)}{\Gamma(\alpha+\beta+N)}
    \]
Beta Distribution Examples

\[
\text{Beta}(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}
\]

\[\alpha = 0.5\]

\[\alpha = 1\]

\[\alpha = 4\]

\[\beta = 0.5\]

\[\beta = 1\]

\[\beta = 4\]
Multinomial Distribution

- A distribution over counts of $K>1$ discrete events (words)
- Domain: $\langle x_1, \ldots, x_K \rangle \in \mathbb{N}^K$

- Parameters: $\langle \theta_1, \ldots, \theta_K \rangle \in \Delta_K = \{ \theta_{1:K} : \theta_k \geq 0, \sum_k \theta_k = 1 \}$

- Distribution: $\text{Mult}(\bar{x} | \bar{\theta}) = \frac{\Gamma\left(\sum x_k + 1\right)}{\prod \Gamma(x_k + 1)} \prod \theta_k^{x_k}$

- Moments: $\langle \theta_1, \ldots, \theta_K \rangle \in \Delta_K = \{ \theta_{1:K} : \theta_k \geq 0, \sum_k \theta_k = 1 \}$
Dirichlet Distribution

- A distribution over a probability simplex
- Domain: \( \langle \theta_1, \ldots, \theta_K \rangle \in \delta^K \)
- Parameters: \( \langle \alpha_1, \ldots, \alpha_K \rangle \in (\mathbb{R}^+)^K \), \( \hat{\alpha} = \sum_k \alpha_k \)
- Distribution: \( \text{Dir}(\tilde{\theta} | \tilde{\alpha}) = \frac{\Gamma(\hat{\alpha})}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k-1} \)
- Moments: \( \mu_k = \frac{\alpha_k}{\hat{\alpha}} \), \( \text{var}_k = \frac{\alpha_k(\hat{\alpha}-\alpha_k)}{(\hat{\alpha})^2(1+\hat{\alpha})} \)
Multinomial/Dirichlet Pair

- Multinomial distribution: \( \text{Mult}(\bar{x} | \vec{\theta}) = \frac{\Gamma\left(\sum x_k + 1\right)}{\prod \Gamma(x_k + 1)} \prod \theta_k^{x_k} \)

- Dirichlet distribution: \( \text{Dir}(\vec{\theta} | \vec{\alpha}) = \frac{\Gamma(\vec{\alpha})}{\prod \Gamma(\alpha_k)} \prod \theta_k^{\alpha_k-1} \)

- Posterior hyper-parameters:
  \[ \langle \hat{\alpha}_1, \ldots, \hat{\alpha}_K \rangle = \langle \alpha_1 + x_1, \ldots, \alpha_K + x_K \rangle \]

- Marginal Distribution:
  \[ p(\bar{x} | \hat{\alpha}) = \frac{\Gamma\left(\sum x_k + 1\right)}{\prod \Gamma(x_k + 1)} \frac{\Gamma(\hat{\alpha})}{\prod \Gamma(\alpha_k)} \prod \frac{\Gamma(\alpha_k + x_k)}{\Gamma(\hat{\alpha} + \sum x_k)} \]
Gaussian/Gaussian-Gamma

- Gaussian distribution: \( \text{Nor}(x \mid \mu, \sigma^2) = (2 \pi \sigma^2)^{1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \)
- Gaussian prior: \( \text{Nor}(\mu \mid m, s^2) \)
- Gamma prior: \( \text{Gam}(\sigma \mid a, b) = \frac{1}{b^a \Gamma(a)} \sigma^{-2a-1} \exp\left(-\frac{1}{b\sigma^2}\right) \)

\[ a > 0, \ b > 0, \ \text{domain} = \mathbb{R}^+ \]

- Posterior hyper-parameters:
  \[ \hat{s} = \left( \frac{1}{s^2} + \frac{1}{\sigma^2} \right)^{-1/2} \]
  \[ \hat{m} = \frac{m/s^2 + \sum_i x_i/s^2}{1/s^2 + N/s^2} \]
  \[ \hat{a} = a + 1/2 \]
  \[ \hat{b} = \left( b^{-1} + \frac{1}{2} \sum_i (x_i - \bar{x})^2 \right)^{-1} \]

- Marginal distribution:
  \[ p(x \mid m, s^2, a, b) = \text{StuT}(m, a, b) \]
Gamma Distribution

\[ \text{Gam}(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp(-bx) \]

\[ \begin{align*}
\mu &= \frac{a}{b} \\
\text{var} &= \frac{a}{b^2}
\end{align*} \]
Conjugate Priors in Action

\[
\beta(1,1) \quad \beta(2,1) \quad \beta(2,2) \quad \beta(2,3)
\]

\[
\beta(3,3) \quad \beta(4,3) \quad \beta(5,3) \quad \beta(6,3)
\]

\[
\beta(6,4) \quad \beta(7,4) \quad \beta(8,4) \quad \beta(9,4)
\]
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Recall our summarization model

\[ z \mid \pi \sim Bin(\pi) \]
\[ w \mid z, \beta \sim Mult(\beta^G)^z Mult(\beta^D)^{1-z} \]

- The problem was that we don't believe that it's okay for \( \pi \) to go to 0 or 1

- Solution? Put a prior on \( \pi \)!

- What's a good prior?
Bayesianified summarization model

\[ p(D | \beta, a, b) = \int U \, d\pi \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1}(1-\pi)^{b-1} \]

\[ \prod_{m} \prod_{n} \sum_{z_{mn} \in \{0,1\}} \pi^{z_{mn}} (1-\pi)^{1-z_{mn}} \]

\[ \prod_{v} (\beta^{G})^{z_{mn}} w_{mnv} (\beta^{D})^{1-z_{mn}} w_{mnv} \]

Conjugacy does not help because of the hidden variables
Interesting Inference Questions

- Predict values of unobserved data:
  \[ P(U | D) \propto \int d\pi(\theta) P(D | \theta) P(U | \theta) \]

- Compute data likelihood:
  \[ P(D) \propto \int d\pi(\theta) P(D | \theta) \]

- Maximize marginal likelihood:
  \[ P(\alpha | D) \propto \int d\pi(\theta | \alpha) P(D | \theta) \]

- Estimate posterior:
  \[ P(\theta | D) = \frac{\pi(\theta) P(D | \theta)}{P(D)} \]

- GENERAL FORM:
  \[ F = \int_X dx p(x) f(x) = E_{x \sim p} \{ f(x) \} \]
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Integration by Summation

- Remember your 9th grade math:

\[ F = \int_{x} dx \, p(x) f(x) \approx \frac{1}{R} \sum_{x \in R} p(x) f(x) \]
Summing in our Model

Simply rewrite the integral as a sum:

\[
p(D | \beta, a, b) = \int_U d\pi \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1}(1-\pi)^{b-1} \prod_m \prod_n \sum_{z_{mn} \in \{0,1\}} \pi^{z_{mn}}(1-\pi)^{1-z_{mn}} \prod_v (\beta^G_v)^{z_{mn} w_{mnv}} (\beta^D_{dv})^{(1-z_{mn}) w_{mnv}}
\]

\[
\approx \sum_{p=1}^{100} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} (p/100)^{a-1}(1-p/100)^{b-1} \prod_m \prod_n \sum_{z_{mn} \in \{0,1\}} (p/100)^{z_{mn}}(1-p/100)^{1-z_{mn}} \prod_v (\beta^G_v)^{z_{mn} w_{mnv}} (\beta^D_{dv})^{(1-z_{mn}) w_{mnv}}
\]
Integration by Summation

- **Pros:**
  - Easy to implement
  - Arbitrarily accurate

- **Cons:**
  - Only works for doubly-bounded regions
  - Intractable for >1 or >2 dimensions
  - Difficult to choose granularity

- Idea: let's choose R differently

\[ F = \int p(x)f(x) \approx \frac{1}{R} \sum_{x \in R} p(x)f(x) \]
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Monte Carlo Integration

- Uniform sampling:
  - Let $R$ be a (multi)set of points drawn uniformly at random

$$F = \int_x dx \, p(x) \cdot f(x) \approx \frac{1}{R} \sum_{x \in R} p(x) \cdot f(x)$$
Uniform Sampling

- **Pros:**
  - Can now work in arbitrarily high dimensions (in theory)
  - Choice is now size of $R$, not the width of windows

- **Cons:**
  - Number of samples required to get near the mode of a spiky distribution is huge: $R \sim 2^{D/2}$
  - True distribution is rarely uniform

$$F = \int_x dx \ p(x) \ f(x) \approx \frac{1}{R} \sum_{x \in R} p(x) \ f(x)$$

See also: MK[29], Was[24.2], Ando3
Importance Sampling

Let R be a set of points drawn from a proposal distribution q

\[ F = \int x \, dx \, p(x) \, f(x) \approx \frac{\sum r w_r f(x_r)}{\sum r w_r} \quad , \quad w_r = \frac{p(x_r)}{q(x_r)} \]
Importance Sampling

- **Pros:**
  - If $q$ can be constructed similar to $p$, then good samples can be had
  - Can scale better than uniform sampling (not saying much)

- **Cons:**
  - Very sensitive to choice of $q$
  - Hard to evaluate whether it has converged
  - Still a lot of samples required:

IS: $R \sim \exp(\sqrt{2D})$

US: $R \sim 2^{D/2}$
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Monte Carlo methods suffer because the proposal density needs to be similar to the true density everywhere.

MCMC methods get around this problem by changing the proposal density after each sample.

General framework:
- Choose a proposal density \( q(x' \mid x) \) parameterized by location \( x \)
- Initialize state \( x \) arbitrarily
- Repeatedly sample by:
  - Propose a new state \( x' \) from \( q(x' \mid x) \)
  - Either accept or reject this new state
    - If accepted, set \( x = x' \)

*New problem:* samples are no longer independent!
Metropolis-Hastings Sampling

- Accept new states with probability: 
  \[ \text{min} \left\{ 1, \frac{p(x')}{p(x)} \frac{q(x|x')}{q(x'|x)} \right\} \]
- Only put every \(N^{\text{th}}\) sample into \(R\)

\[ F = \int_x dx \, p(x) \, f(x) \approx \frac{1}{R} \sum_{x \in R} f(x) \]
MH in our Model

- Invent a proposal distribution $q$
  
  $$\log a' \mid a \sim \text{Nor}(\log(a), 1)$$
  $$\log b' \mid b \sim \text{Nor}(\log(b), 1)$$
  $$\sigma(\pi') \mid \pi \sim \text{Nor}(\sigma(\pi), 1)$$
  $$z_{mn}' \mid z \sim \text{Bin}(0.5)$$

- Or, condition on all variables:
  
  $$\log a' \sim \text{Nor}(\log(a), 1)$$
  $$\log b' \sim \text{Nor}(\log(b), 1)$$
  $$\sigma(\pi') \sim \text{Beta}(\pi' \mid a, b) \prod_{m,n} \text{Bin}(z_{mn} \mid \pi')$$
  $$z_{mn}' \sim \text{Bin}(z_{mn}' \mid \pi) p(w_{mn} \mid z_{mn}', \beta)$$

- Now we can compute expectations of $z$ easily and use these for the M-step of EM
Metropolis-Hastings Sampling

- **Pros:**
  - No longer need to specify a universally good proposal distribution; only locally good
  - Simple proposal distributions can go far

- **Cons:**
  - Hard to tell now far to space samples:
    - Suppose we use spherical proposals and, then we need at least
      \[ N \geq \left( \frac{\sigma_{\text{max}}}{\sigma_{\text{min}}} \right)^2 \]
      where *sigmas* are lengths of the major density in \( p \)
  - Auto-correlation to track this:
    \[
    R_k = \frac{\sum_{i=1}^{N-k} (X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^{N} (X_i - \bar{X})^2}
    \]
Gibbs Sampling

- Defined only for multidimensional problems
- Useful when you can take out one variable and explicitly sample the rest

\[ F = \int_x dx \frac{p(x)}{f(x)} \approx \frac{1}{R} \sum_{x \in R} f(x) \]
Gibbs Sampling

- Typically our params are: $\tilde{\theta} = \langle \theta_1, \ldots, \theta_D \rangle$
- If, for each $i$, we can draw a sample from:
  
  $$p(\theta_i | \theta_{-i}) = p(\theta_i | \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_D)$$

  
  then we can use Gibbs sampling.

- In graphical models, only depends on the Markov blanket:
  
  $$p(\theta_i | \theta_{-i}) = p(\theta_i | \text{par}(\theta_i)) \prod_{j : \theta_j \in \text{par}(\theta_i)} p(\theta_j | \text{par}(\theta_j))$$

  
  $$p(d | \theta_{-d}) = p(d | a, b) p(e | d) p(f | d, c)$$
Gibbs in our Model

- Compute conditional probabilities

\[
\begin{align*}
a, b & \mid \neg a, b \sim \text{Beta}(\pi \mid a, b) \\
\pi & \mid \neg \pi \sim \text{Beta}(\pi \mid a, b) \prod_{m,n} \text{Bin}(z_{mn} \mid \pi) \\
z_{mn} & \mid \neg z_{mn} \sim \text{Bin}(z_{mn} \mid \pi) p(w_{mn} \mid z_{mn}, \beta)
\end{align*}
\]

- Now we can compute expectations of \(z\) easily and use these for the M-step of EM
  - Alternatively, we could propose values for LMs in the sampling
Gibbs Sampling

- **Pros:**
  - Designed to work in high dimensional spaces
  - Terribly simple to implement
  - Automatable

- **Cons:**
  - Hard to judge convergence, can require many many samples to get an independent one (often worse than MH)
  - Only applicable when conditional distributions are 'nice'
    - (Though there are ways around this)
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Laplace (Saddlepoint) Approximation

- Idea: approximate the expectation by a quadratic (Taylor expansion) and use the normalizing constant from the resulting Gaussian distribution

\[
F = \int_x dx \, p(x) f(x) \approx g(x_0) \sqrt{\frac{2 \pi}{c}} \left[ -\frac{\partial^2}{\partial x^2} \ln g(x) \right]_{x=x_0}
\]

\[
p(x)f(x) = g(x)
\]
Laplace Approximation

- Find a mode $x_0$ of the high-dimensional distribution $g$
- Approximate $\ln g(x)$ by a Taylor expansion around this mode:
  $$\ln g(\bar{x}) \approx \ln g(\bar{x}_0) - \frac{1}{2} (\bar{x} - \bar{x}_0)^T A(\bar{x} - \bar{x}_0)$$

- Compute the matrix $A$ of second derivatives
  $$A_{ij} = -\left[ \frac{\partial^2}{\partial x_i \partial x_j} \ln g(\bar{x}) \right]_{x=x_0}$$
- The exponential form is a Gaussian distribution; use the Gaussian normalizing constant:
  $$F \approx g(x_0) \sqrt{2\pi / c}$$
  $$c = -\left[ \frac{\partial^2 \ln g(x)}{\partial x^2} \right]_{x=x_0}$$

See also: MK[27]
Laplace in our Model

- Compute second derivatives:

\[
\int_U d\pi Z_{ab} \pi^{a-1}(1-\pi)^{b-1} \prod_m \prod_n \sum_{z_{mn}} \pi^{z_{mn}}(1-\pi)^{1-z_{mn}} p(w_{mn} | z_{mn}, \beta)
\]

\[
g(\pi)
\]

\[
\frac{\partial \log g}{\partial \pi} = \frac{(a-1)}{\pi} - \frac{(b-1)}{1-\pi} + \sum_{m,n} \left[ \frac{z_{mn}}{\pi} - \frac{(1-z_{mn})}{1-\pi} \right]
\]

\[
\frac{\partial^2 \log g}{\partial \pi^2} = -\frac{(a-1)}{\pi^2} - \frac{(b-1)}{(1-\pi)^2} - \sum_{m,n} \left[ \frac{z_{mn}}{\pi^2} + \frac{(1-z_{mn})}{(1-\pi)^2} \right]
\]

\[
\frac{\pi_0}{1-\pi_0} = \frac{a-1 + \sum_{m,n} z_{mn}}{b-1 + \sum_{m,n} (1-z_{mn})}
\]

\[
F = \int x \, dx \, p(x) f(x) \approx g(x_0) \sqrt{\frac{2\pi}{c}} , \quad c = -\left[ \frac{\partial^2}{\partial x^2} \ln g(x) \right]_{x=x_0}
\]
Laplace Approximation

- **Pros:**
  - Deterministic
  - Efficient if $A$ is of a suitable form (i.e., diagonal or block-diagonal)
  - Can apply transformations to make quadratic approximation more reasonable

- **Cons:**
  - Poor fit for multimodal distributions
  - Often, $\text{det } A$ cannot be found efficiently

\[
F \approx g(x_0) \sqrt{2\pi/c}
\]

\[
c = -\left[ \frac{\partial^2 \ln g(x)}{\partial x^2} \right]_{x=x_0}
\]
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Variational Approximation

- Basic idea: replace intractable $p$ with tractable $q$
- Old Problem:
  - We cannot come up with a good, single, $q$ to approximate $p$
- Key Idea:
  - Consider a family of distributions $Q = \{ q(\cdot | \phi) : \phi \in \Phi \}$ with 'variational parameters' $\phi$
  - Choose a member $q$ from $Q$ that is closest to $p$

- New problems:
  - How do we choose $Q$?
  - How do we measure 'closeness' between $q$ and $p$?

See also: MK[33], Wain03, Min03
Recall EM and Jensen's Inequality

- Jensen gives us:

\[
\log p(x \mid \theta) = \log \int_z dz \ p(x, z \mid \theta) \\
= \log \int_z dz \ q(z) \frac{p(X, z \mid \theta)}{q(z)} \\
\geq \int_z dz \ q(z) \log \frac{p(X, z \mid \theta)}{q(z)} \\
= \int_z q(z) \log p(x, z \mid \theta) - \int_z q(z) \log q(z) \\
= E_{z \sim q} \{ \log p(x, z \mid \theta) \} - E_{z \sim q} \{ \log q(z) \}
\]

Where we chose \( q(z) = p(z \mid x, \theta) \) to turn the inequality into an equality. But we can also compute:

\[
\log p(x \mid \theta) = \mathcal{L} + KL(q(z) \mid \mid p(z \mid x, \theta))
\]

for any choice of \( q \)
Variational EM

- Parameterize $q$ and directly optimize:

$$
\log p(x | \theta) = \mathbb{E}_{z \sim q}[\log p(x, z | \theta)] - \mathbb{E}_{z \sim q}[\log q(z)] + KL(q(z | \tilde{\theta}) \| p(z | x, \theta))
$$

- Iterate:
  - **V-Step:** Compute variational parameters $\tilde{\theta}$ to minimize KL
  - **E-Step:** Compute expectations of hidden variables wrt $q(\tilde{\theta})$
  - **M-Step:** Maximize $\mathcal{L}$ wrt true parameters $\theta$

- Art: inventing $q$ so that this is all tractable
Variational EM in Pictures

\[ p(x)f(x) = g(x) \]

\[ q(x | \tilde{\theta}) \]
Variational: Choosing Q

Mixture model:

\[ p(w, \pi, z | \rho, a, b) = \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \pi^{a-1} (1-\pi)^{b-1} \]

\[ \prod_{m,n} \pi^{z_{nm}} (1-\pi)^{1-z_{nm}} \prod_{i} \prod_{v} [\beta_v^i]^{z_{mini}} \]

\[ q(\pi, z | \tilde{\alpha}, \tilde{\beta}, \tilde{\pi}) = \frac{\Gamma(\tilde{\alpha}+\tilde{\beta})}{\Gamma(\tilde{\alpha}) \Gamma(\tilde{\beta})} \prod_{m,n} \prod_{i} \pi_{mi}^{\tilde{\alpha}_i-1} \tilde{\pi}_{mn} z_{mini} \]

Key: \( \pi \) and \( z \) are now not tied in the \( q \) distribution!
VEM in our Model

- Iterate:
  - Optimize variational parameters:
    \[ \tilde{\tau}_{mni} \propto \exp \left[ \mathcal{E}_i + \omega_{mni} \right] \]
    \[ \tilde{a}_i = a_i + \sum_{m,n} \tilde{\tau}_{mni} \]
    \[ \mathcal{E}_i = \Psi(\tilde{a}_i) - \Psi\left(\sum_i \tilde{a}_i\right) \quad \omega_{mni} = \sum_j w_{mni} \log \beta^i_j \]

  - Optimize model parameters:
    \[ \beta^i_v \propto \sum_{m,n} \tilde{\tau}_{mni} w_{mnv} \]
    \[ a, b \sim \text{generic optimization techniques} \]
Variational EM Summed Up

- **Steps:**
  - Write down conditional likelihood and choose an approximating distribution (e.g., by factoring everything) with variational parameters
  - Iterate between optimizing the VPs and model parameters

- **Pros:**
  - Efficient, deterministic, often quite accurate

- **Cons:**
  - At its heart, still a mode-based technique
  - Often underestimates the spread of a distribution
  - Approximation is *local*
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Message Passing Algorithms

- Two major choices:
  - What approximating distribution should we use?
  - What cost should we minimize?

\[
D_a(p||q) = \frac{1}{\beta(1-\beta)} \int dx \beta p + (1-\beta) q - p^\beta q^\beta \\
\beta = \frac{1}{2}(1+a)
\]

\[
KL(p||q) = D_1(p||q) \\
KL(q||p) = D_{-1}(p||q)
\]
Empirical Evaluation of Methods

- Query-focused summarization model:

\[ w^Q_{qn} \sim \text{Mult}(\beta^Q_q) \]
\[ \pi_{ms} \sim \text{Dir}(a) \]
\[ z_{msn} \sim \text{Mult}(\pi_{ms}) \]
\[ w_{msn} \sim \text{Mult}(\beta^G_{z_{msn}^1}) \]
\[ \prod_m \text{Mult}(\beta^D_m)^{z_{msn}(m+1)} \]
\[ \prod_q \text{Mult}(\beta^Q_q)^{z_{msn}(q+M+1)} \]
Evaluation Data

- All TREC data
  - Queries 51-350 and 401-450 (35k words)
  - All relevant documents (43k docs, 2.1m sents, 65.8m words)
  - Asked 7 annotators to select up to 4 sentences for an extract
    - Each annotated 25 queries (166 total)
  - Systems produce ranked lists of sentences
    - Compared on mean average precision, mean reciprocal rank and precision at 2

- Computation Time:
  - MAP-EM (2 hours)
  - Summing (2 days)
  - Monte Carlo (2 days)
  - MCMC (1 day)
  - Laplace (5 hours)
  - Variational (4 hours)
  - EP (2.5 hours)
Evaluation Results

Mean Average Precision

- Random
- Position
- IR
- MAP
- Summing
- Monte Carlo
- MCMC
- Laplace
- Variational
- EP

See also: DM06
Tutorial Outline

- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions
Latent Dirichlet Allocation

- Unigram model of documents
- Each document is a mixture over topics
- Each topic is a mixture over words

Generative model for each document (M total):
- Choose a single topic mixture: $\theta \sim \text{Dir}(\alpha)$
- For each word (N total):
  - Choose a topic for this word: $z \sim \text{Mult}(\theta)$
  - Choose the word itself: $w \sim \text{Mult}(\beta^z)$

[Blei, Ng + Jordan, JMLR 03]
LDA: Geometric Interpretation

Doc 1: topography
Doc 2: finance

[Blei, Ng + Jordan, JMLR 03]
LDA: Inference

\[ P(D) = \int_{\Delta_v} dP(\beta) \int_{\mathbb{R}^+} dP(\alpha) \prod_{m=1}^{M} \int_{\Delta_K} d\theta \left[ \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{j=1}^{K} \theta_k^{\alpha-1} \right] \]

\[
\prod_{n=1}^{N} \sum_{z_{mn}=1}^{K} \prod_{i=1}^{V} \prod_{j=1}^{K} \beta_{ji} \mathbf{1}[w_{mn}=i] \mathbf{1}[z_{mn}=j]
\]

Desired: either $\beta$s or $z$s

[Blei, Ng + Jordan, JMLR 03]
LDA: Naïve Gibbs Sampler

[Griffiths + Tenenbaum, CogSci 03]

\[
\begin{align*}
\alpha & \sim P(\alpha) \prod_m \text{Dir}(\theta_m | \alpha) \\
\beta_j & \sim P(\beta_j) \prod_{mn} \text{Mult}(w_{mn} | \beta_j)^1[z_{mn}=j] \\
\theta_m & \sim \text{Dir}(\theta_m | \alpha) \prod_n \text{Mult}(z_{mn} | \theta_m) \\
z_{mn} & \sim \text{Mult}(z_{mn} | \theta_m) \text{Mult}(w_{mn} | \beta_{z_{mn}})
\end{align*}
\]

Can collapse this step!
LDA Results

[Blei, Ng + Jordan, JMLR 03]

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.
Integrating Topics and Syntax

For each document $M$:
- Choose a topic mixture $\theta$
- For each word $N$:
  - Choose topic $z$
  - Choose class $s$
  - Choose $w$ from:
    - $\beta_z$ if $s=0$
    - $\zeta_s$ otherwise

[Griffiths, Steyvers, Blei + Tenenbaum, NIPS 2004]
**LDA versus Topics+Syntax**

<table>
<thead>
<tr>
<th>LDA</th>
<th>Topics</th>
<th>Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>blood of a the the the the the</td>
<td>breath</td>
<td>the a for it new have made would years</td>
</tr>
<tr>
<td>, and of of to in in in in and and and game</td>
<td></td>
<td>say can time</td>
</tr>
<tr>
<td>of in in to water in drink ball</td>
<td></td>
<td>made would years</td>
</tr>
<tr>
<td>body a in in in in and and game</td>
<td></td>
<td>new have made would years</td>
</tr>
<tr>
<td>heart in land and to water in drink ball</td>
<td></td>
<td>have made would years</td>
</tr>
<tr>
<td>and trees to classes picture is story alcohol and</td>
<td></td>
<td>made would years</td>
</tr>
<tr>
<td>in tree farmers government film and is to team</td>
<td></td>
<td>made would years</td>
</tr>
<tr>
<td>to with for a image matter to bottle to</td>
<td></td>
<td>made would years</td>
</tr>
<tr>
<td>is on farm state lens are as in play</td>
<td></td>
<td>made would years</td>
</tr>
</tbody>
</table>

[Griffiths, Steyvers, Blei + Tenenbaum, NIPS 2004]
Matching Words and Pictures

For each image/caption pair $M$
- Draw a topic mixture $\theta \sim \text{Dir}(\alpha)$
  - For each image region $P$
    - Draw a topic $z \sim \text{Mult}(\theta)$
    - Draw the region $r \sim \text{Gaussian}(\mu, \sigma^2)$
  - For each word $N$
    - Draw a image region $y \sim \text{Unif}(1..P)$
    - Draw the word $w \sim \text{Mult}(\beta_{zy})$

[Barnard, Duygulu, de Freitas, Forsyth, Blei + Jordan, JMLR 2003]

- People, tree
- Sky, jet
- Sky, clouds
- Sky, mountain
- Plane, jet
- Plane, jet
Matching Words and Pictures

[Barnard, Duygulu, de Freitas, Forsyth, Blei + Jordan, JMLR 2003]
Conclusions

- Bayesian methods provide efficient, effective models
- Graphical models are an easy language
- Plug and play of Multinomial/Dirichlet/Beta/Gamma leads to models that admit efficient Gibbs sampling methods
- For faster inference, the variational approximation is effective
- Bayesian models of text problems is largely unexplored

- Many topics not discussed:
  - Alternative inference techniques (belief/expectation propagation)
  - Classifiers/discriminative models (Gaussian Processes \( \approx \) SVMs)
  - Infinite models (Dirichlet Processes, Chinese Restaurant Processes)
Bayes in Action (NLP/IR/Text)

Barnard, Duygulu, de Freitas, Forsyth, Blei + Jordan. *Matching words and pictures*. JMLR03.

Daumé III + Marcu, *Bayesian Query-Focused Summarization*, ACL06.


McCallum, Corrada-Emmanuel + Wang, *Topic and Role Discovery in Social Networks*. IJCAI05.

Zhang, Callan + Minka, *Novelty and Redundancy Detection in Adaptive Filtering*. SIGIR02.
For Further Information (Books)


For Further Information (Tutorials)


Murphy, *A Brief Introduction to Graphical Models and Bayesian Networks*. [www.cs.ubc.ca/~murphyk/Bayes/bayes.html](http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html)

Other References


http://bayes.hal3.name/
http://nlpers.blogspot.com