# Beyond EM:
# Bayesian Techniques for HLT

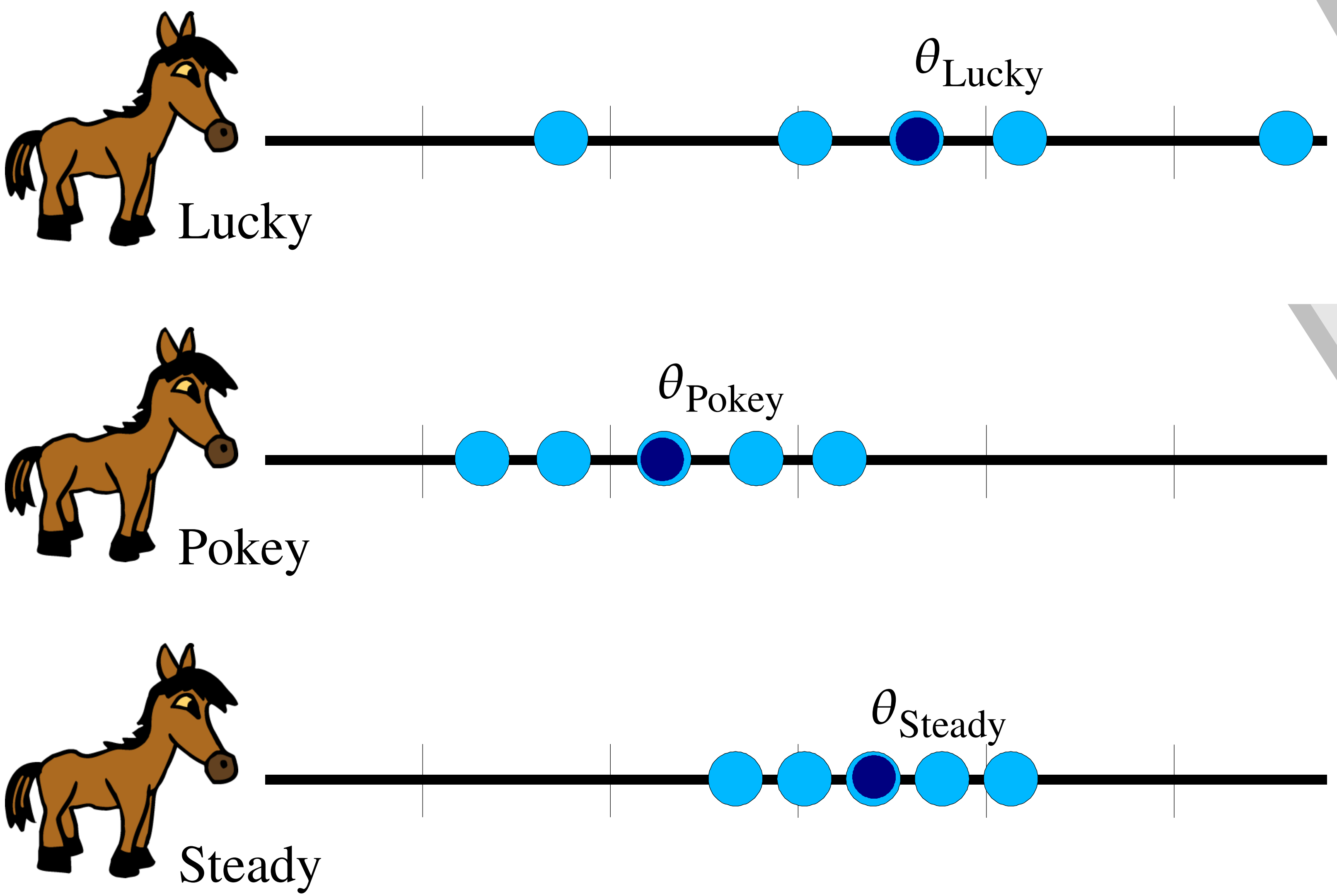
**Hal Daumé III**

`me@hal3.name`

`http://bayes.hal3.name/`

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

# Horse Racing



Hal Daumé III (me@hal3.name)

# 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..."

Hal Daumé III (me@hal3.name)

# Tutorial Goals

**Understand when to be Bayesian**

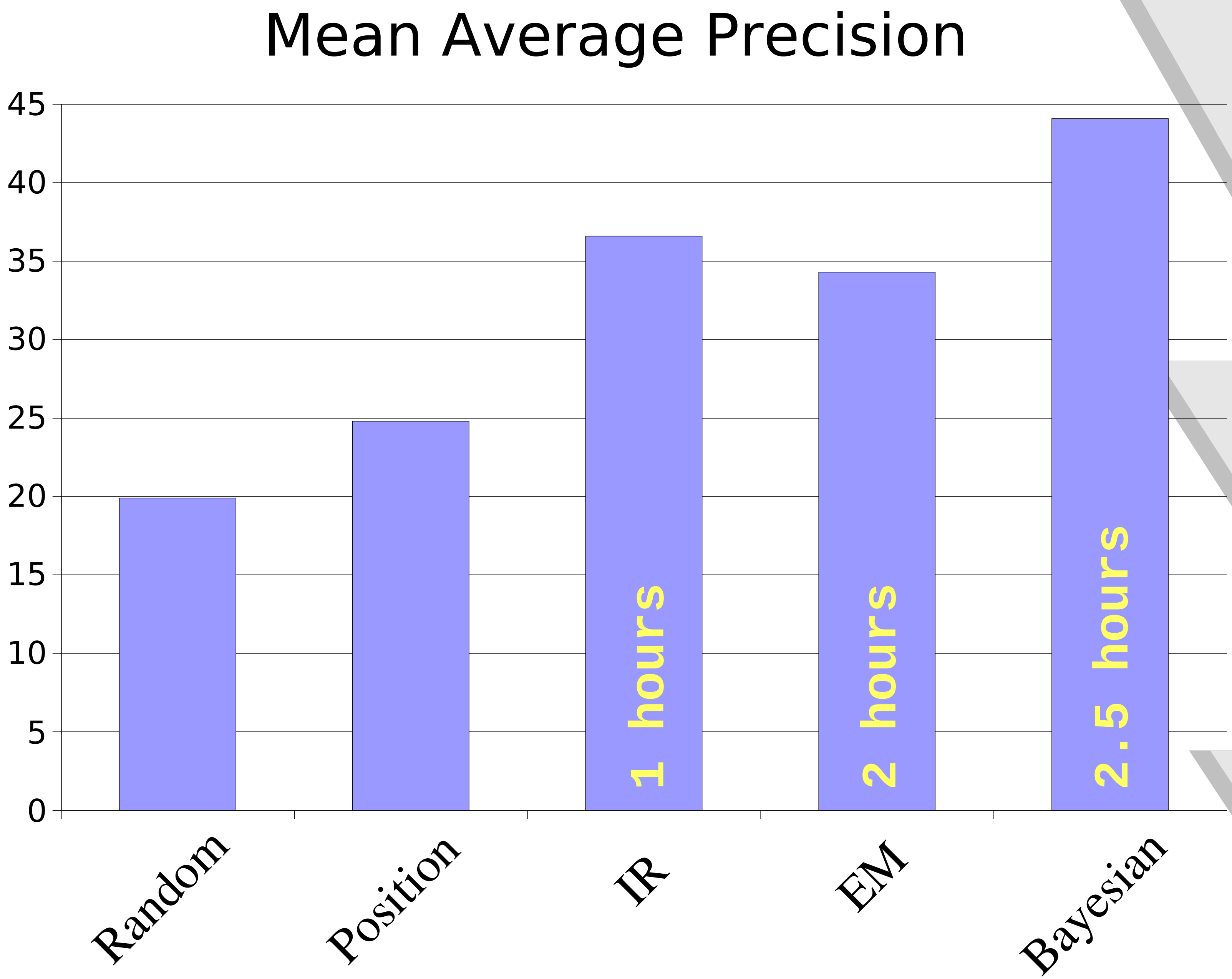
**Know the natural prior distributions**

**Draw complex graphical models**

**Implement a Gibbs sampler for LDA**

**Read NIPS/UAI/etc. papers**

# Empirical Motivation




*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

| | |
|---|---|
| Iraq's National Assembly approved a list of Cabinet members for a transitional government Thursday, three months after national elections. | (0.5, 0.2, 0.2, 0.1) |
| Three ministries – Defense, Oil and Electricity – were filled with temporary appointments because of a last minute failure to reach a compromise. | (0.1, 0.6, 0.1, 0.1) |
| Prime minister Ibrahim al-Jaafari assumed his post with the creation of his government | (0.2, 0.2, 0.3, 0.3) |
| The approval of the Cabinet represents the end of a major political impasse in the country. | (0.4, 0.4, 0.1, 0.1) |
| On Wednesday, al-Jaafari told a news conference that he had submitted his proposal Cabinet to President Jalal Talabani, who had to approve the names before the transitional National Assembly voted on them. | (0.1, 0.2, 0.5, 0.1) |
| Al-Jaafari's announcement came a short time after gunment shot and killed an assembly member on her doorsep in Baghdad… | (0.2, 0.4, 0.1, 0.3) |

*See also: DM06*

Hal Daumé III (me@hal3.name)

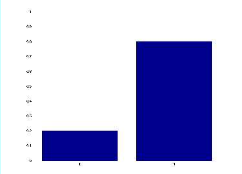
# Tutorial Outline

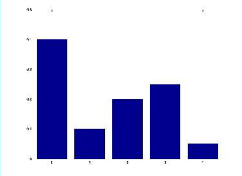
- Introduction to the Bayesian Paradigm
- Background Material
  - Graphical Models
  - Maximum Likelihood
  - Expectation Maximization
- Priors, priors, priors (subjective, conjugate, reference, etc.)
- Inference Problem and Solutions
  - Summing
  - Monte Carlo
  - Markov Chain Monte Carlo
  - Laplace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

Hal Daumé III (me@hal3.name)

# A Brief Refresher

**Distributions**

Binomial Binary  $Bin(x \mid N, \theta) \propto \theta^n (1-\theta)^{N-n}$

Multinomial K classes  $Mult(\bar{x} \mid \bar{\theta}) \propto \prod \theta_k^{x_k}$

**Expectations:**

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

**Probability Calculus:**

$$p(x_{1:N}) = \prod_n p(x_n \mid x_{1:n-1}) \qquad p(a \mid b) = \frac{p(a)\, p(b \mid a)}{p(b)}$$

Hal Daumé III (me@hal3.name)

# The Bayesian Paradigm

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

Prior

Likelihood

Posterior

$$\underbrace{P(params \mid data)}_{\text{Posterior}} = \frac{\overbrace{P(params)}^{\text{Prior}}\,\overbrace{P(data \mid params)}^{\text{Likelihood}}}{\underbrace{P(data)}_{\text{Marginal}}}$$

Marginal

- 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

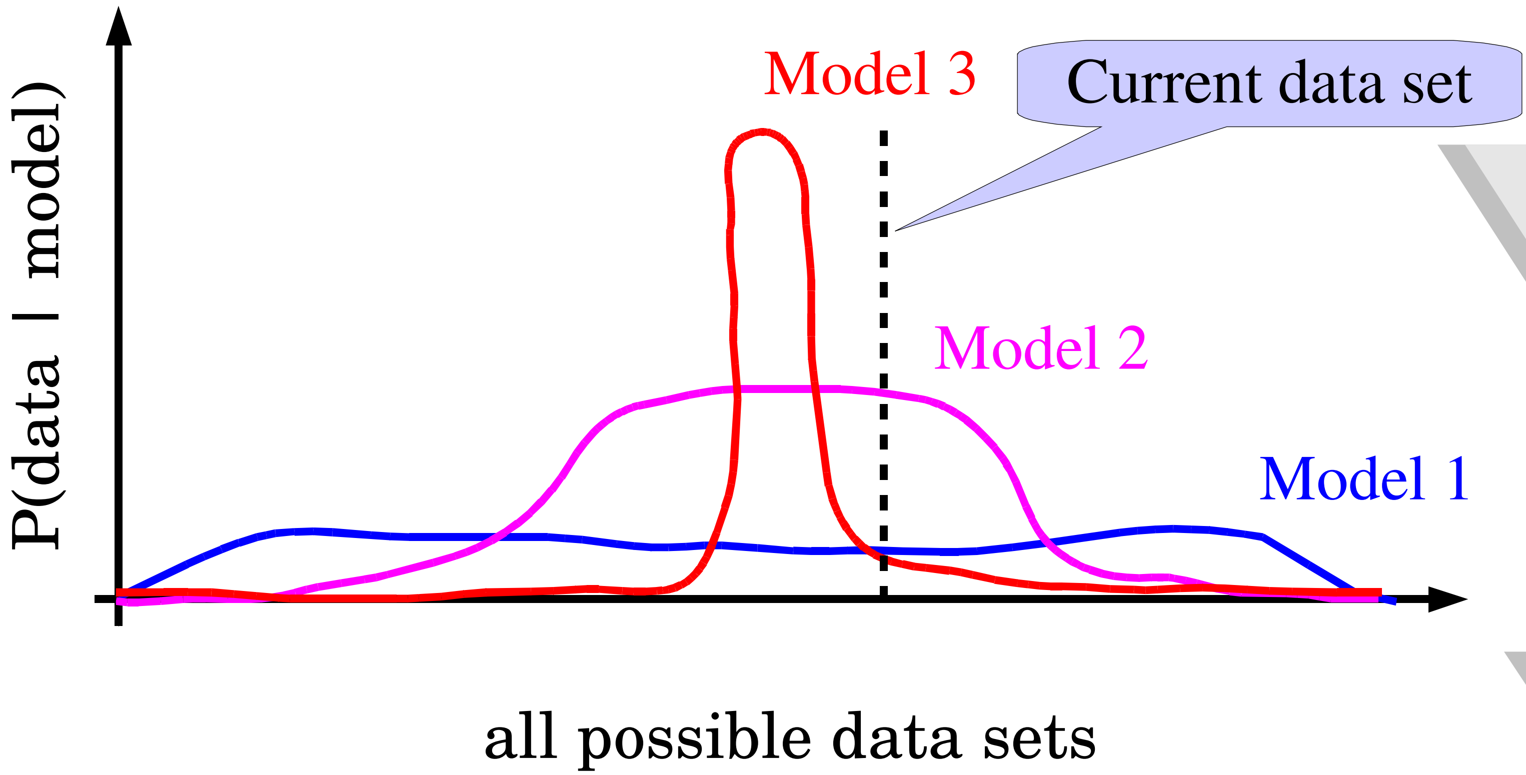
---

- Inference problems:
  - Estimate parameters (or their distribution)
  - Estimate missing data (prediction)
  - Find a good model

Hal Daumé III (me@hal3.name)

# 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):



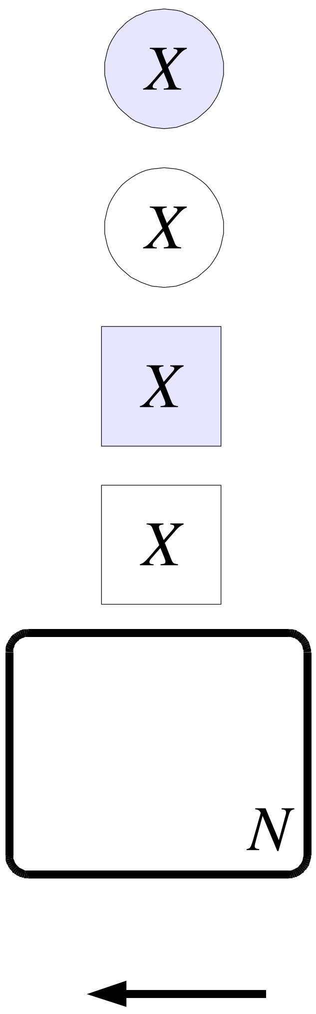

*See also: MK[28.1]*

# 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
  - 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

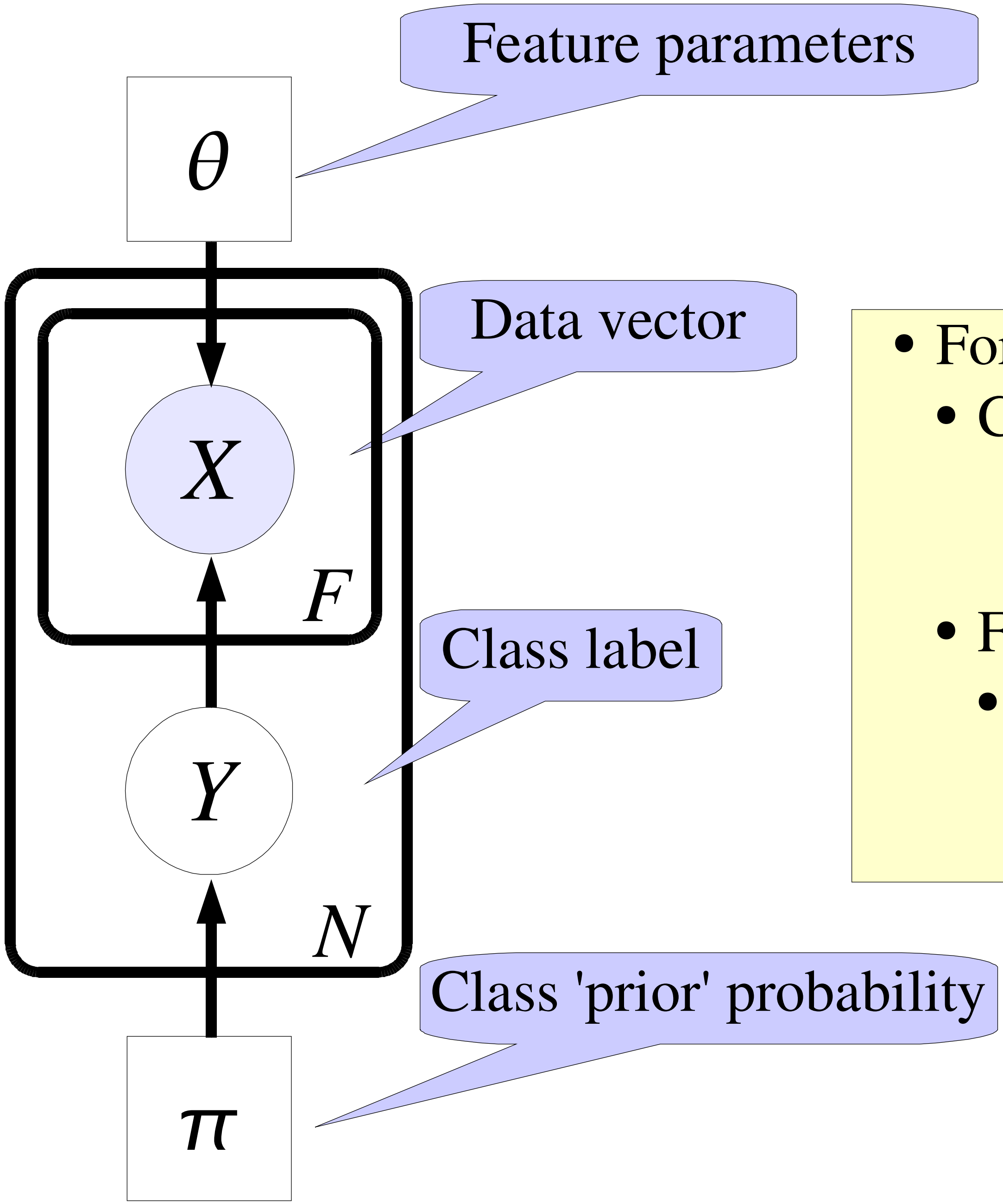
A unobserved/unknown parameter

A submodel replicated *N* times

An indication of conditional dependence

*See also: Murphy*

# Example 1: Naïve Bayes




$$X \mid \theta, Y \sim \text{Binomial}(X \mid \theta^Y)$$
$$Y \mid \pi \sim \text{Multinomial}(\pi)$$

- For each example $n$:
  - Choose a class $Y$ by:

    $$p(Y = y \mid \pi) \propto \pi_y$$

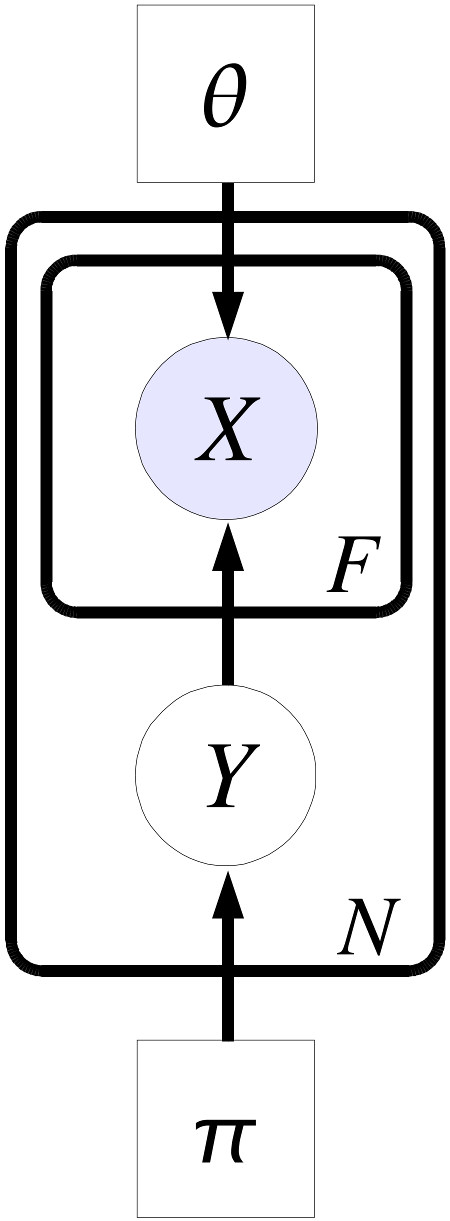
  - For each feature $f$:
    - Choose $X$ by:

      $$p(X_f \mid \theta^Y) \propto \theta_f^Y$$

*See also: Murphy*

Hal Daumé III (me@hal3.name)

# Example 1: Naïve Bayes



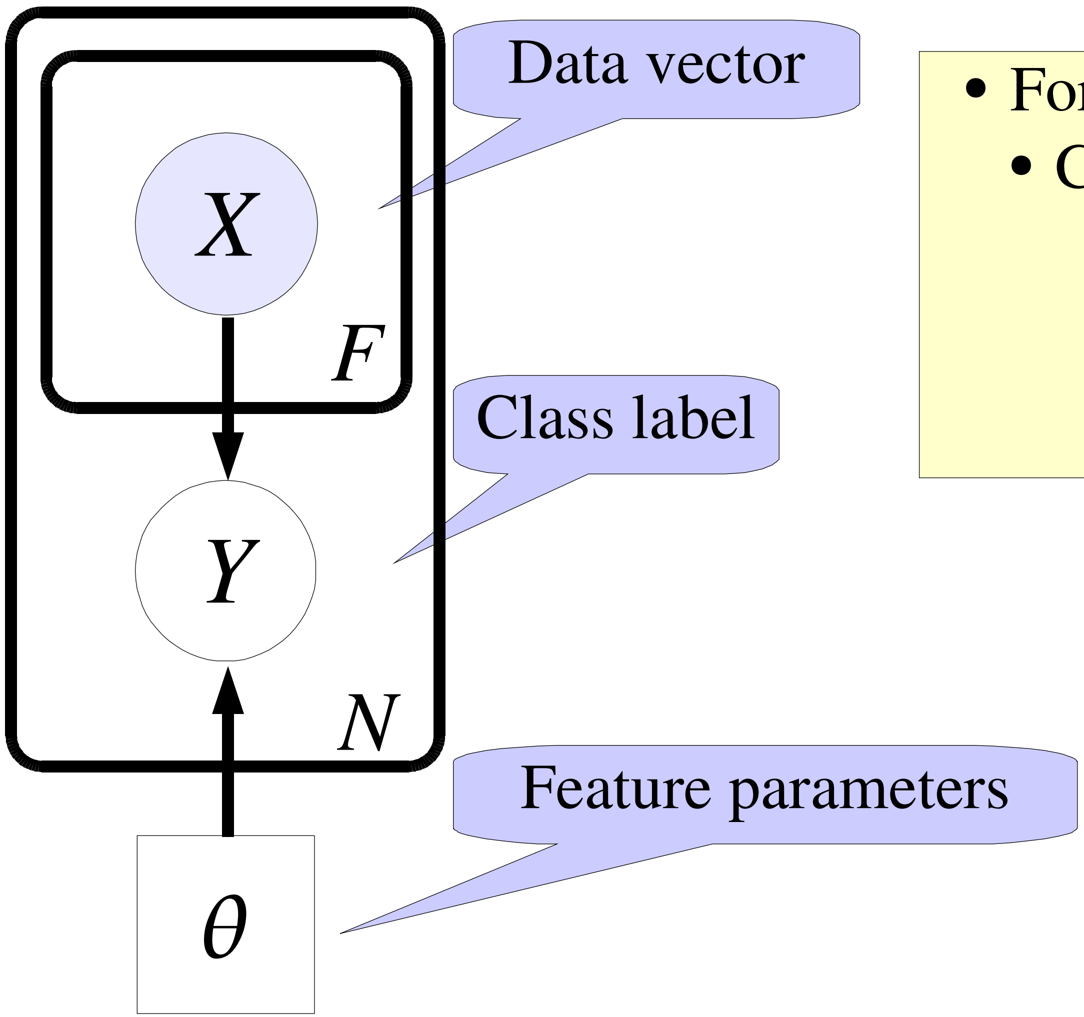

$$p(D \mid \theta, \pi) = \prod_n p(y_n \mid \pi) \prod_f p(x_{nf} \mid y_n, \theta)$$

$$= \prod_n \underbrace{\pi^{y_n} (1-\pi)^{1-y_n}}_{\begin{cases} \pi & \text{if } y_n = 1 \\ 1\text{-}\pi & \text{if } y_n = 0 \end{cases}} \underbrace{\prod_f \prod_v \theta_{y_n f v}^{x_{nfv}}}_{\theta_{yfv} \quad \text{if } x_{nfv} = 1}$$

$\theta_{yfv}$ = probability that feature $f$ takes value $v$ if the class is $y$

*See also: Murphy*

# Example 2: Maximum Entropy

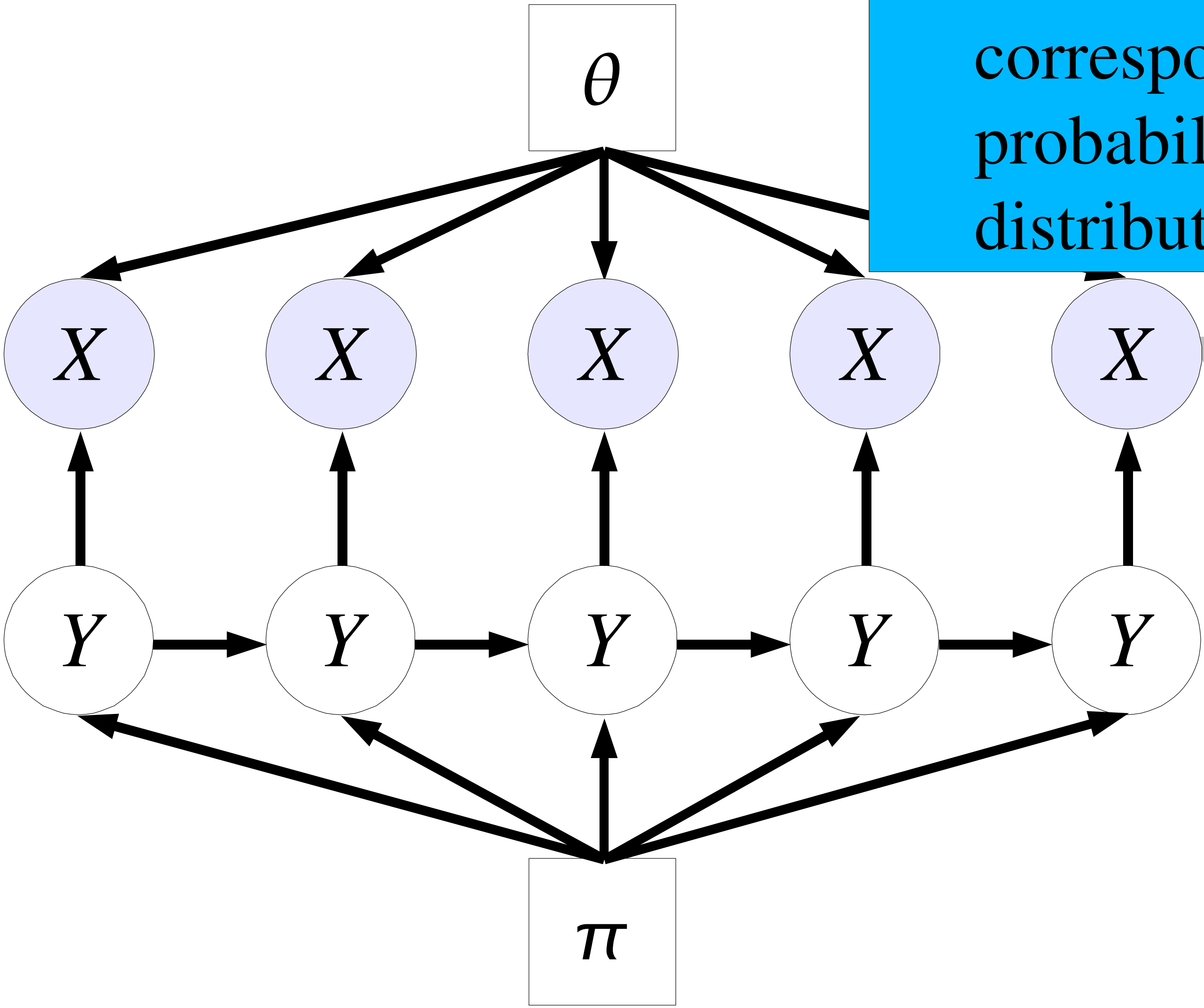



- For each example $n$:
  - Choose a class $Y$ by:

$$p(Y=y \mid X,\theta) \propto \exp\left[\sum_f X_f \theta_f\right]$$

Hal Daumé III (me@hal3.name)

# Example 3: Hidden Markov Models




Task: Write out corresponding probability distribution.

*See also: Bilmes*

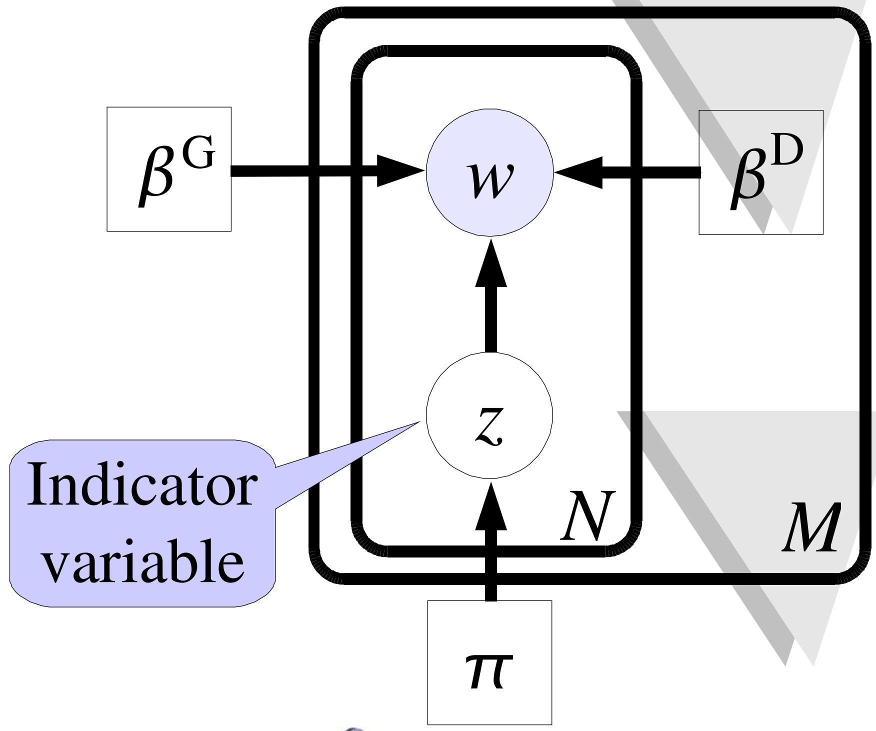
# Example for Summarization

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

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

$$p(w|\beta^G)^{z_{mn}}$$

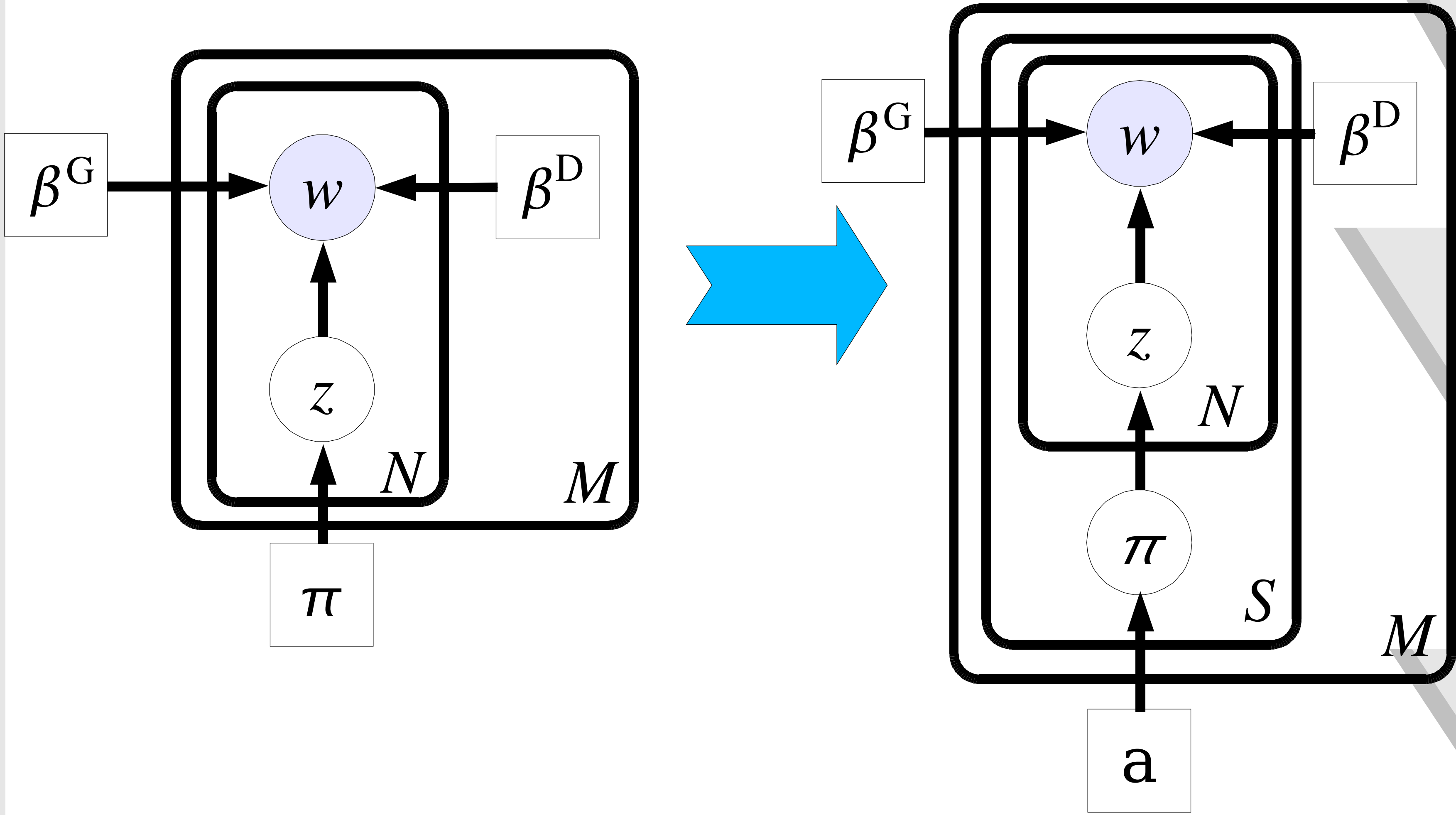
$$p(w|\beta^D_m)^{1-z_{mn}}$$



Hal Daumé III (me@hal3.name)

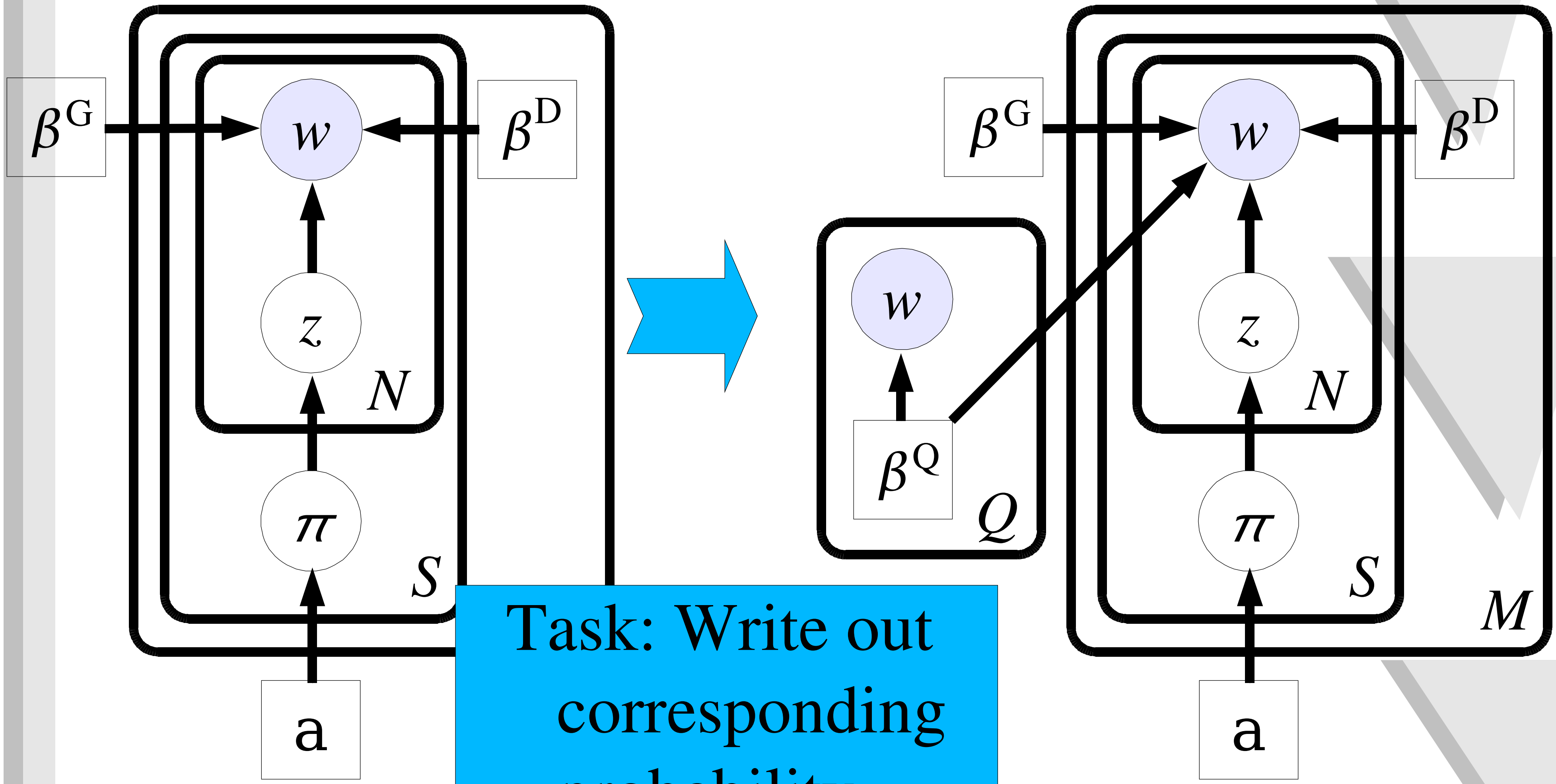
# Fun with Graphical Models

- 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
  - Laplace Approximation
  - Variational Approximation
  - Message Passing...
- Survey of Popular Models
- Pointers to Literature
- Conclusions

Hal Daumé III (me@hal3.name)

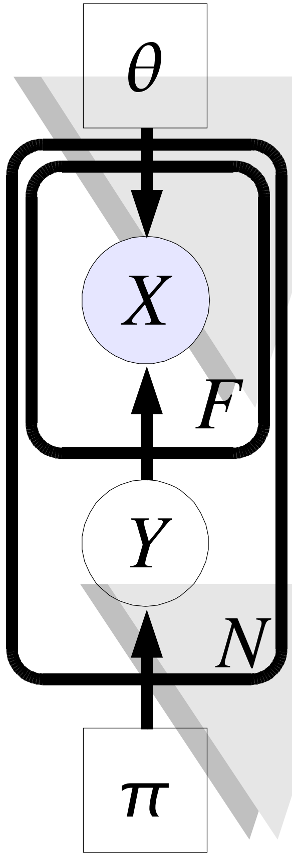
# Maximum Likelihood Estimators (MLE)

- 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):

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

$$l(\theta,\pi) = \sum_n \sum_k \left(Y_{nk} \log \pi_k + (1-Y_{nk}) \log(1-\pi_k)\right) + \sum_n \sum_f \left(X_{nf} \log \theta_f^{Y_n} + (1-X_{nf}) \log(1-\theta_f^{Y_n})\right)$$

$$\frac{\partial l}{\partial \pi} = \sum_n \sum_k \left[ \frac{Y_{nk}}{\pi_k} - \frac{1-Y_{nk}}{1-\pi_k} \right] \qquad \frac{\partial l}{\partial \theta^k} = \sum_{n:Y_n=k} \sum_f \left[ \frac{X_{nf}}{\theta_f^k} - \frac{1-X_{nf}}{1-\theta_f^k} \right]$$




See also: Was

# 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

Hal Daumé III (me@hal3.name)

# MLE with hidden variables

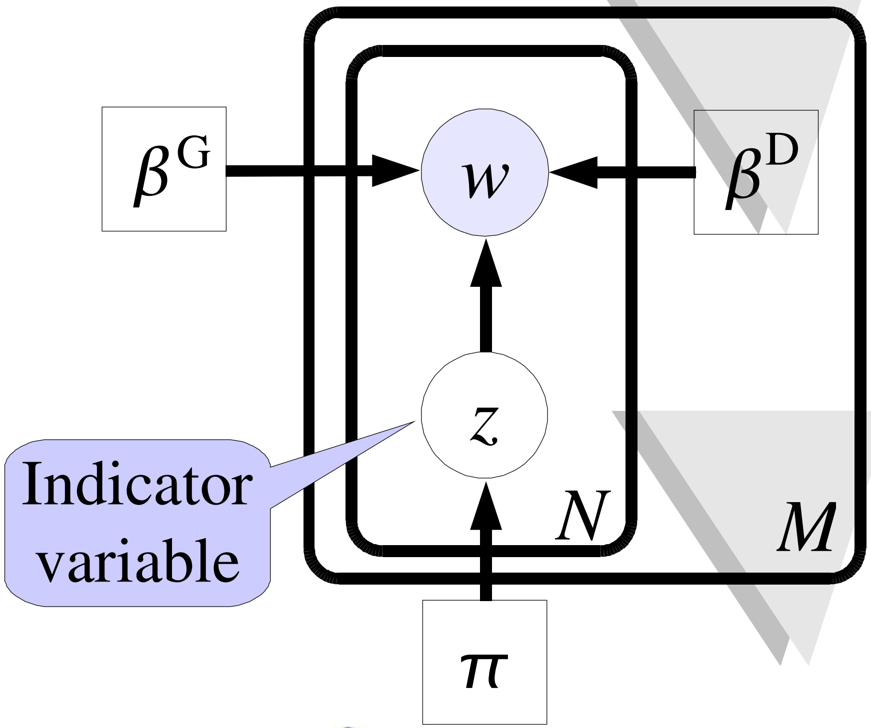
- Consider a stupid summarization model:
  - Each word in a document is drawn independently
  - Each word is draw either from a *general English* model, or a document specific model
  - We don't know which words are drawn from which

$$p(w \mid \pi, \beta^G, \beta^D) = \prod_m \prod_n \sum_{z_{mn}} p(z_{mn}|\pi)\, p(w|\beta^G)^{z_{mn}}\, p(w|\beta^D_m)^{1-z_{mn}}$$

  - Compute log likelihood:

$$l(\pi, \beta | w) = \sum_m \sum_n \log \sum_{z_{mn}} \ldots$$

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




*See also: Was, MK[22]*

Hal Daumé III (me@hal3.name)

# Expectation Maximization

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

$$
\begin{aligned}
\log p(x \mid \theta) &= \log \int_Z dz\, p(x, z \mid \theta) \\
&= \log \int_Z dz\, q(z) \frac{p(X, z \mid \theta)}{q(z)} \\
&\geq \int_Z dz\, q(z) \log \frac{p(X, z \mid \theta)}{q(z)} \\
&= \int_Z q(z) \log p(x, z \mid \theta) - \int_Z q(z) \log q(z) \\
&= \boldsymbol{E}_{z \sim q}\{\log p(x, z \mid \theta)\} - \boldsymbol{E}_{z \sim q}\{\log q(z)\}
\end{aligned}
$$

- For *any* distribution $Q$ (with the same support)
- How should we choose $Q$?

*See also: Was[9.13], Murphy, MK[22]*

Hal Daumé III (me@hal3.name)

# Expectation Maximization

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

$$
\begin{aligned}
\int_Z dz\, q(z) \log \frac{p(x, z \mid \theta)}{q(z)} &= \int_Z dz\, p(x \mid z, \theta) \log \frac{p(x, z \mid \theta)}{p(x \mid z, \theta)} \\
&= \int_Z dz\, p(x \mid z, \theta) \log \frac{p(z \mid x, \theta)\, p(x \mid \theta)}{p(x \mid z, \theta)} \\
&= \int_Z dz\, p(x \mid z, \theta) \log p(x \mid \theta) \\
&= \log p(x \mid \theta) \int_Z dz\, p(x \mid z, \theta) \\
&= \log p(x \mid \theta)
\end{aligned}
$$

- So, when computing $\boldsymbol{E}_{z \sim q}\{\log p(x, z \mid \theta)\}$, the expectation should be taken with respect to the true posterior

*See also: Was[9.13], Murphy, MK[22]*

# EM in Practice

- Recall, we wanted to estimate parameters for:

$$\begin{aligned} p(w \mid \pi, \beta^G, \beta^D) &= \prod_m \prod_n \sum_{z_{mn}} p(z_{mn}|\pi)\, p(w|\beta^G)^{z_{mn}}\, p(w|\beta^D_m)^{1-z_{mn}} \\ &= \prod_m \prod_n \boldsymbol{E}_{z_{mn}\sim\pi} \{ p(w|\beta^G)^{z_{mn}}\, p(w|\beta^D_m)^{1-z_{mn}} \} \end{aligned}$$

- So we replace the hidden variables with their expectations:

$$l(\beta \mid w) \geq \sum_m \sum_n \boldsymbol{E}\{z_{mn}\} \log p(w|\beta^G) + (1 - \boldsymbol{E}\{z_{mn}\}) \log p(w|\beta^D_m)$$

- All we need to do is calculate the expectations:

$$\boldsymbol{E}\{z_{mn}\} \propto p(z_{mn}=1 \mid \pi)\, p(w \mid \beta^G)$$

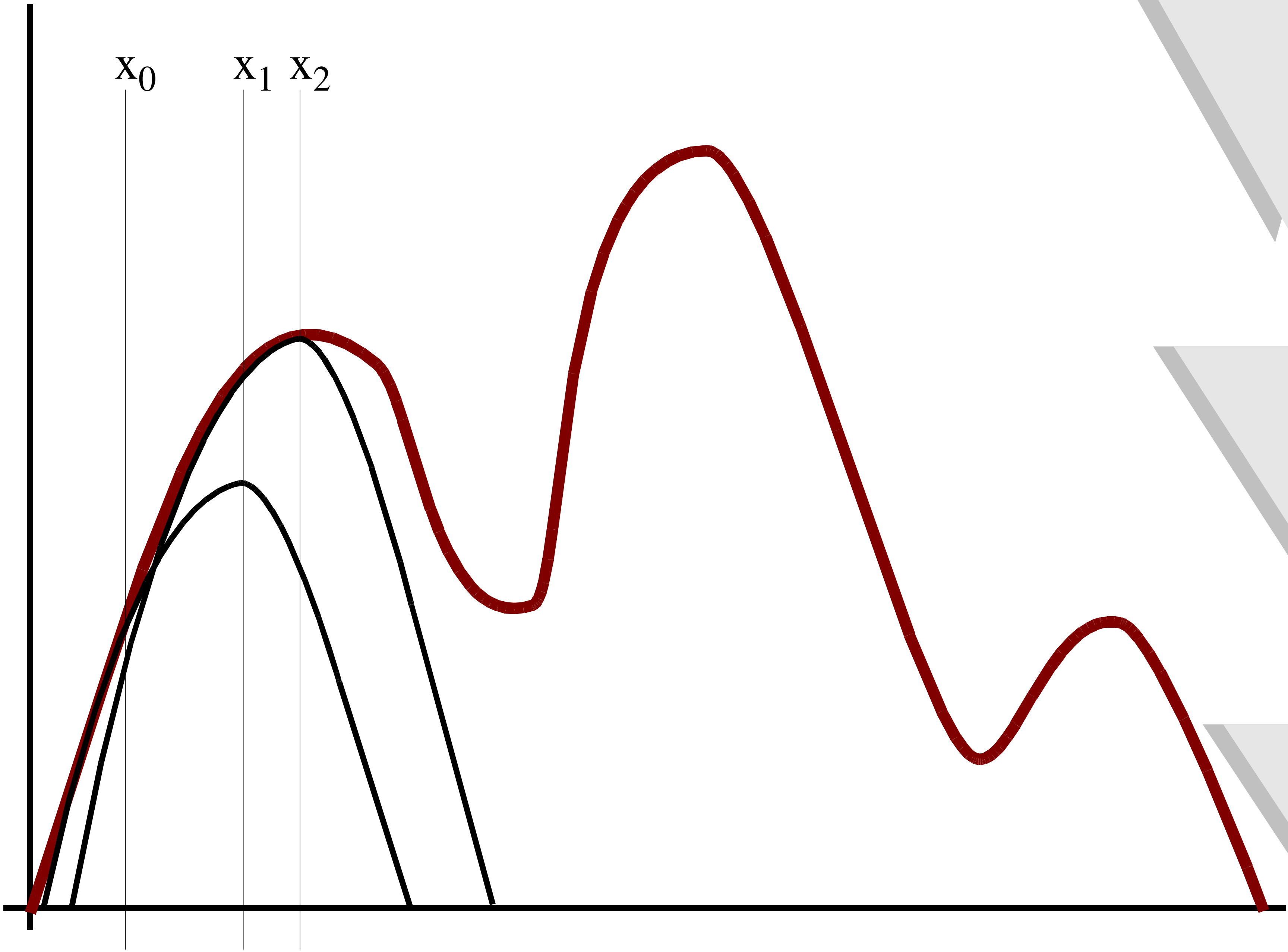
- And now the computation proceeds as in the no-hidden-variable setting

# EM Summed Up

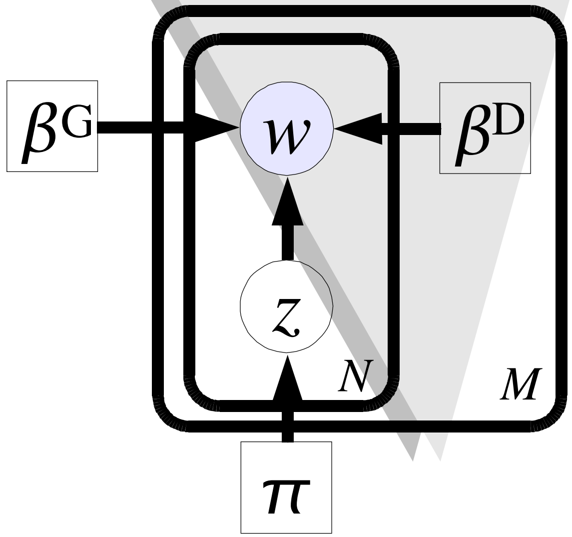
- Initialize parameters however you desire
- Repeat:
  - *E-STEP:*
    Compute expectations of hidden variables under the current parameter settings
  - *M-STEP:*
    Optimize parameters given those expectation

- 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: $\boldsymbol{E}\{z_{mn}\} \propto p(z_{mn}=1 \mid \pi)\, p(w \mid \beta^G)$

$$\boldsymbol{E}\{z_{11}\} = \frac{\pi \beta^G_A}{\pi \beta^G_A + (1-\pi)\beta^D_{1A}} = \frac{0.5*1/3}{0.5*1/3+0.5*1/3} = 0.5$$

$$\boldsymbol{E}\{z_{12}\} = \boldsymbol{E}\{z_{21}\} = \boldsymbol{E}\{z_{22}\} = 0.5$$

  - M-step:

$$\beta^G_A = \frac{1}{Z}\left[\boldsymbol{E}\{z_{11}\} + \boldsymbol{E}\{z_{21}\}\right] = \frac{1}{2} \qquad \beta^G_B = \frac{1}{Z}\left[\boldsymbol{E}\{z_{12}\}\right] = \frac{1}{4} \qquad \beta^G_C = \frac{1}{Z}\left[\boldsymbol{E}\{z_{22}\}\right] = \frac{1}{4}$$

$$\beta^D_{1A} = \frac{1}{Z}\left[1 - \boldsymbol{E}\{z_{11}\}\right] = \frac{1}{2} \qquad \beta^D_{1B} = \frac{1}{Z}\left[1 - \boldsymbol{E}\{z_{12}\}\right] = \frac{1}{2} \qquad \beta^G_{1C} = 0$$

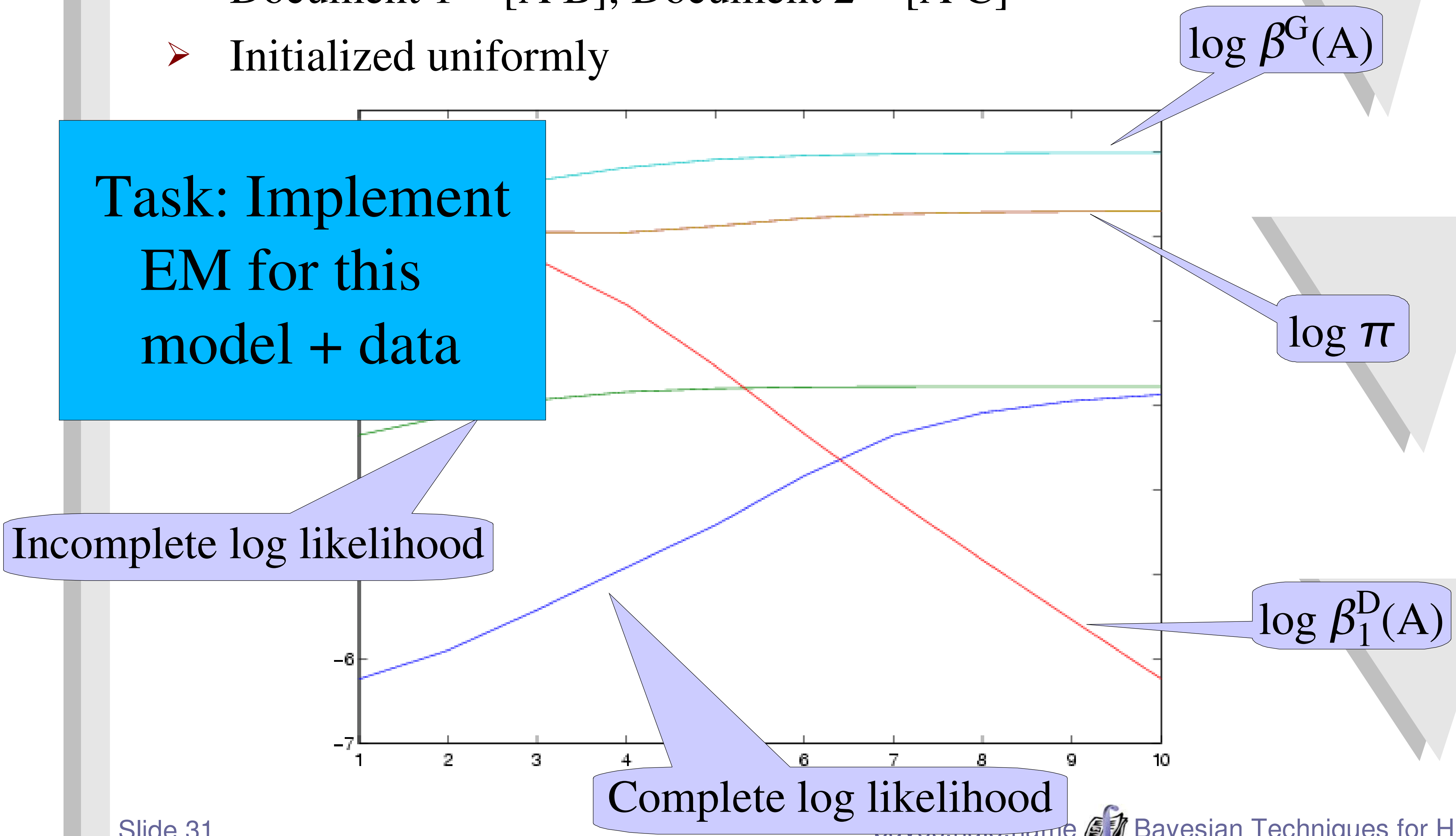
$$\beta^D_{2A} = \frac{1}{Z}\left[1 - \boldsymbol{E}\{z_{21}\}\right] = \frac{1}{2} \qquad \beta^D_{2B} = 0 \qquad \beta^G_{2C} = \frac{1}{Z}\left[1 - \boldsymbol{E}\{z_{22}\}\right] = \frac{1}{2}$$

$$\pi = \frac{E\{z_{11}\} + E\{z_{21}\}}{E\{z_{11}\} + E\{z_{21}\} + E\{z_{12}\} + E\{z_{22}\}} = \frac{1}{2}$$

# 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



Hal Daumé III (me@hal3.name)

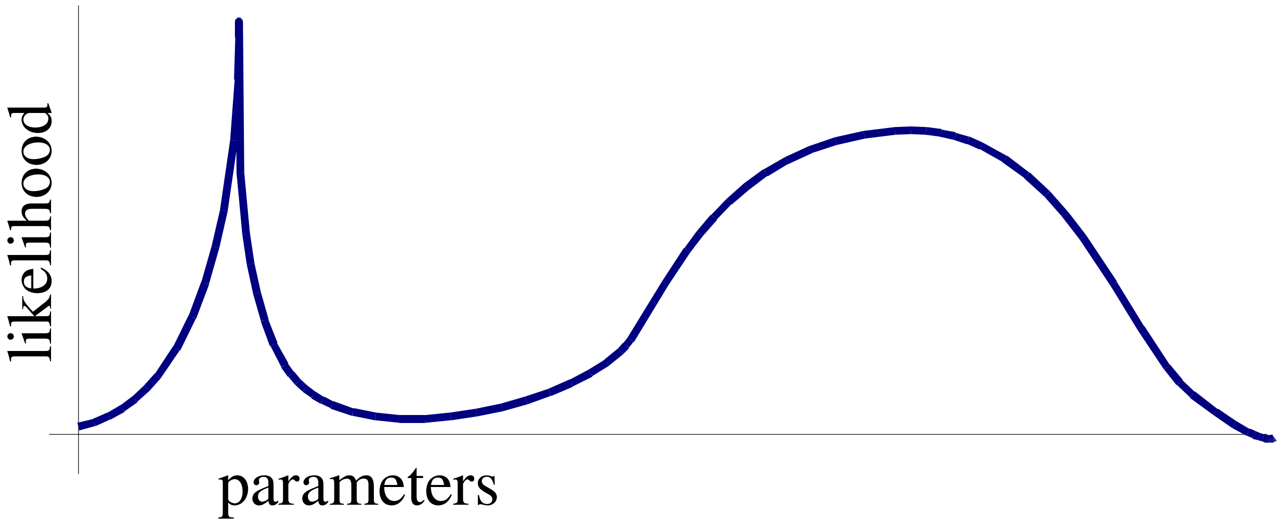
# Problems with Maximum Likelihood

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

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

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

**Background knowledge is 0/1**




*See also: MK[3], Ber[3.1, 2.3, 4.8]*

# 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

Hal Daumé III (me@hal3.name)

# What is a Prior?

- Recall Bayes' Rule:

Prior

Likelihood

Posterior

$$P(\theta \mid D) = \frac{P(\theta)\ P(D \mid \theta)}{\int_{\Theta} d\theta\, P(\theta)\ P(D \mid \theta)}$$

Marginal

- A *prior* is a specification of our *beliefs* about the values parameters can take, *before seeing any data*

*See also: Ber[3.1-6, 4.7], MK[22,43], , Was[11.1]*

Hal Daumé III (me@hal3.name)

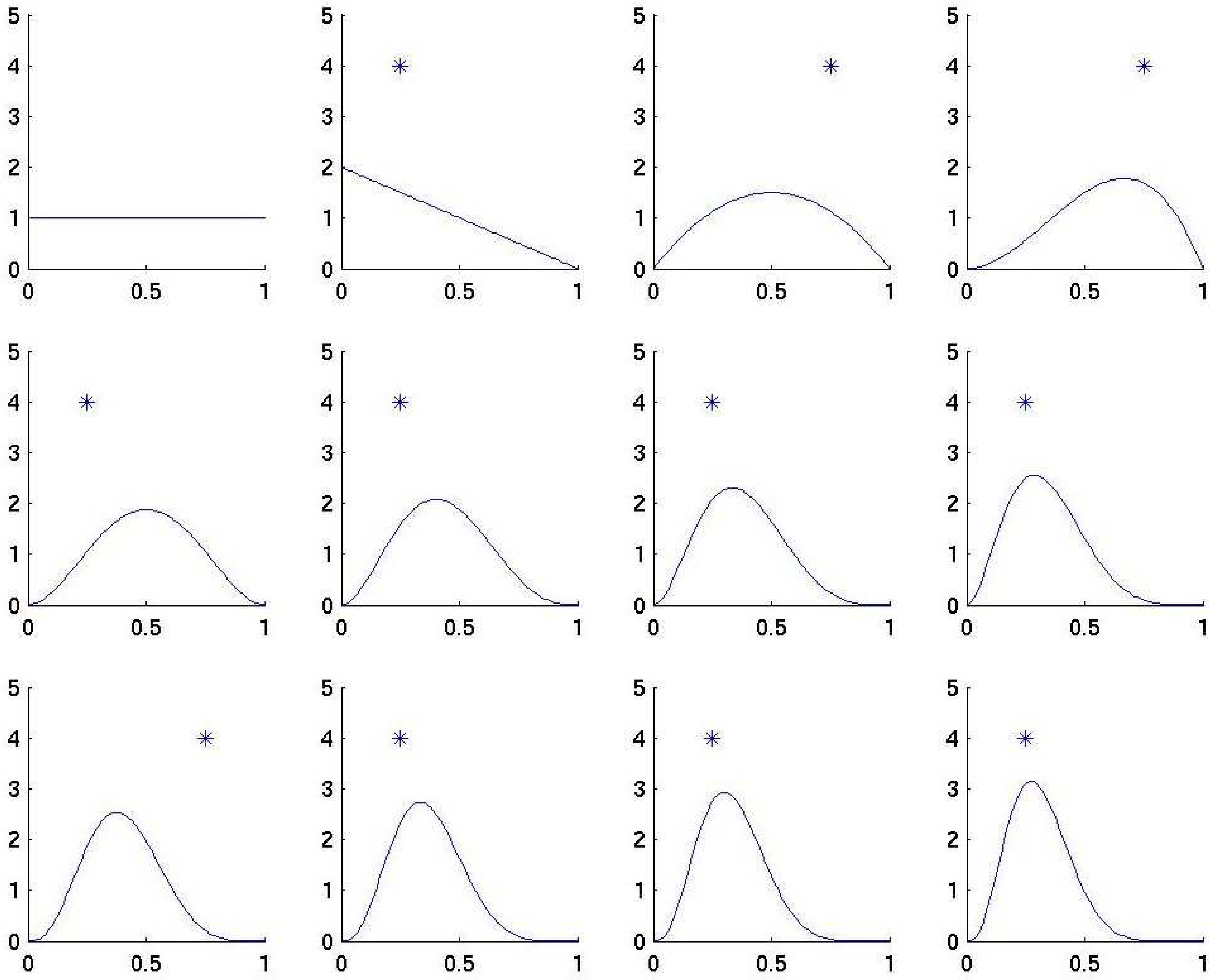
# How Does the Posterior Behave?

Take sequence of data $x_1$, ..., $x_N$...

$$
\begin{aligned}
p(\theta) &= \textit{just the prior} \\
p(\theta \mid x_1) &= \frac{p(\theta)\, p(x_1 \mid \theta)}{\int d\theta\, p(\theta)\, p(x_1 \mid \theta)} \\
p(\theta \mid x_1, x_2) &= \frac{p(\theta \mid x_1)\, p(x_2 \mid \theta)}{\int d\theta\, p(\theta \mid x_1)\, p(x_2 \mid \theta)} \\
&\vdots \\
p(\theta \mid x_{1:N}) &= \frac{p(\theta \mid x_{1:N-1})\, p(x_N \mid \theta)}{\int d\theta\, p(\theta \mid x_{1:N-1})\, p(x_N \mid \theta)} \\
&= \frac{p(\theta) \prod_n p(x_n \mid \theta)}{\int d\theta\, p(\theta) \prod_n p(x_n \mid \theta)}
\end{aligned}
$$

Hal Daumé III (me@hal3.name)

# Binomial Example



Hal Daumé III (me@hal3.name)

# 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 $\pi(\theta) = \pi(\theta \mid \alpha)$ 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):

$$\Gamma = \{ \pi : \pi(\theta) = g(\theta \mid \alpha) \}$$

- Estimate the prior by maximizing the marginal likelihood:

$$\begin{aligned} \hat{\pi} &= \max_{\pi \in \Gamma} p(x \mid \pi) \\ &= \max_{\alpha \in A} \int_{\Theta} d\theta \, \pi(\theta \mid \alpha) p(x \mid \theta) \end{aligned}$$

Hal Daumé III (me@hal3.name)

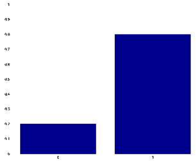
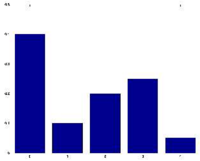
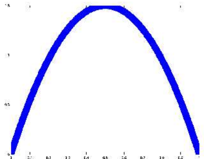
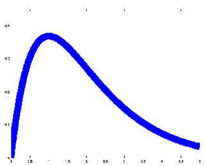
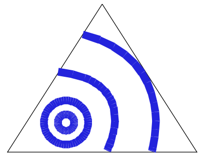
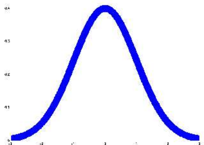
# Conjugate (convenient) Priors

- Recall: $p(\theta \mid x_{1:N}) = \dfrac{p(\theta)\prod_n p(x_n \mid \theta)}{\int d\theta\, p(\theta)\prod_n p(x_n \mid \theta)}$
- Given a distribution $p(x \mid \theta)$
- And a prior $\pi(\theta \mid \alpha)$
- The prior is *conjugate* if:

$$p(\theta \mid \alpha, x) = \frac{\pi(\theta \mid \alpha)\, p(x \mid \theta)}{\int_{\Theta} F^{\pi(\alpha)}(\theta)\, p(x \mid \theta)} = \pi(\theta \mid \mathring{\alpha})$$

*See also: Ber[4.2.2], MK[23]*

# Summary of Distributions

| Distribution | Domain | Picture | Parametric Form |
|---|---|---|---|
| Binomial | Binary |  | $Bin(x \mid N, \theta) \propto \theta^{n} (1-\theta)^{N-n}$ |
| Multinomial | K classes |  | $Mult(\bar{x} \mid \bar{\theta}) \propto \prod \theta_k^{x_k}$ |
| Beta | [0,1] |  | $Beta(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1} (1-\theta)^{\beta-1}$ |
| Gamma | $[0,\infty)$ |  | $Gam(x \mid a, b) \propto x^{-a-1} \exp(-bx)$ |
| Dirichlet | Simplex |  | $Dir(\bar{\theta} \mid \bar{\alpha}) \propto \prod \theta_k^{\alpha_k-1}$ |
| Gaussian | Reals |  | $Nor(x \mid \mu, \sigma^2) \propto \exp\left((x-\mu)^2/2\sigma^2\right)$ |

Hal Daumé III (me@hal3.name)

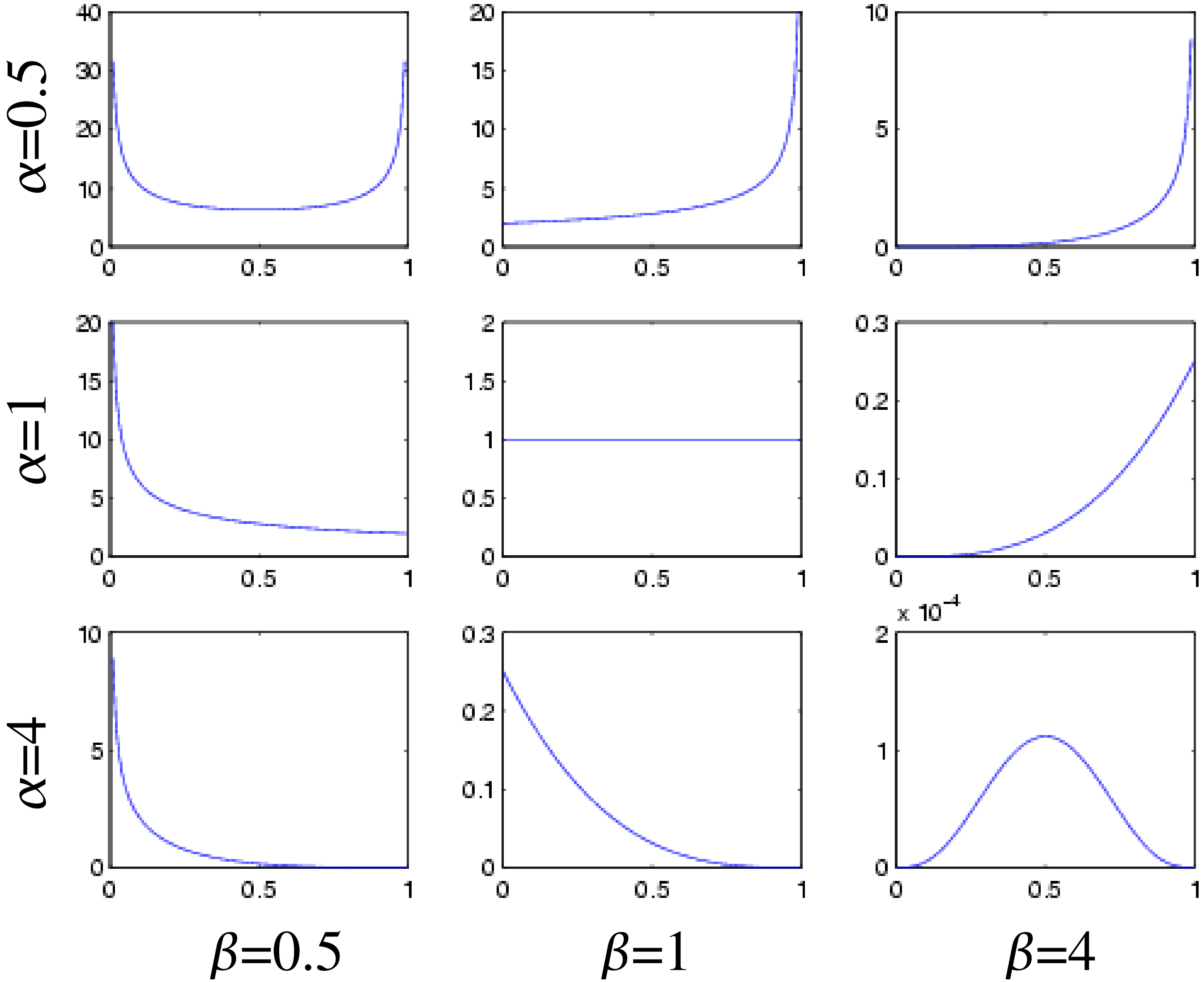
# Binomial and Beta Distributions

- 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: $Bin(x \mid N, \theta) = \binom{N}{x} \theta^n (1-\theta)^{N-n}$
  - 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: $Beta(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$
  - Moments: $\mu = \frac{\alpha}{\alpha+\beta}, \ \ var = \frac{\alpha\beta}{(\alpha+\beta)^2 (\alpha+\beta+1)}$
- Beta is conjugate to binomial:
  - Posterior parameters: $\mathring{\alpha} = \alpha + x, \ \ \mathring{\beta} = \beta + N - x$
  - 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

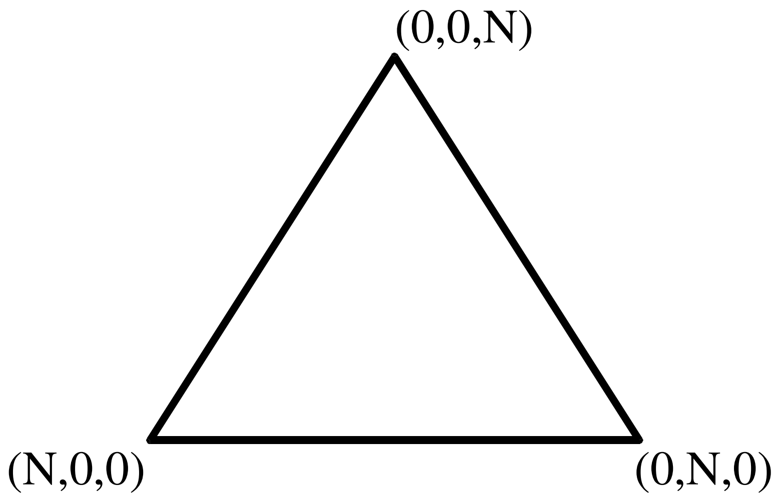
$$Beta(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$



Hal Daumé III (me@hal3.name)

# Multinomial Distribution

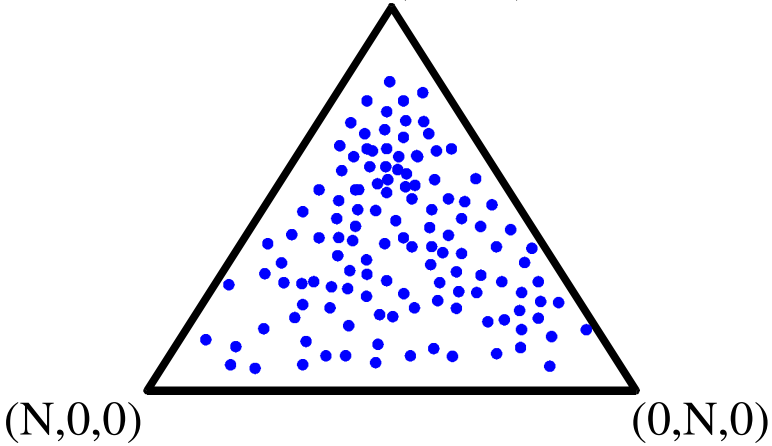
- A distribution over counts of K>1 discrete events (words)
- Domain: $\langle x_1, \ldots, x_K \rangle \in \mathbb{N}^K$
- Parameters: $\langle \theta_1, \ldots, \theta_K \rangle \in \Delta_K = \{\theta_{1:K} : \theta_k \geq 0, \sum_k \theta_k = 1\}$
- Distribution: $Mult(\bar{x} \mid \bar{\theta}) = \frac{\Gamma\left(\sum x_k + 1\right)}{\prod \Gamma(x_k + 1)} \prod \theta_k^{x_k}$
- Moments: $\langle \theta_1, \ldots, \theta_K \rangle \in \Delta_K = \{\theta_{1:K} : \theta_k \geq 0, \sum_k \theta_k = 1\}$



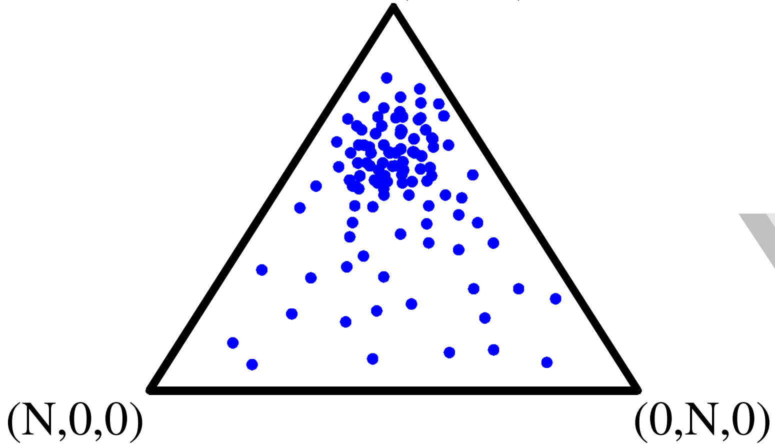
Hal Daumé III (me@hal3.name)

# Dirichlet Distribution

- A distribution over a probability simplex
- Domain: $\langle \theta_{1,}\ldots,\theta_K \rangle \in \delta^K$
- Parameters: $\langle \alpha_1,\ldots,\alpha_K \rangle \in (\mathbb{R}^+)^K \;,\quad \hat{\alpha} = \sum_k \alpha_k$
- Distribution: $Dir(\bar{\theta} \mid \bar{\alpha}) = \frac{\Gamma(\hat{\alpha})}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$
- Moments: $\mu_k = \frac{\alpha_k}{\hat{\alpha}},\quad var_k = \frac{\alpha_k(\hat{\alpha}-\alpha_k)}{(\hat{\alpha})^2(1+\hat{\alpha})}$




[1,1,2]




[5,5,10]

# Multinomial/Dirichlet Pair

- Multinomial distribution: $Mult(\bar{x} \mid \bar{\theta}) = \frac{\Gamma\left(\sum x_k + 1\right)}{\prod \Gamma(x_k + 1)} \prod \theta_k^{x_k}$
- Dirichlet distribution: $Dir(\bar{\theta} \mid \bar{\alpha}) = \frac{\Gamma(\hat{\alpha})}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$
- Posterior hyper-parameters:

$$\langle \mathring{\alpha}_1, \ldots, \mathring{\alpha}_K \rangle = \langle \alpha_1 + x_1, \ldots, \alpha_K + x_K \rangle$$

- Marginal Distribution:

$$p(\bar{x} \mid \bar{\alpha}) = \frac{\Gamma\left(\sum x_k + 1\right)}{\prod \Gamma(x_k + 1)} \frac{\Gamma(\hat{\alpha})}{\prod \Gamma(\alpha_k)} \frac{\prod \Gamma(\alpha_k + x_k)}{\Gamma\left(\hat{\alpha} + \sum x_k\right)}$$

# Gaussian/Gaussian-Gamma

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

$$a > 0,\ \ b > 0,\ \ domain = \mathbb{R}^+$$

- Posterior hyper-parameters:

$$\mathring{s} = \left(\frac{1}{s^2} + \frac{1}{\sigma^2}\right)^{-1/2} \qquad \mathring{m} = \frac{m/s^2 + \sum_i x_i/\sigma^2}{1/s^2 + N/\sigma^2}$$

$$\mathring{a} = a + 1/2 \qquad \mathring{b} = \left(b^{-1} + \frac{1}{2}\sum_i (x_i - \bar{x})^2\right)^{-1}$$

- Marginal distribution:

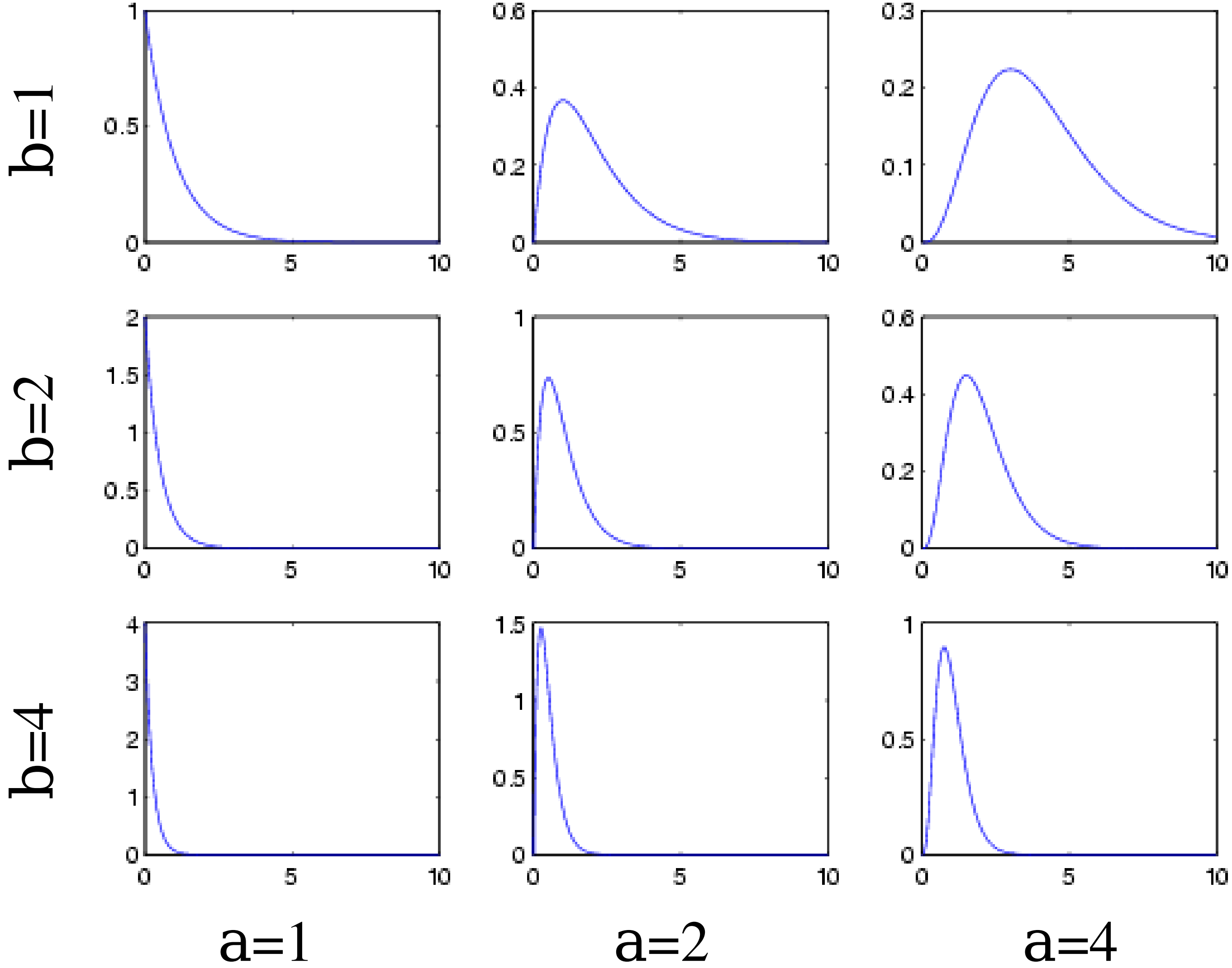
$$p(x \mid m, s^2, a, b) = StuT(m, a, b)$$

Non-standard
Student's T distribution

# Gamma Distribution

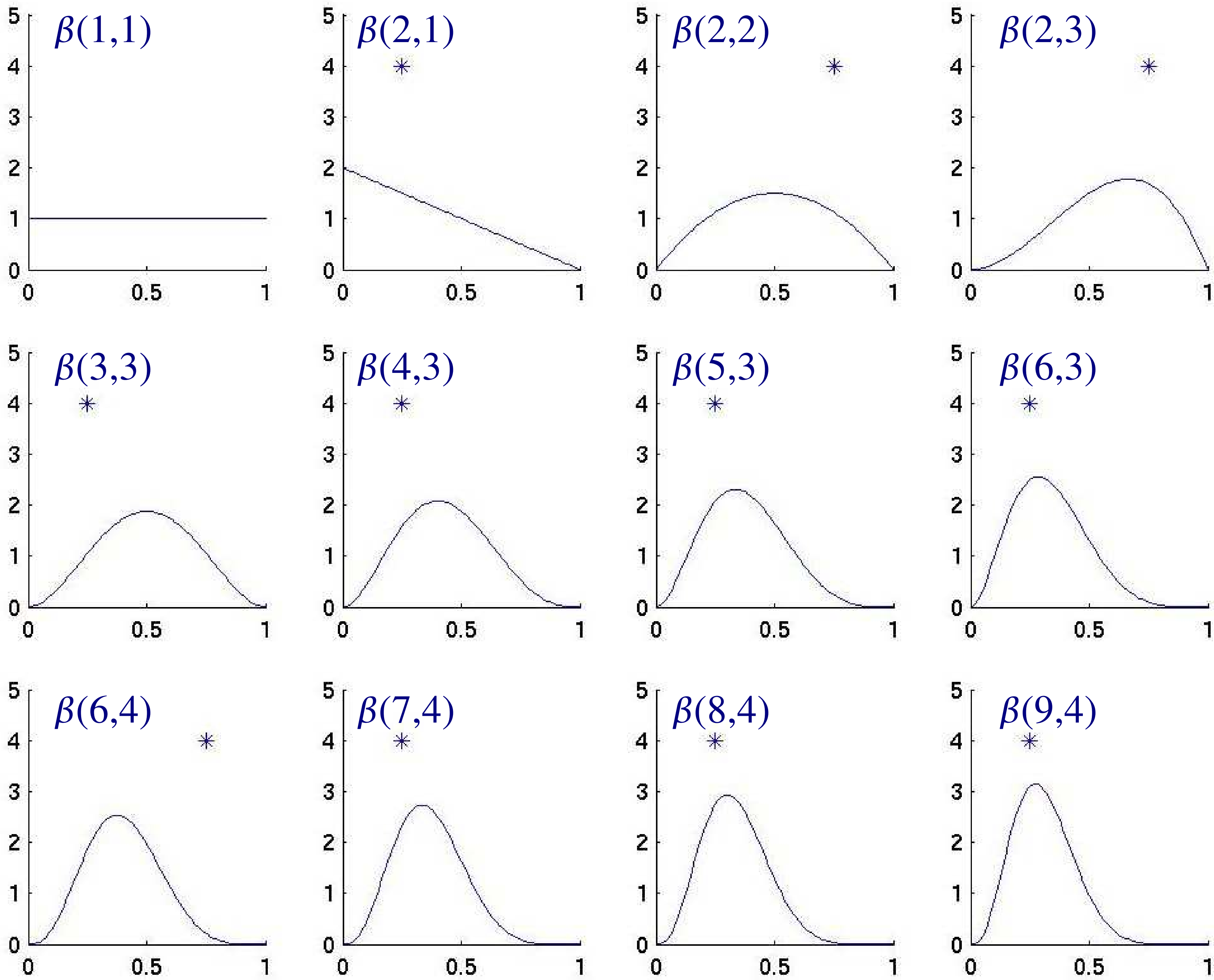
$$Gam(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp(-bx)$$

$$\begin{aligned} \mu &= a/b \\ var &= a/b^2 \end{aligned}$$



Hal Daumé III (me@hal3.name)

# Conjugate Priors in Action



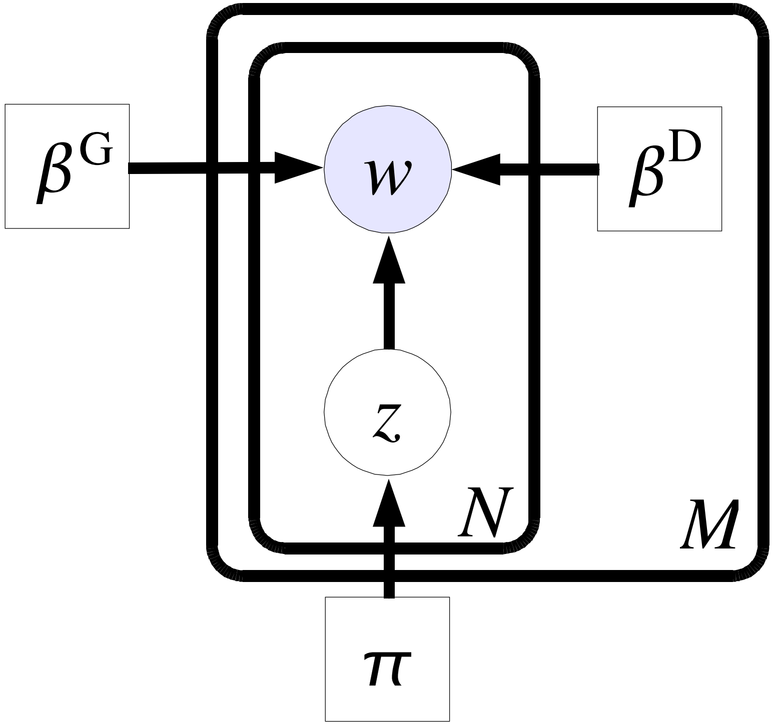
Hal Daumé III (me@hal3.name)

# 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

Hal Daumé III (me@hal3.name)

# Recall our summarization model

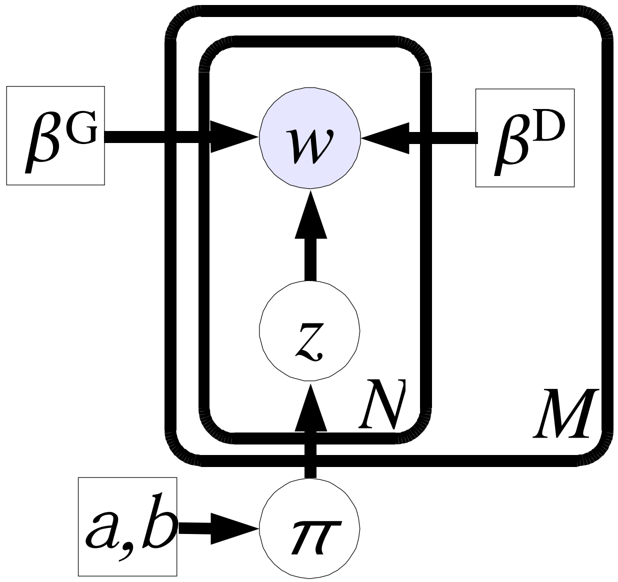



$$
\begin{array}{rcll}
z & | & \pi & \sim & Bin(\pi) \\
w & | & z,\beta & \sim & Mult(\beta^G)^z\, Mult(\beta^D)^{1-z}
\end{array}
$$

- 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?

Hal Daumé III (me@hal3.name)

# Bayesianified summarization model




$$
\begin{array}{lcl}
\pi \mid a,b & \sim & Beta(a,b) \\
z \mid \pi & \sim & Bin(\pi) \\
w \mid z,\beta & \sim & Mult(\beta^G)^z \, Mult(\beta^D)^{1-z}
\end{array}
$$

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

Conjugacy does not help because of the hidden variables

Hal Daumé III (me@hal3.name)

# Interesting Inference Questions

- Predict values of unobserved data:

$$P(U \mid D) \propto \int_{\Theta} d\pi(\theta) P(D \mid \theta) P(U \mid \theta)$$

- Compute data likelihood:

$$P(D) \propto \int_{\Theta} d\pi(\theta) P(D \mid \theta)$$

- Maximize marginal likelihood:

$$P(\alpha \mid D) \propto \int_{\Theta} d\pi(\theta \mid \alpha) P(D \mid \theta)$$

- Estimate posterior:

$$P(\theta \mid D) = \frac{\pi(\theta) P(D \mid \theta)}{P(D)}$$

- **GENERAL FORM:**

$$F = \int_X dx\, p(x) f(x) = \boldsymbol{E}_{x \sim p}\{f(x)\}$$

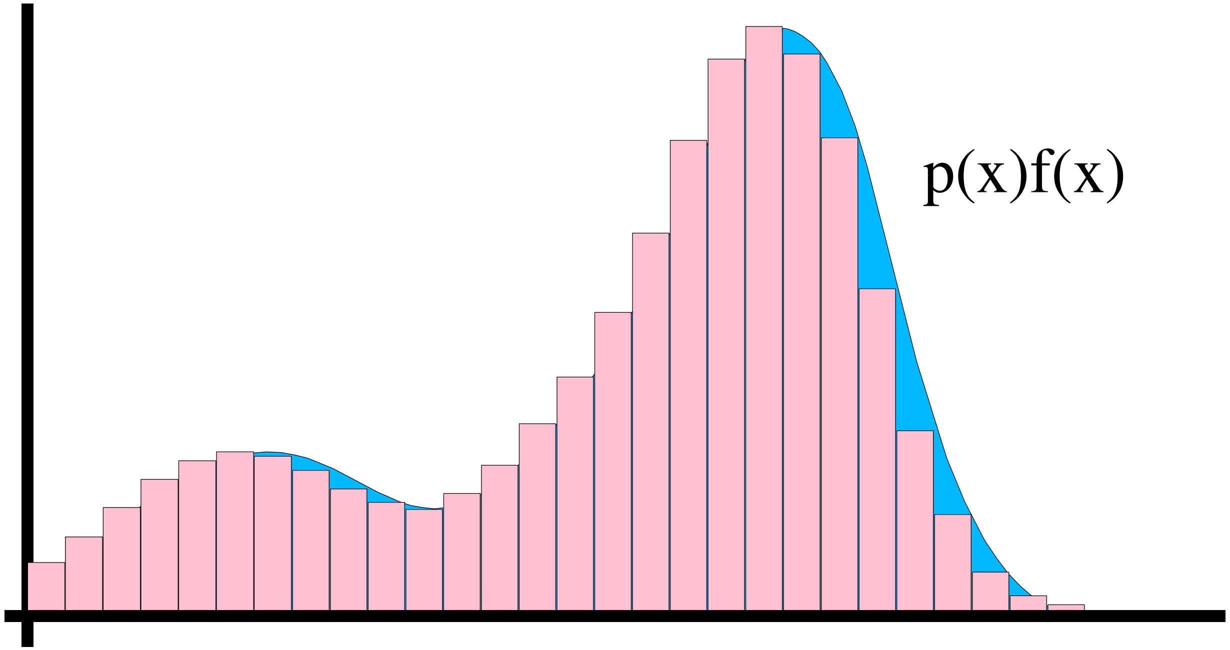
Hal Daumé III (me@hal3.name)

# 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:

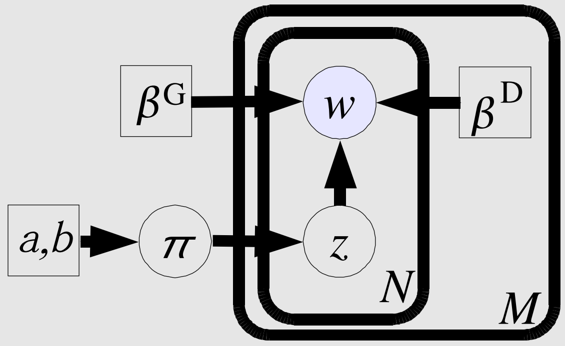



$$F = \int_X dx\, p(x)\, f(x) \approx \frac{1}{R} \sum_{x \in R} p(x)\, f(x)$$

*See also: MK[24]*

# Summing in our Model

- Simply rewrite the integral as a sum:

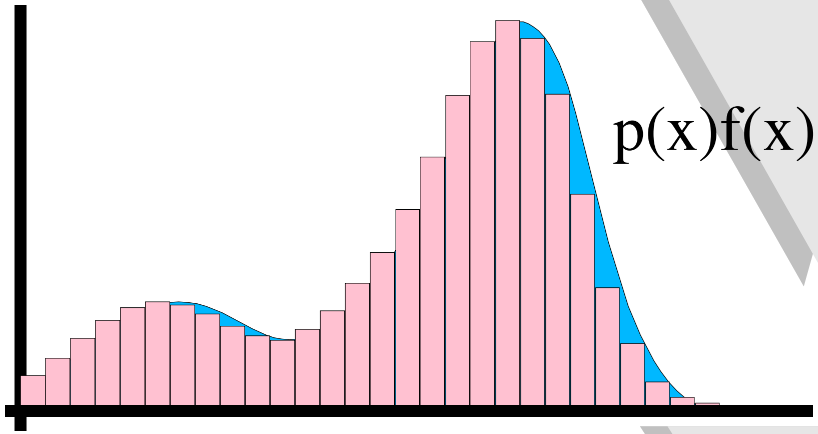



$$
\begin{array}{lcl}
\pi & | \; a,b & \sim \; Beta(a,b) \\
z & | \; \pi & \sim \; Bin(\pi) \\
w & | \; z,\beta & \sim \; Mult(\beta^G)^z Mult(\beta^D)^{1-z}
\end{array}
$$

$$
\begin{aligned}
p(D|\beta,a,b) &= \int_U d\pi \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1}(1-\pi)^{b-1} \\
&\quad \prod_m \prod_n \sum_{z_{mn}\in\{0,1\}} \pi^{z_{mn}}(1-\pi)^{1-z_{mn}} \prod_v (\beta_v^G)^{z_{mn} w_{mnv}} (\beta_{dv}^D)^{(1-z_{mn}) w_{mnv}} \\
&\approx \frac{\sum_{p=1}^{100} \Gamma(a+b)}{\Gamma(a)\Gamma(b)} (p/100)^{a-1}(1-p/100)^{b-1} \\
&\quad \prod_m \prod_n \sum_{z_{mn}\in\{0,1\}} (p/100)^{z_{mn}}(1-p/100)^{1-z_{mn}} \\
&\quad \prod_v (\beta_v^G)^{z_{mn} w_{mnv}} (\beta_{dv}^D)^{(1-z_{mn}) w_{mnv}}
\end{aligned}
$$

# 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




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

- Idea: let's choose R differently

Hal Daumé III (me@hal3.name)

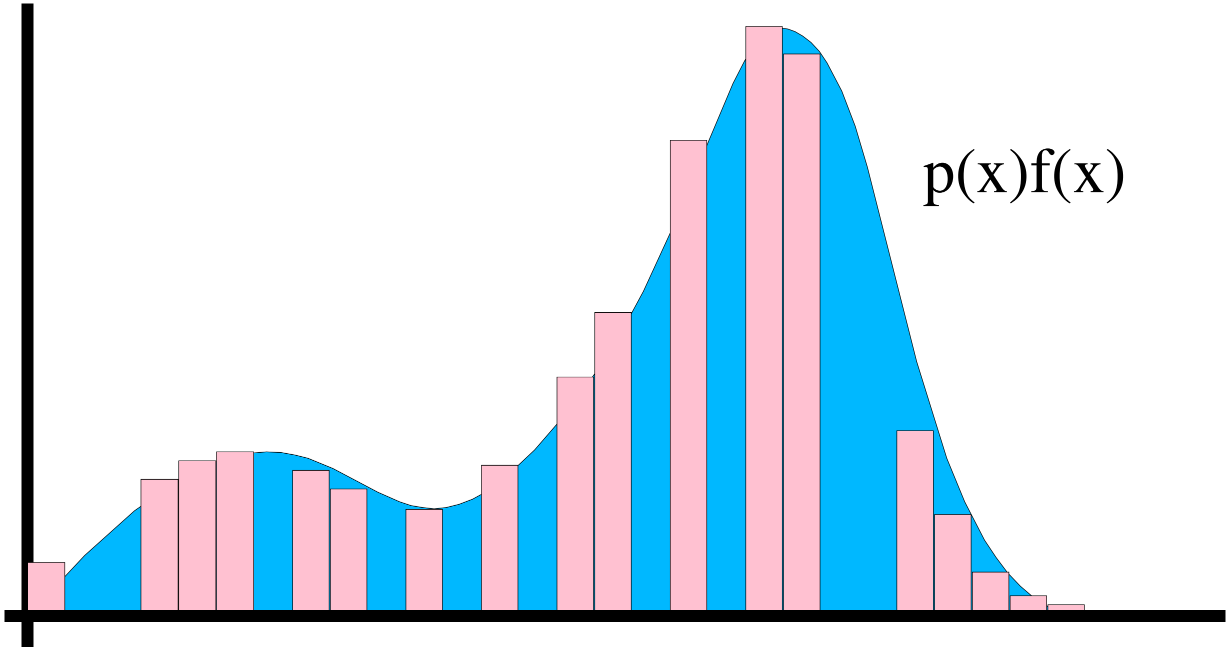
# Tutorial Outline

- 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

Hal Daumé III (me@hal3.name)

# Monte Carlo Integration

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

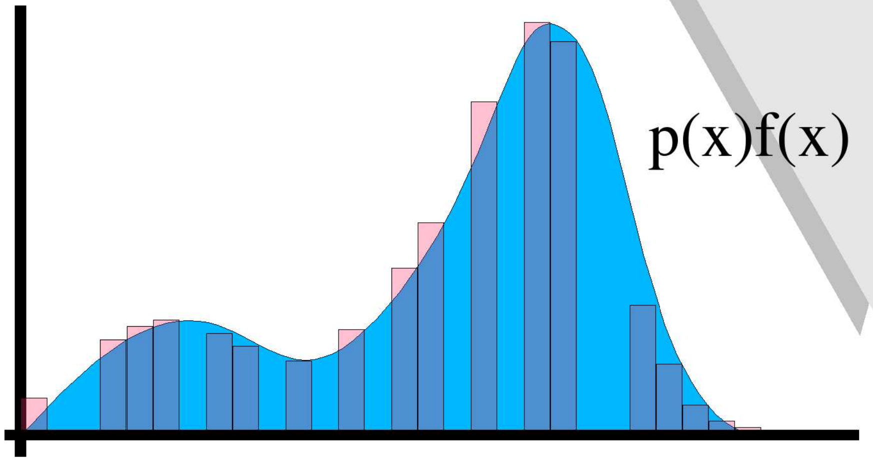



$$F = \int_X dx\, p(x)\, f(x) \approx \frac{1}{R} \sum_{x \in R} p(x)\, f(x)$$

See also: MK[29], Was[24.2], And03

# Uniform Sampling

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



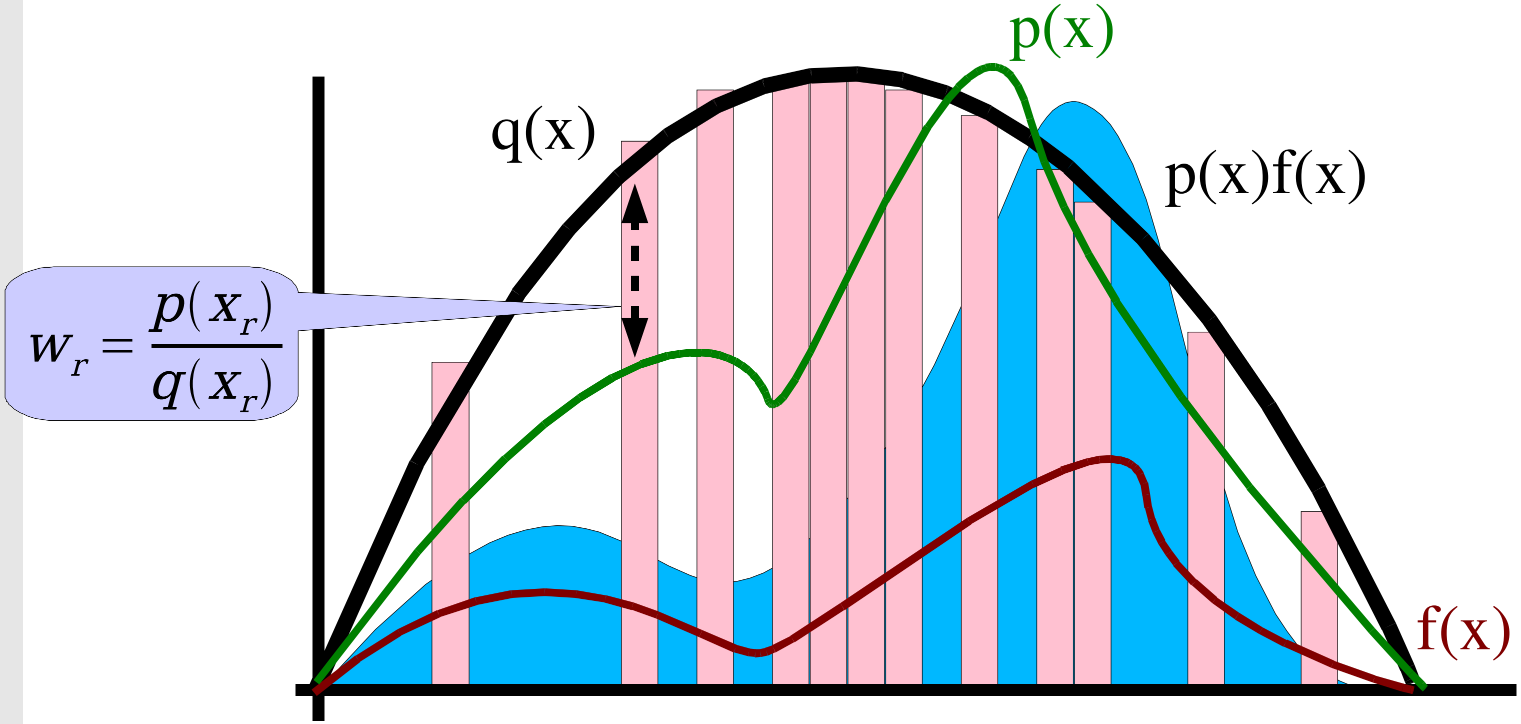

$$F = \int_X dx\, p(x)\, f(x) \approx \frac{1}{R} \sum_{x \in R} p(x)\, f(x)$$

- 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

*See also: MK[29], Was[24.2], And03*

# Importance Sampling

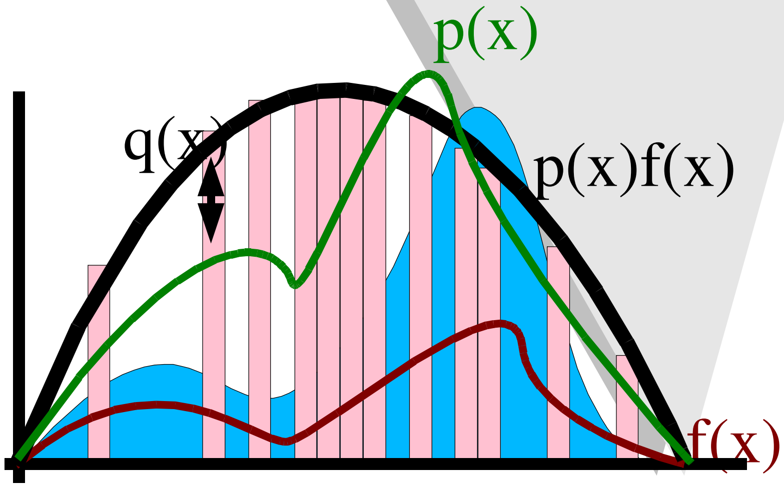
- Let R be a set of points drawn from a proposal distribution q




$$F = \int_X dx\, p(x)\, f(x) \approx \frac{\sum_r w_r f(x_r)}{\sum_r w_r} \quad , \quad w_r = \frac{p(x_r)}{q(x_r)}$$

*See also: MK[29], Was[24.3], And03*

# Importance Sampling




$$F \approx \frac{\sum w_r f(x_r)}{\sum w_r}, \quad w_r = \frac{p(x_r)}{q(x_r)}$$

- 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: $R \sim 2^{D/2}$

Hal Daumé III (me@hal3.name)

# 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

# Markov Chain Monte Carlo

- 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 $q(\ |\ x)$ 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 $x = x'$

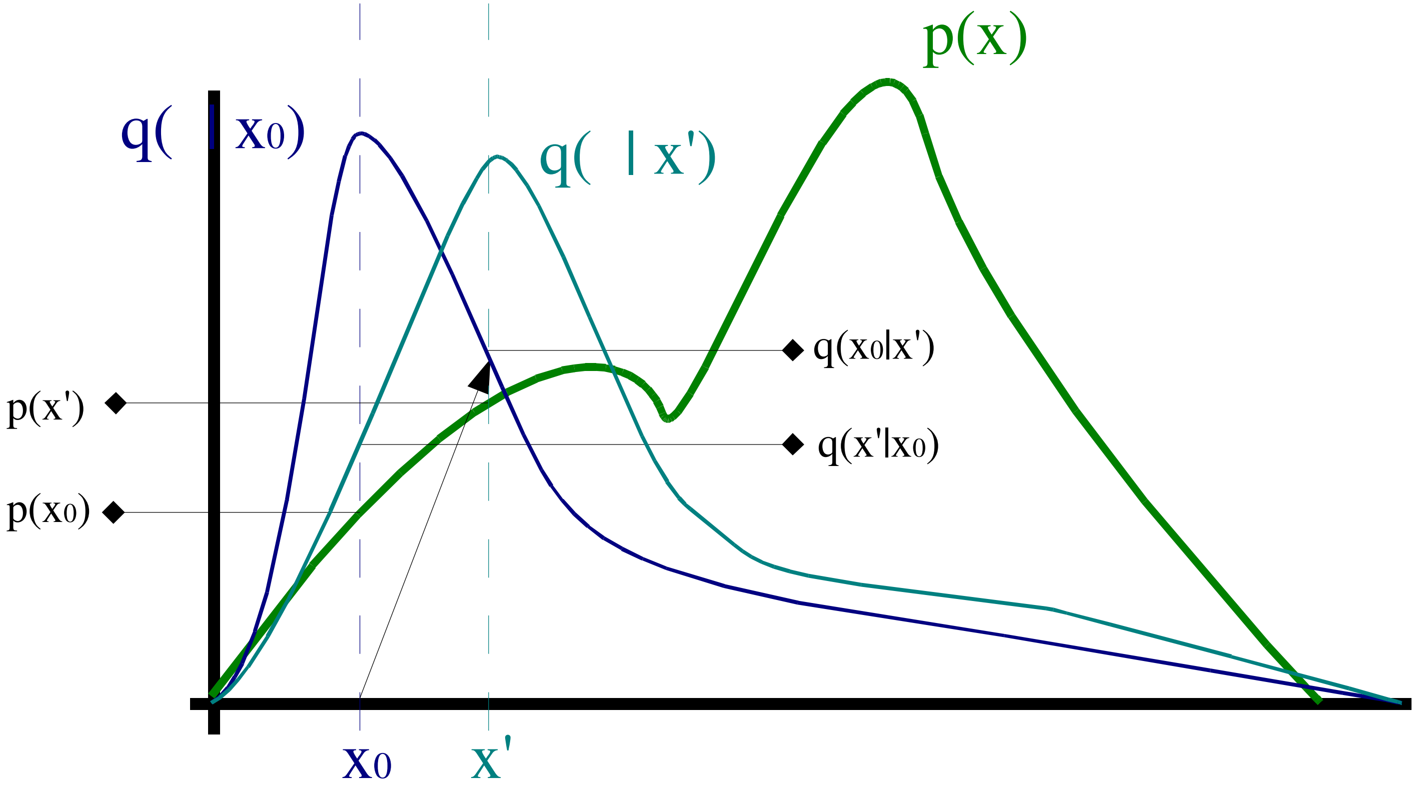
- *New problem:* samples are no longer independent!

*See also: MK[30], Was[24.4], And03*

Hal Daumé III (me@hal3.name)

# Metropolis-Hastings Sampling

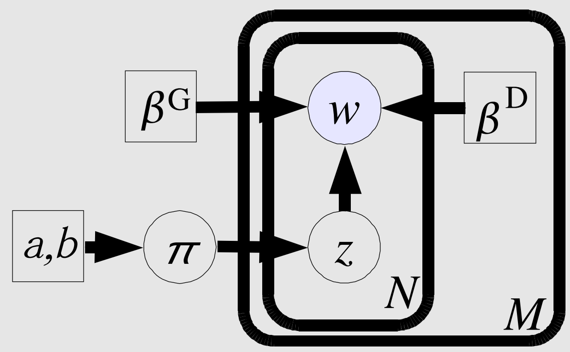
- Accept new states with probability: $min\left\{1, \frac{p(x')}{p(x)} \frac{q(x \mid x')}{q(x' \mid x)}\right\}$
- Only put every N$^{th}$ sample into R




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

*See also: MK[30], Was[24.4], And03*

# MH in our Model




$$\begin{array}{lcll} \pi & \mid & a,b & \sim & Beta(a,b) \\ z & \mid & \pi & \sim & Bin(\pi) \\ w & \mid & z,\beta & \sim & Mult(\beta^G)^z Mult(\beta^D)^{1-z} \end{array}$$

- Invent a proposal distribution $q$

$$\begin{array}{lcll} \log a' & \mid & a & \sim & Nor(\log(a),1) \\ \log b' & \mid & b & \sim & Nor(\log(b),1) \\ \sigma(\pi') & \mid & \pi & \sim & Nor(\sigma(\pi),1) \\ z_{mn}' & \mid & z & \sim & Bin(0.5) \end{array}$$

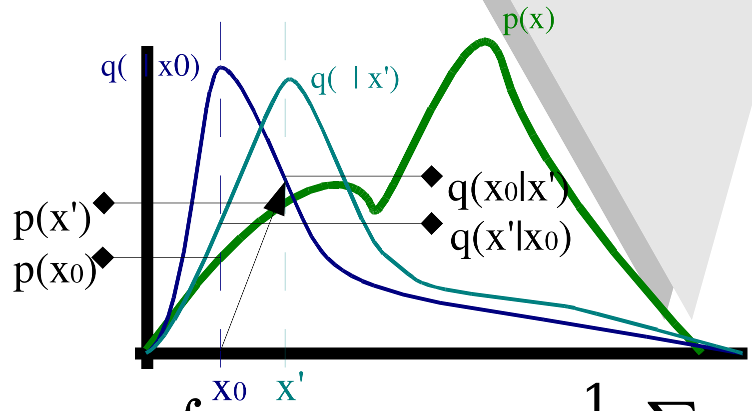
- Or, condition on all variables:

$$\begin{array}{lcl} \log a' & \sim & Nor(\log(a),1) \\ \log b' & \sim & Nor(\log(b),1) \\ \sigma(\pi') & \sim & Beta(\pi' \mid a,b) \prod_{m,n} Bin(z_{mn} \mid \pi') \\ z_{mn}' & \sim & Bin(z_{mn}' \mid \pi)\, p(w_{mn} \mid z_{mn}', \beta) \end{array}$$

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

Hal Daumé III (me@hal3.name)

# Metropolis-Hastings Sampling




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

- 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

      $$N \geq (\sigma_{max}/\sigma_{min})^2$$

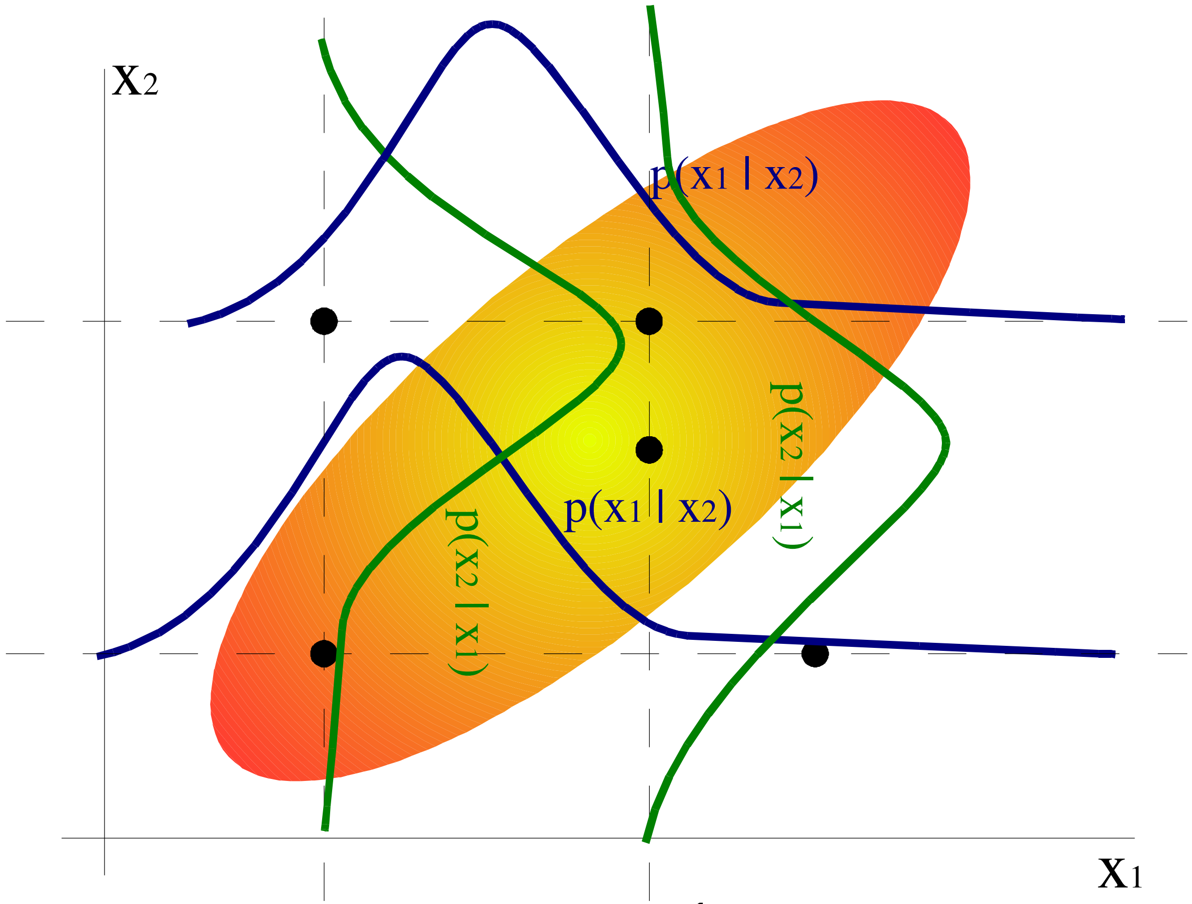
      where *sigma*s are lengths of the major density in *p*
  - Auto-correlation to track this:

$$r_k = \frac{\sum_{i=1}^{N-k}(x_i-\bar{x})(x_{i+k}-\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

Hal Daumé III (me@hal3.name)

# Gibbs Sampling

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




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

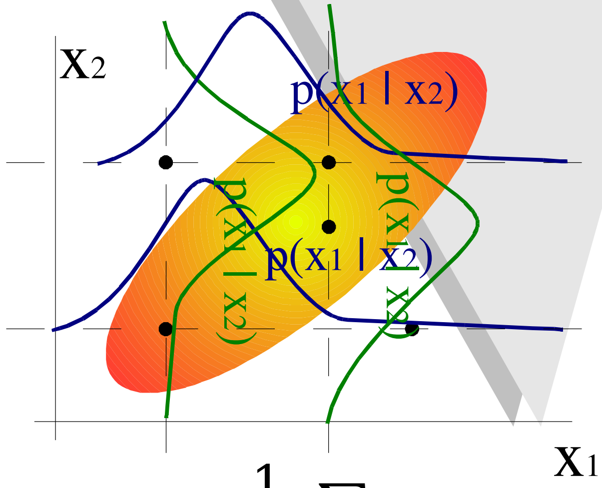
*See also: MK[30], Was[24.5], And03*

# Gibbs Sampling

- Typically our params are: $\bar{\theta} = \langle \theta_1, \ldots, \theta_D \rangle$
- If, for each $i$, we can draw a sample from:

$$p(\theta_i \mid \theta_{-i}) = p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_D)$$

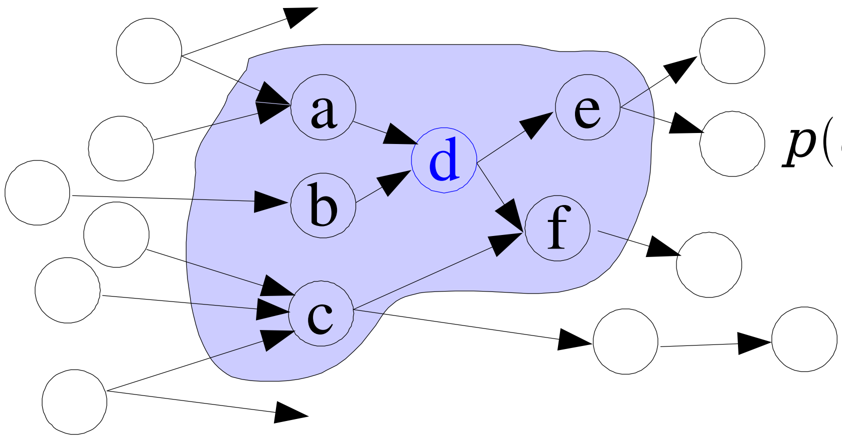
then we *can* use Gibbs sampling




$$F \approx \frac{1}{R} \sum_{x \in R} f(x)$$

- In graphical models, only depends on the *Markov blanket*:

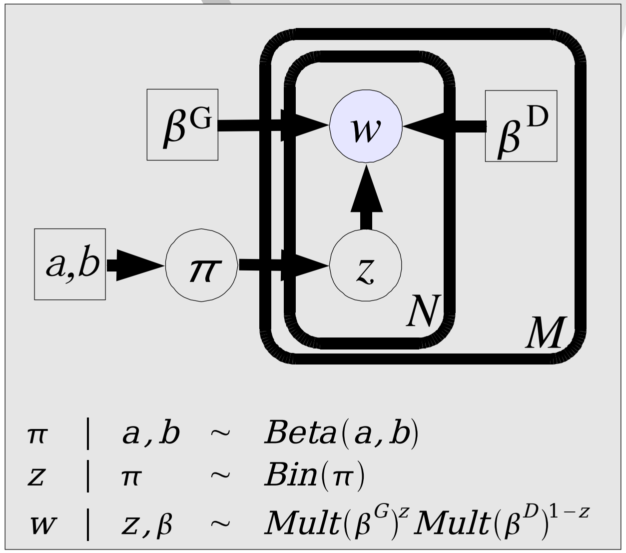
$$p(\theta_i \mid \theta_{-i}) = p(\theta_i \mid par(\theta_i)) \prod_{j : \theta_i \in par(\theta_j)} p(\theta_j \mid par(\theta_j))$$




$$p(d \mid \theta_{-d}) = p(d \mid a, b)\, p(e \mid d)\, p(f \mid d, c)$$

Hal Daumé III (me@hal3.name)

# Gibbs in our Model

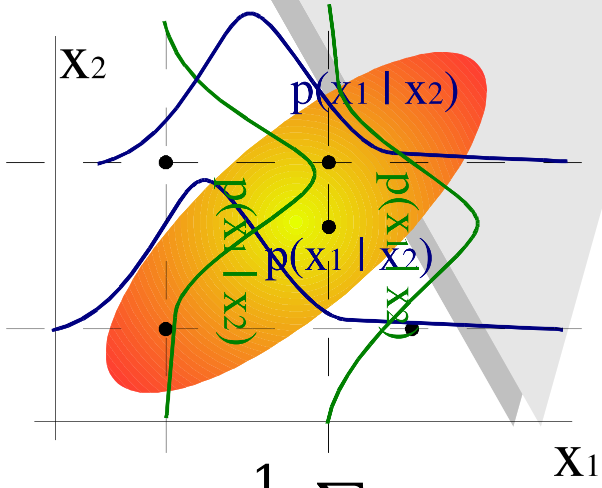



- Compute conditional probabilities

$$\begin{array}{lcll} a,b & \mid & \neg a,b & \sim & Beta(\pi \mid a,b) \\ \pi & \mid & \neg \pi & \sim & Beta(\pi \mid a,b) \prod_{m,n} Bin(z_{mn} \mid \pi) \\ z_{mn} & \mid & \neg z_{mn} & \sim & Bin(z_{mn} \mid \pi) p(w_{mn} \mid z_{mn}, \beta) \end{array}$$

- 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




$$F \approx \frac{1}{R} \sum_{x \in R} f(x)$$

- 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)

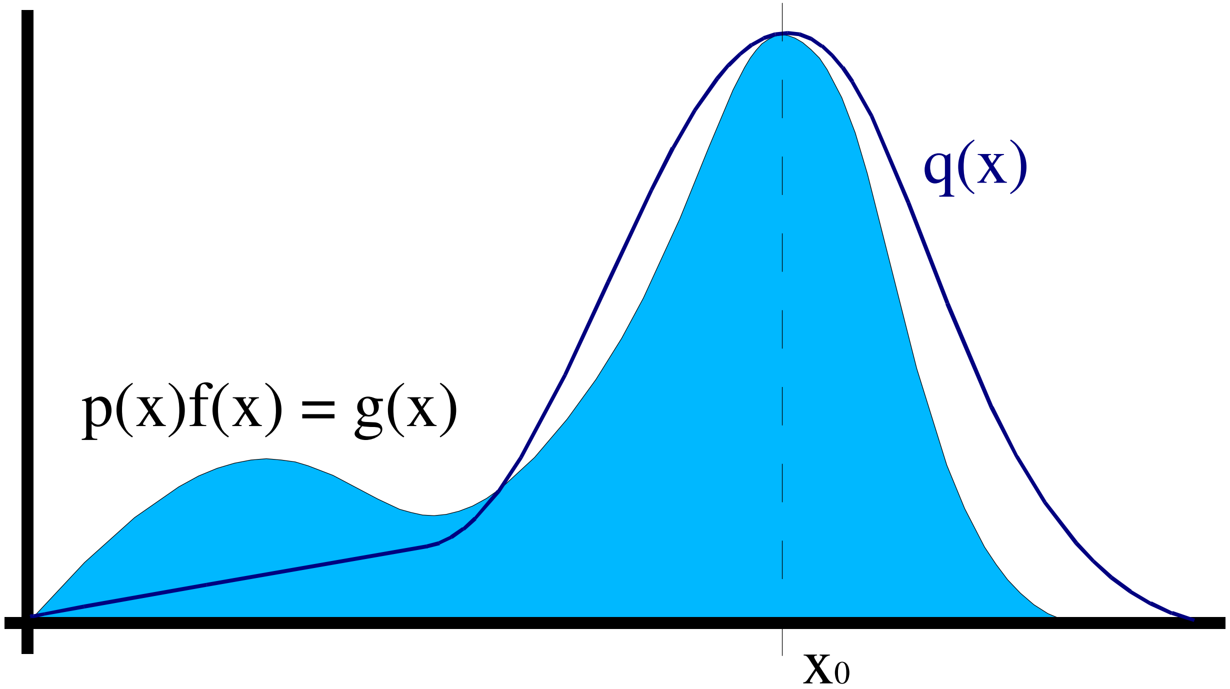
# Tutorial Outline

- 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

Hal Daumé III (me@hal3.name)

# Laplace (Saddlepoint) Approximation

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

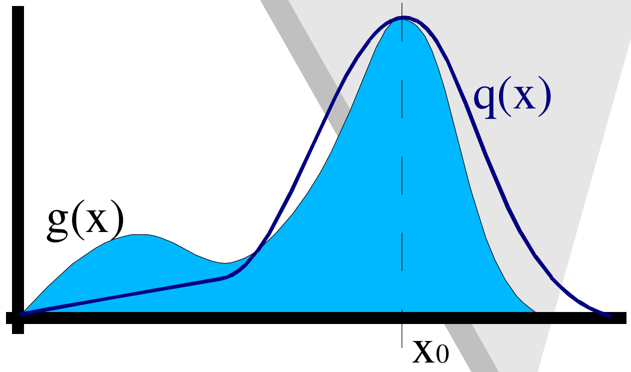



$$F = \int_X dx\, p(x) f(x) \approx g(x_0)\sqrt{\frac{2\pi}{c}} \quad , \quad c = -\left[\frac{\partial^2}{\partial x^2} \ln g(x)\right]_{x=x_0}$$

*See also: MK[27]*

Hal Daumé III (me@hal3.name)

# Laplace Approximation




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

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

- Find a mode x0 of the high-dimensional distribution *g*
- Approximate ln *g(x)* by a Taylor expansion around this mode:

$$\ln g(\bar{x}) \approx \ln g(\bar{x}_0) - \frac{1}{2}(\bar{x}-\bar{x}_0)^T \boldsymbol{A}(\bar{x}-\bar{x}_0)$$

- Compute the matrix **A** of second derivatives

$$A_{ij} = -\left[\frac{\partial^2}{\partial x_i \partial x_j} \ln g(\bar{x})\right]_{\bar{x}=\bar{x}_0}$$

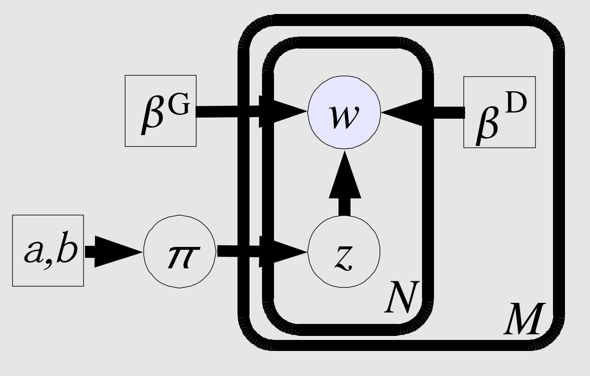
- The exponential form is a Gaussian distribution; use the Gaussian normalizing constant:

$$F = \int_{\mathbb{R}^D} dx\, g(x) \approx g(\bar{x}_0)\sqrt{\frac{(2\pi)^D}{\det \boldsymbol{A}}}$$

*See also: MK[27]*

Hal Daumé III (me@hal3.name)

# Laplace in our Model




$$\begin{array}{lcl} \pi \mid a,b & \sim & Beta(a,b) \\ z \mid \pi & \sim & Bin(\pi) \\ w \mid z,\beta & \sim & Mult(\beta^G)^z\, Mult(\beta^D)^{1-z} \end{array}$$

- Compute second derivatives:

$$\int_U d\pi\, \underbrace{Z_{ab}\pi^{a-1}(1-\pi)^{b-1}\prod_m\prod_n\sum_{z_{mn}}\pi^{z_{mn}}(1-\pi)^{1-z_{mn}}p(w_{mn}\mid z_{mn},\beta)}_{g(\pi)}$$

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

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

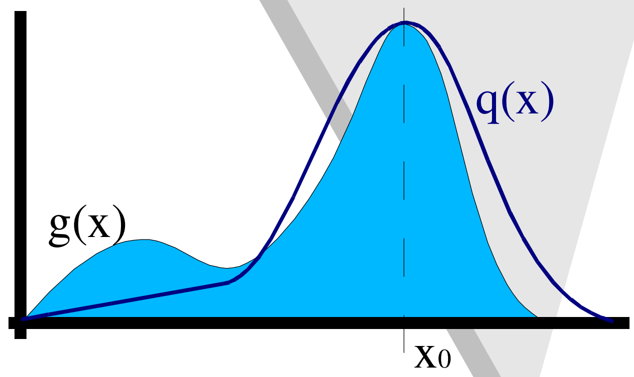
$$\frac{\pi_0}{(1-\pi_0)} = \frac{a-1+\sum_{m,n} z_{mn}}{b-1+\sum_{m,n}(1-z_{mn})}$$

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

Hal Daumé III (me@hal3.name)

# 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




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

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

- Cons:
  - Poor fit for multimodal distributions
  - Often, det **A** cannot be found efficiently

# 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 $Q = \{q(\cdot \mid \phi) : \phi \in \Phi\}$ 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:

$$\begin{aligned}
\log p(x \mid \theta) &= \log \int_Z dz\, p(x, z \mid \theta) \\
&= \log \int_Z dz\, q(z) \frac{p(X, z \mid \theta)}{q(z)} \\
&\geq \int_Z dz\, q(z) \log \frac{p(X, z \mid \theta)}{q(z)} \\
&= \int_Z q(z) \log p(x, z \mid \theta) - \int_Z q(z) \log q(z) \\
&= \underbrace{\boldsymbol{E}_{z \sim q}\{\log p(x, z \mid \theta)\} - \boldsymbol{E}_{z \sim q}\{\log q(z)\}}_{\mathcal{L}(x \mid \theta)}
\end{aligned}$$

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

$$\log p(x \mid \theta) = \mathcal{L} + KL\left(q(z) \,\|\, p(z \mid x, \theta)\right)$$

for *any* choice of $q$

Hal Daumé III (me@hal3.name)

# Variational EM

- Parameterize $q$ and directly optimize:

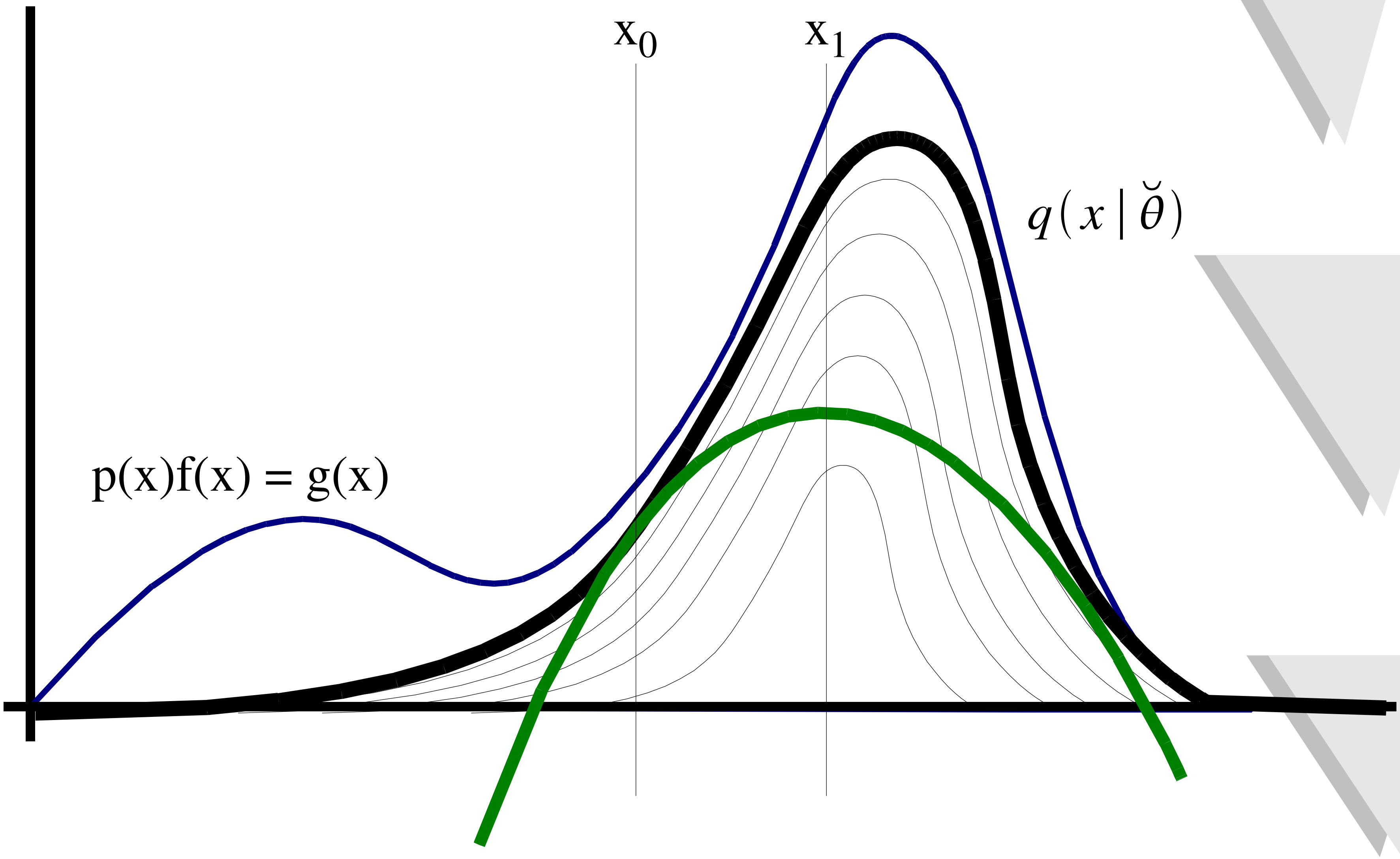
$$\log p(x \mid \theta) = \boldsymbol{E}_{z \sim q}\{\log p(x, z \mid \theta)\} - \boldsymbol{E}_{z \sim q}\{\log q(z)\} + KL\left(q(z \mid \breve{\theta}) \,\|\, p(z \mid x, \theta)\right)$$

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

- Art: inventing $q$ so that this is all tractable

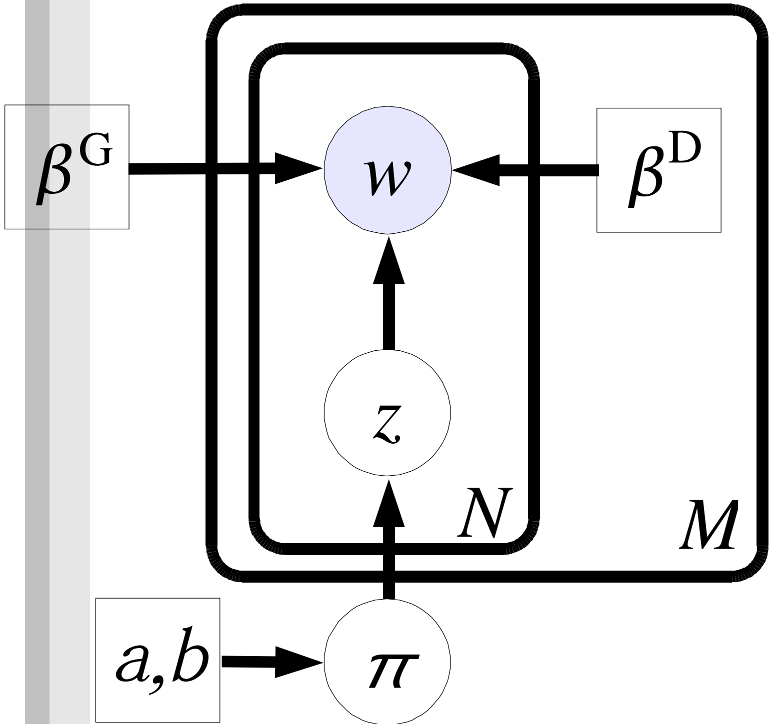
Hal Daumé III (me@hal3.name)

# Variational EM in Pictures

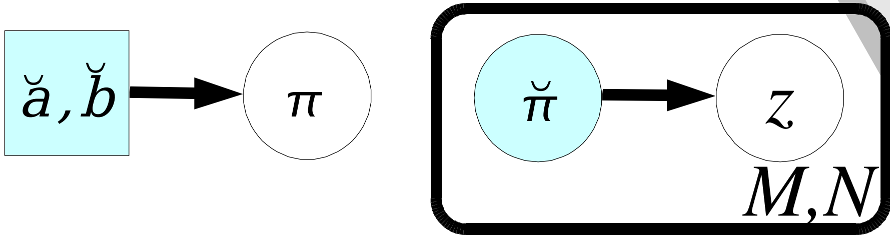


# Variational: Choosing Q

- Mixture model:




$$p(w,\pi,z \mid \rho,a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1}(1-\pi)^{b-1} \prod_{m,n} \pi^{z_{nm}}(1-\pi)^{1-z_{nm}} \prod_i \prod_v \left[\beta_v^i\right]^{z_{mni}}$$

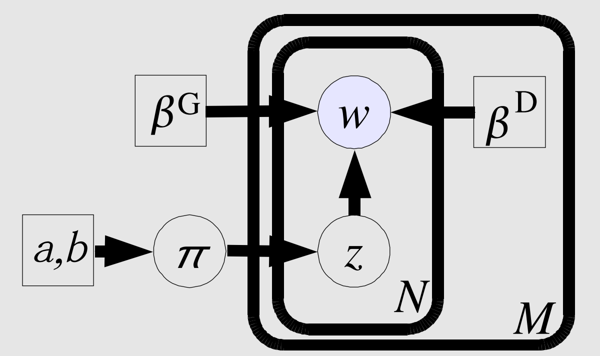



$$\begin{array}{lclcl} \pi & \mid & \breve{a},\breve{b} & \sim & Beta(\breve{a},\breve{b}) \\ z & \mid & \breve{\pi} & \sim & Bin(\breve{\pi}) \end{array}$$

$$q(\pi,z \mid \breve{a},\breve{b},\breve{\pi}) = \frac{\Gamma(\breve{a}+\breve{b})}{\Gamma(\breve{a})\Gamma(\breve{b})} \prod_{m,n} \prod_i \pi_{mni}^{\breve{a}_i-1} \breve{\pi}_{mn}^{z_{nmi}}$$

Key: $\pi$ and z are now not tied in the q distribution!

# VEM in our Model




$$\begin{array}{lllll} \pi & | & a,b & \sim & Beta(a,b) \\ z & | & \pi & \sim & Bin(\pi) \\ w & | & z,\beta & \sim & Mult(\beta^G)^z Mult(\beta^D)^{1-z} \end{array}$$

- Iterate:
  - Optimize variational parameters:

$$\breve{\pi}_{mni} \propto \exp\left[\Xi_i + \omega_{mni}\right]$$
$$\breve{a}_i = a_i + \sum_{m,n} \breve{\pi}_{mni}$$

$$\boxed{\Xi_i = \Psi(\breve{a}_i) - \Psi\left(\sum_i \breve{a}_i\right) \qquad \omega_{mni} = \sum_j w_{mnj} \log \beta_j^i}$$

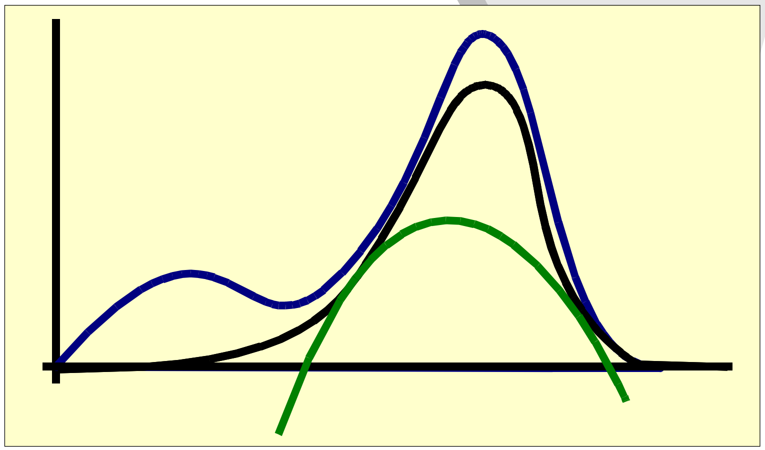
  - Optimize model parameters:

$$\beta_v^i \propto \sum_{m,n} \breve{\pi}_{mni} w_{mnv}$$

$a, b \sim$ generic optimization techniques

Hal Daumé III (me@hal3.name)

# Variational EM Summed Up



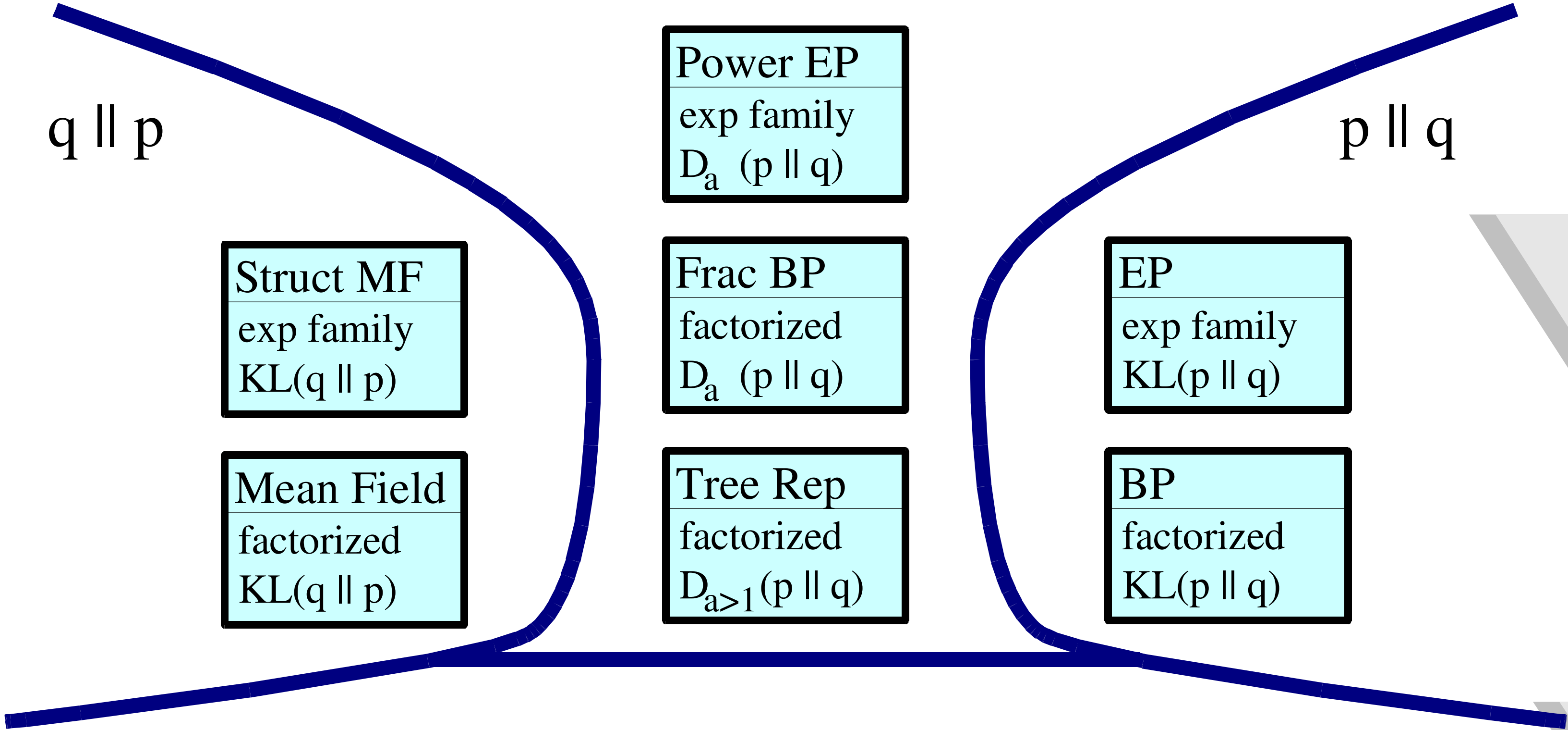
- Steps:
  - Write down conditional likelihood and choose an approximating distribution (eg, by factoring everything) with variational parameters
  - Iterate between optimizing the VPs and model parameters
- Pros:
  - Efficient, deterministic, often quite accurate
- Cons:
  - At it's heart, still a mode-based 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?




$$D_a(p||q) = \frac{1}{\beta(1-\beta)}\int dx\, \beta p + (1-\beta) q - p^{\beta} q^{\beta}$$

$$\beta = \frac{1}{2}(1+a)$$

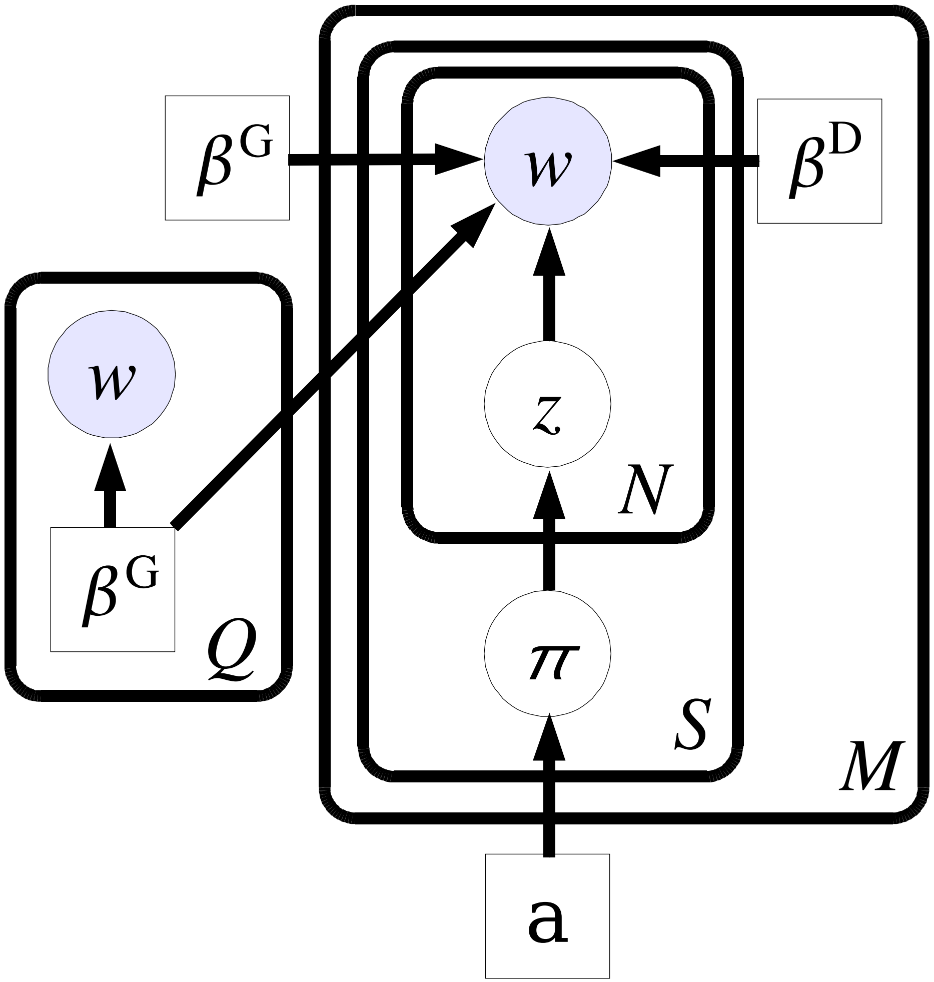
$$KL(p||q) = D_1(p||q) \qquad KL(q||p) = D_{-1}(p||q)$$

*See also: Min05*

Hal Daumé III (me@hal3.name)

# Empirical Evaluation of Methods

- Query-focused summarization model:




$$
\begin{aligned}
w^{Q}_{qn} &\sim Mult(\beta^{Q}_{q})\\
\pi_{ms} &\sim Dir(a)\\
z_{msn} &\sim Mult(\pi_{ms})\\
w_{msn} &\sim Mult(\beta^{G})^{z_{msn1}}\\
&\quad \prod_{m} Mult(\beta^{D}_{m})^{z_{msn}(m+1)}\\
&\quad \prod_{q} Mult(\beta^{Q}_{q})^{z_{msn}(q+M+1)}
\end{aligned}
$$

*See also: DM06*

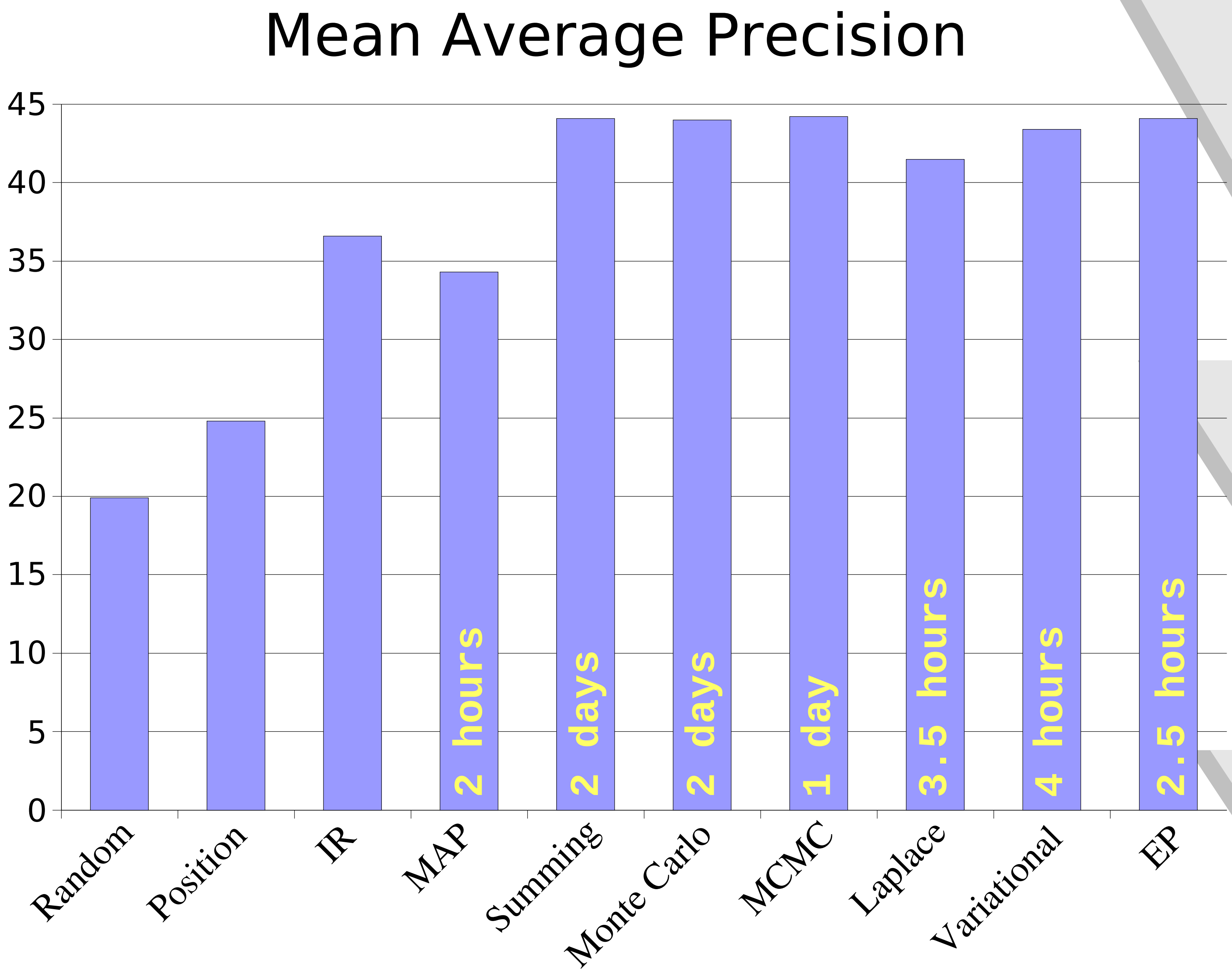
Hal Daumé III (me@hal3.name)

# 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)

*See also: DM06*

# Evaluation Results




See also: DM06

Hal Daumé III (me@hal3.name)

# 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

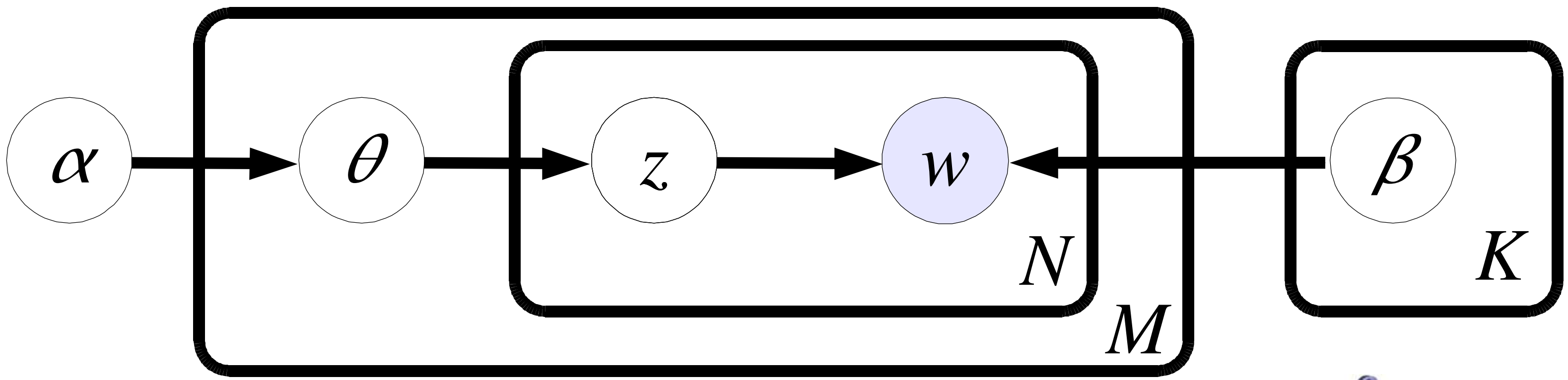
Hal Daumé III (me@hal3.name)

# Latent Dirichlet Allocation

[Blei, Ng + Jordan, JMLR 03]

- 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: $\theta \sim \mathtt{Dir}(\alpha)$
  - For each word (N total):
    - Choose a topic for this word: z $\sim \mathtt{Mult}(\theta)$
    - Choose the word itself: w $\sim \mathtt{Mult}(\beta^z)$

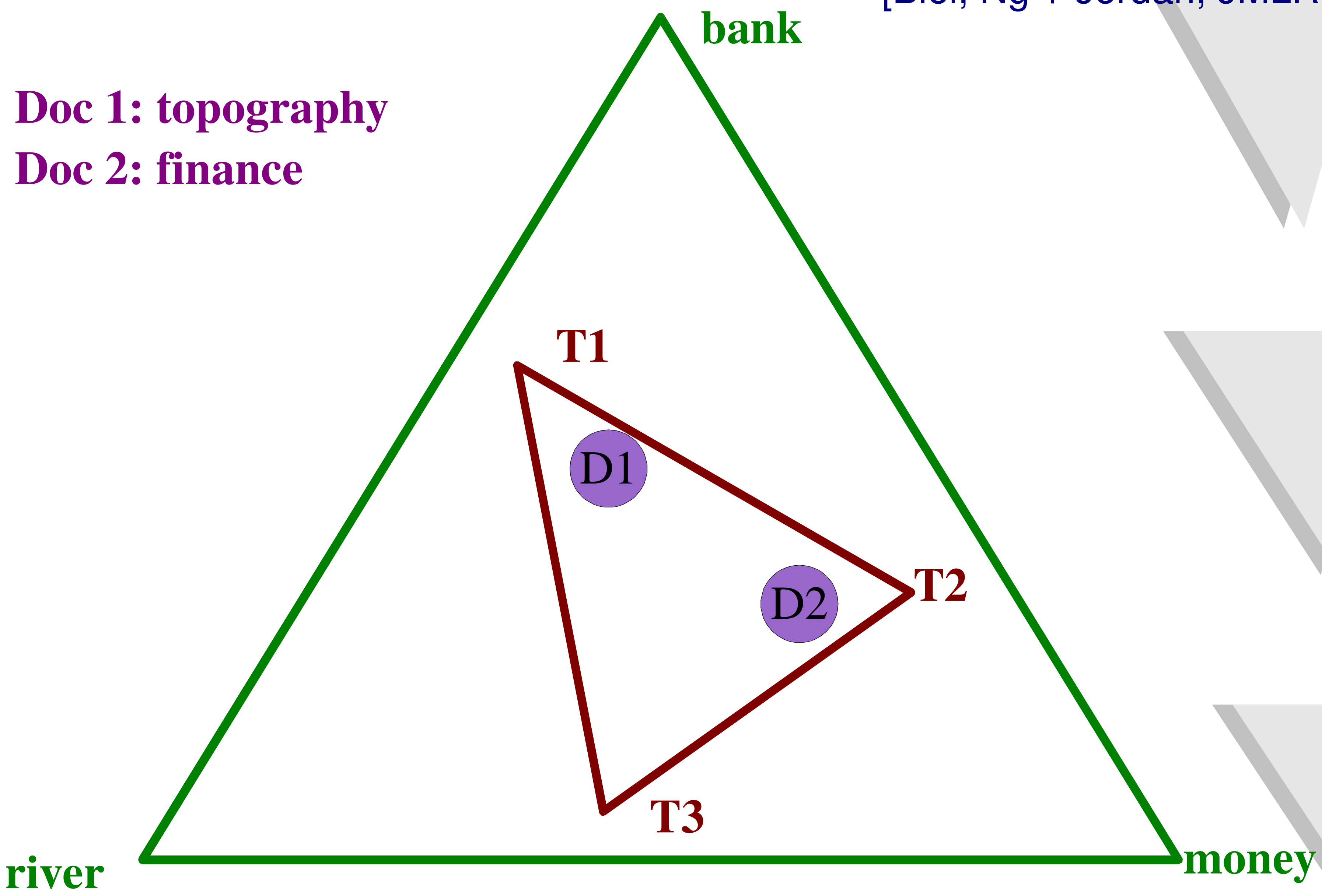


# LDA: Geometric Interpretation

[Blei, Ng + Jordan, JMLR 03]

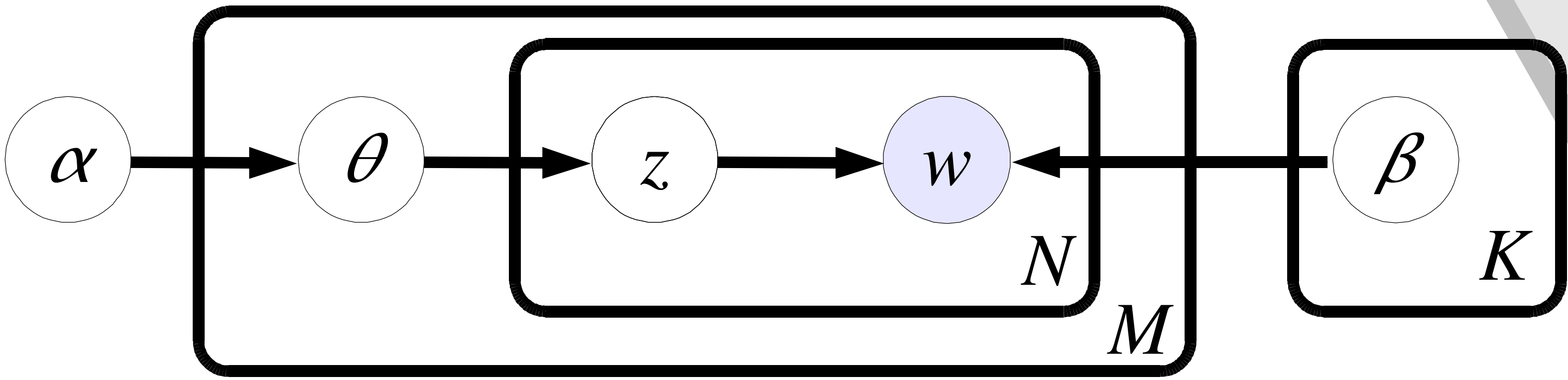
**Doc 1: topography**

**Doc 2: finance**



# LDA: Inference

[Blei, Ng + Jordan, JMLR 03]

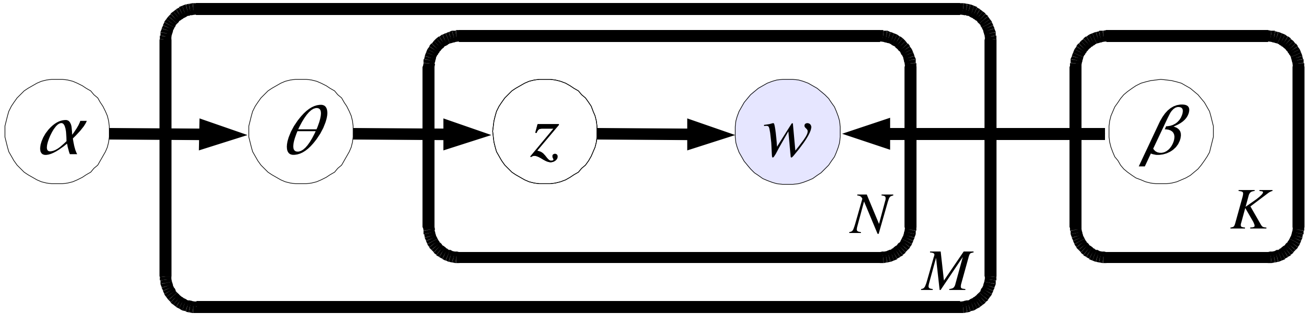



$$P(D) = \int_{\Delta_V} dP(\beta) \int_{\mathbb{R}^+} dP(\alpha) \prod_{m=1}^{M} \int_{\Delta_K} d\theta \left[ \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{j=1}^{K} \theta_k^{\alpha-1} \right] \prod_{n=1}^{N} \sum_{z_{mn}=1}^{K} \prod_{i=1}^{|V|} \prod_{j=1}^{K} \beta_{ji}^{\mathbf{1}[w_{mn}=i]\mathbf{1}[z_{mn}=j]}$$

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

# LDA: Naïve Gibbs Sampler

[Griffiths + Tenenbaum, CogSci 03]




$$\begin{aligned}
\alpha &\sim P(\alpha) \prod_m Dir(\theta_m \mid \alpha) \\
\beta_j &\sim P(\beta_j) \prod_{mn} Mult(w_{mn} \mid \beta_j)^{\mathbf{1}[z_{mn}=j]} \\
\theta_m &\sim Dir(\theta_m \mid \alpha) \prod_n Mult(z_{mn} \mid \theta_m) \\
z_{mn} &\sim Mult(z_{mn} \mid \theta_m)\, Mult(w_{mn} \mid \beta_{z_{mn}})
\end{aligned}$$

Can collapse this step!

# LDA Results

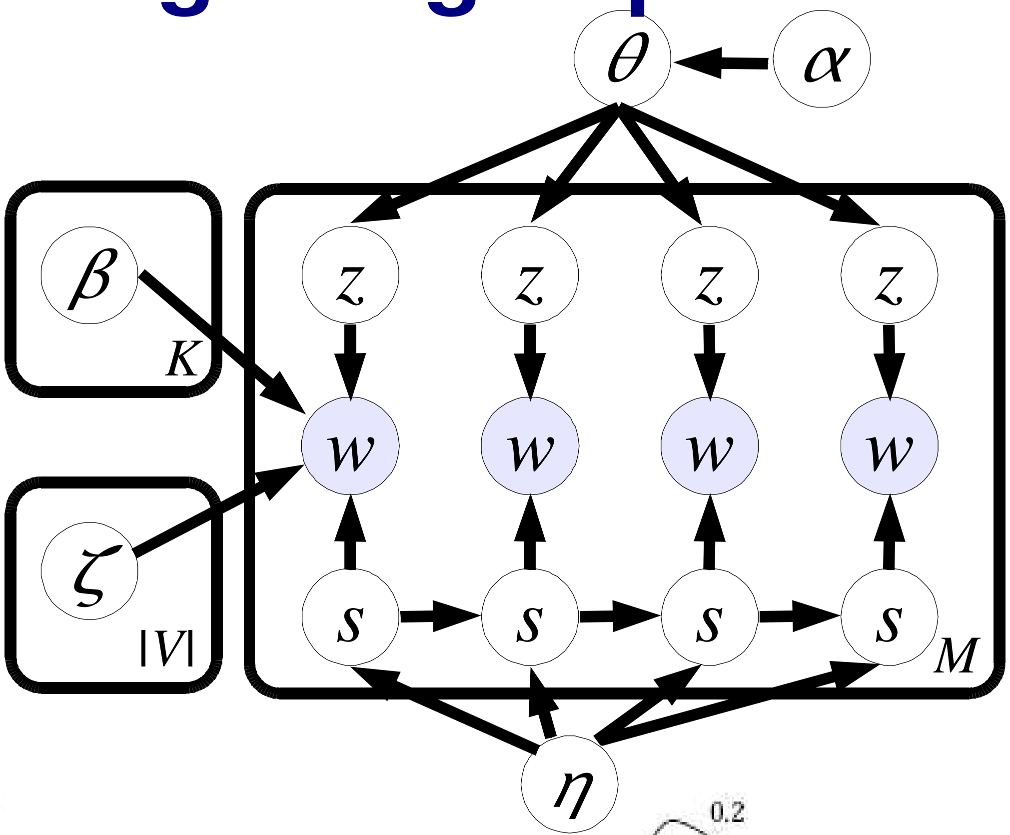
[Blei, Ng + Jordan, JMLR 03]

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

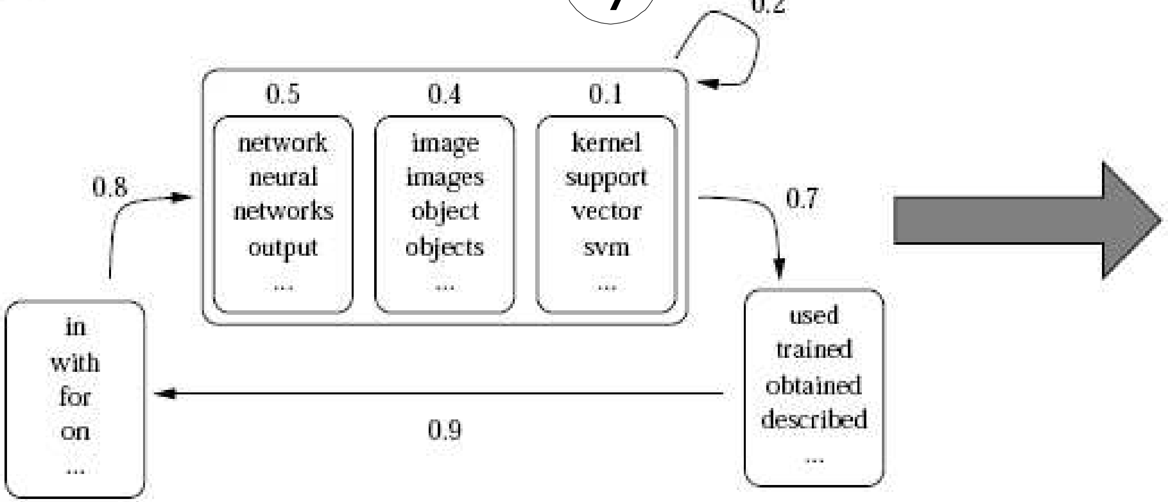
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our 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

[Griffiths, Steyvers, Blei + Tenenbaum, NIPS 2004]




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




network used for images
image obtained with kernel
output described with objects
neural network trained with svm images

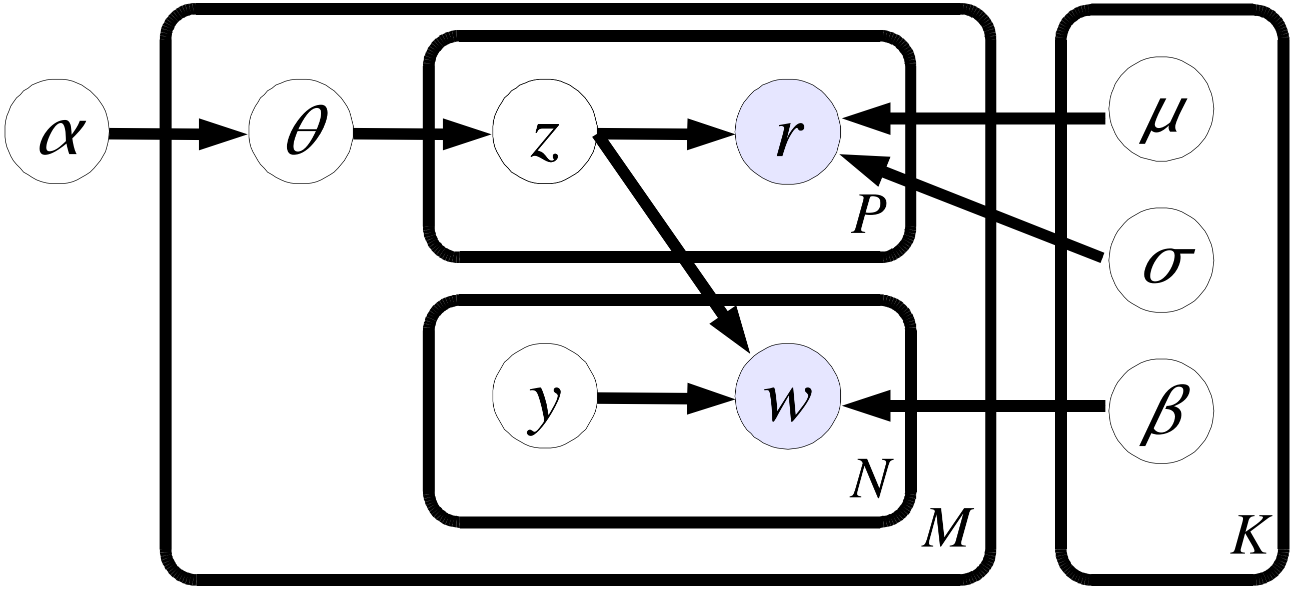
Hal Daumé III (me@hal3.name)

# LDA versus Topics+Syntax

[Griffiths, Steyvers, Blei + Tenenbaum, NIPS 2004]

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **LDA** | the | the | the | the | the | a | the | the | the |
| | blood | , | , | of | a | the | , | , | , |
| | , | and | and | , | of | of | of | a | a |
| | of | of | of | to | , | , | a | of | in |
| | body | a | in | in | in | in | and | and | game |
| | heart | in | land | and | to | water | in | drink | ball |
| | and | trees | to | classes | picture | is | story | alcohol | and |
| | in | tree | farmers | government | film | and | is | to | team |
| | to | with | for | a | image | matter | to | bottle | to |
| | is | on | farm | state | lens | are | as | in | play |
| **Topics** | blood | forest | farmers | government | light | water | story | drugs | ball |
| | heart | trees | land | state | eye | matter | stories | drug | game |
| | pressure | forests | crops | federal | lens | molecules | poem | alcohol | team |
| | body | land | farm | public | image | liquid | characters | people | * |
| | lungs | soil | food | local | mirror | particles | poetry | drinking | baseball |
| | oxygen | areas | people | act | eyes | gas | character | person | players |
| | vessels | park | farming | states | glass | solid | author | effects | football |
| | arteries | wildlife | wheat | national | object | substance | poems | marijuana | player |
| | * | area | farms | laws | objects | temperature | life | body | field |
| | breathing | rain | corn | department | lenses | changes | poet | use | basketball |
| **Syntax** | the | in | he | * | be | said | can | time | , |
| | a | for | it | new | have | made | would | way | ; |
| | his | to | you | other | see | used | will | years | ( |
| | this | on | they | first | make | came | could | day | : |
| | their | with | i | same | do | went | may | part | ) |
| | these | at | she | great | know | found | had | number | |
| | your | by | we | good | get | called | must | kind | |
| | her | from | there | small | go | | do | place | |
| | my | as | this | little | take | | have | | |
| | some | into | who | old | find | | did | | |

# Matching Words and Pictures




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

1. People, tree
2. Sky, jet
3. Sky, clouds
4. Sky, mountain
5. Plane, jet
6. Plane, jet

For each image/caption pair $M$

Draw a topic mixture $\theta \sim \text{Dir}(\alpha)$

For each image region $P$

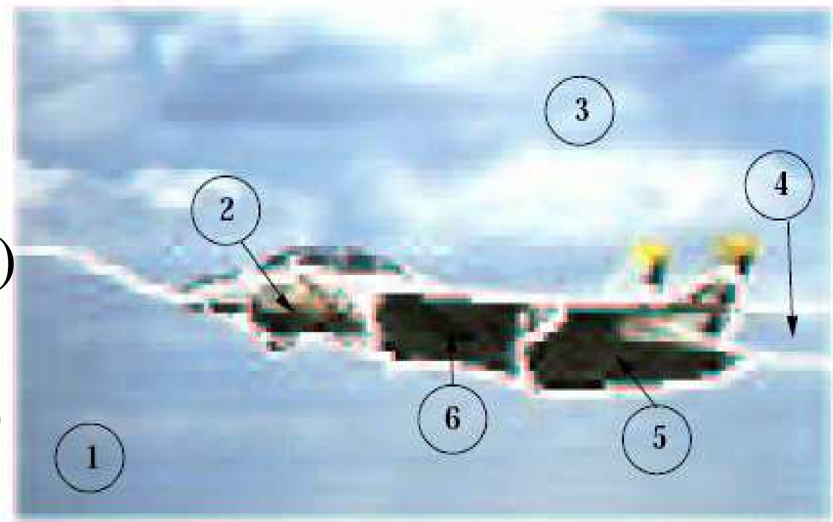
Draw a topic $z \sim \text{Mult}(\theta)$

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 $w \sim \text{Mult}(\beta_{zy})$



# Matching Words and Pictures

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

**True caption**
market people
**Corr–LDA**
people market pattern textile display

**True caption**
scotland water
**Corr–LDA**
scotland water flowers hills tree

**True caption**
sky tree water
**Corr–LDA**
tree water sky people buildings

**True caption**
birds tree
**Corr–LDA**
birds nest leaves branch tree

**True caption**
fish reefs water
**Corr–LDA**
fish water ocean tree coral

**True caption**
clouds jet plane
**Corr–LDA**
sky plane jet mountain clouds

# 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 ≈ SVMs)
  - Infinite models (Dirichlet Processes, Chinese Restaurant Processes)

Hal Daumé III (me@hal3.name)

# Bayes in Action (NLP/IR/Text)


Blei, Ng + Jordan, *Latent Dirichlet allocation*, JMLR03.

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

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

Griffiths, Steyvers, Blei, Tenenbaum, *Integrating topics and syntax.* NIPS04.

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)

James O. Berger, *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

David MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Larry Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2003.

Christopher Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

# For Further Information (Tutorials)

Andreiu, de Freitas, Doucet + Jordan, *An Introduction to MCMC for Machine Learning*. ML 2003

Wainwright + Jordan, *Graphical models, exponential families and variational inference*. UCB Stat TR#649, 2003.

Murphy, *A Brief Introduction to Graphical Models and Bayesian Networks*. **www.cs.ubc.ca/~murphyk/Bayes/bayes.html**

Minka, *Using lower bounds to approximate integrals*. 2003.

**www.research.microsoft.com/~minka/papers/rem.html**.

# Other References

Lawrence, *Fast sparse Gaussian process methods: the informative vector machine*. NIPS 2003.

Minka, *Expectation Propagation for Approximate Bayesian Inference.* UAI 2001.

Minka, *Divergence Measures and Message Passing*. AI-Stats 2005.

Neal, *Markov chain sampling methods for Dirichlet process mixture models*, TR. 9815, Dept. of Statistics, University of Toronto.

**http://bayes.hal3.name/**

**http://nlpers.blogspot.com**