CS598JHM: Advanced NLP (Spring 2013)
http://courses.engr.illinois.edu/cs598jhm/

Lecture 3: Comparing frequentist and Bayesian estimation techniques

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center
Office hours: by appointment

Text classification

The task: binary classification (e.g. sentiment analysis)

Assign (sentiment) label $L_i \in \{+,-\}$ to a document $\boldsymbol{W}_i = (w_{i1} \dots w_{iN})$.

$\boldsymbol{W}_1$= *"This is an amazing product: great battery life, amazing features and it's cheap."*

$\boldsymbol{W}_2$= *"How awful. It's buggy, saps power and is way too expensive."*

The data: A set $\boldsymbol{D}$ of N documents with (or without) labels

The model: Naive Bayes

We will use a frequentist model and a Bayesian model and compare supervised and unsupervised estimation techniques for them.

A Naive Bayes model

The task:

Assign (sentiment) label $L_i \in \{+,-\}$ to document $\mathbf{W}_i$.

$\mathbf{W}_1$= “This is an amazing product: great battery life, amazing features and it’s cheap.”
$\mathbf{W}_2$= “How awful. It’s buggy, saps power and is way too expensive.”

The model:

$$L_i = \text{argmax}_L\ P(L \mid \mathbf{W}_i) = \text{argmax}_L\ P(\mathbf{W}_i \mid L)P(L)$$

Assume $\mathbf{W}_i$ is a “bag of words”:

$\mathbf{W}_1 = \{$an:1, and: 1, amazing: 2, battery: 1, cheap: 1, features: 1, great: 1,…$\}$
$\mathbf{W}_2 = \{$awful: 1, and: 1, buggy: 1, expensive: 1,…$\}$

$P(\mathbf{W}_i \mid L)$ is a multinomial distribution: $\mathbf{W}_i \sim \text{Multinomial}(\boldsymbol{\theta}_L)$

With a vocabulary of V words, $\boldsymbol{\theta}_L = (\theta_1,\ldots, \theta_V)$

$P(L)$ is a Bernoulli distribution: $L \sim \text{Bernoulli}(\pi)$

The frequentist (maximum-likelihood) model

The frequentist model

The frequentist model has specific parameters $\boldsymbol{\theta}_L$ and π

$$L_i = \text{argmax}_L \, P(\mathbf{W}_i \mid \boldsymbol{\theta}_L) P(L \mid \pi)$$

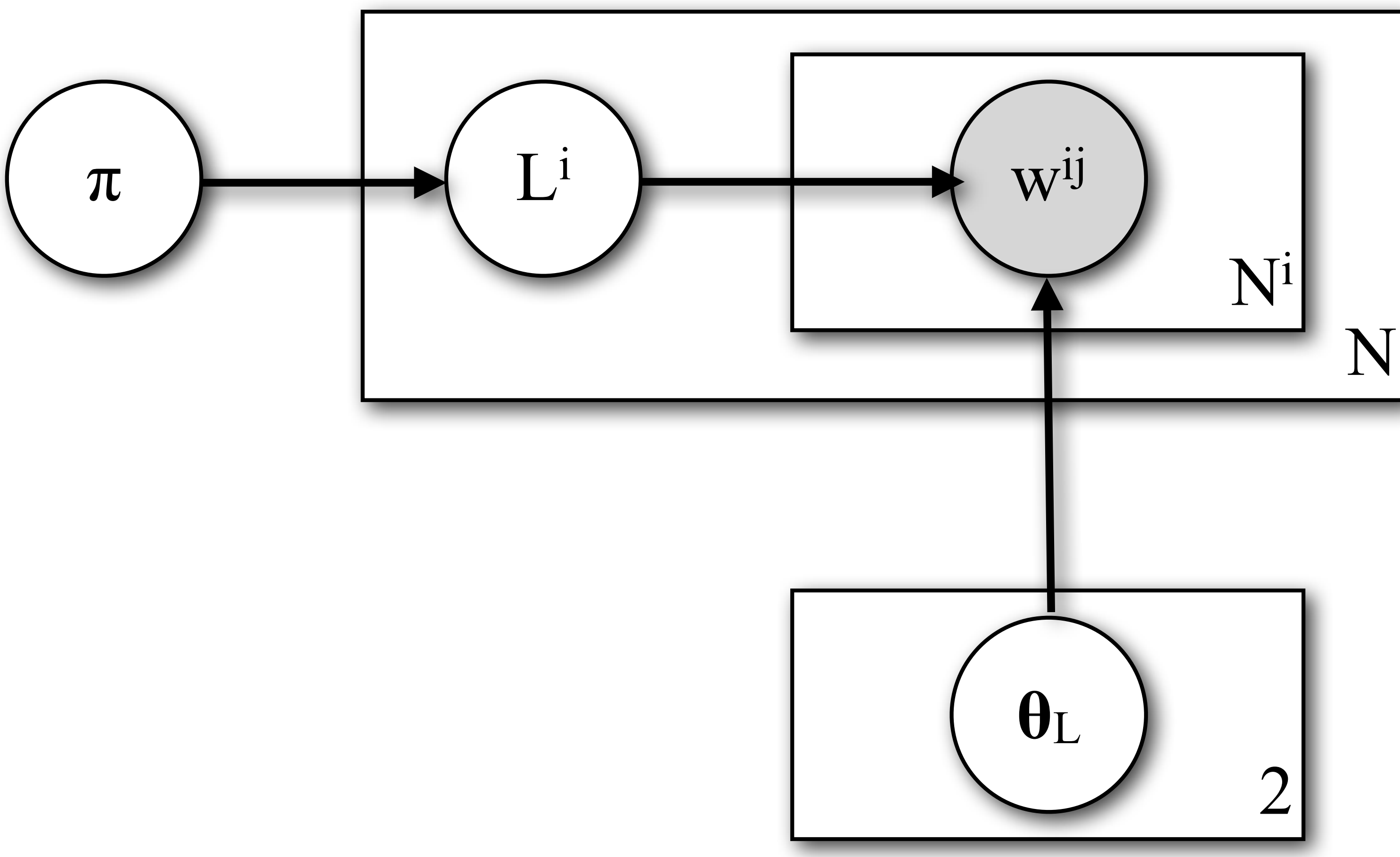
$P(\mathbf{W}_i \mid \boldsymbol{\theta}_L)$ is a multinomial over V words
with parameter $\boldsymbol{\theta}_L = (\theta_1, \ldots, \theta_V)$:

$$\mathbf{W}_i \sim \text{Multinomial}(\boldsymbol{\theta}_L)$$

$P(L \mid \pi)$ is a Bernoulli distribution with parameter π:

$$L \sim \text{Bernoulli}(\pi)$$

The frequentist model



Supervised MLE

The data is labeled:

We have a set $\mathbf{D}$ of D documents $\mathbf{W}_1 ... \mathbf{W}_d$ with N words

Each document $W_{\mathbf{i}}$ has N^i words

D^+ documents (subset $\mathbf{D}^+$) have a positive label and N^+ words

D^- documents (subset $\mathbf{D}^-$) have a negative label and N^- words

Each word w_i appears $N^+(w_i)$ times in $\mathbf{D}^+$, $N^-(w_i)$ times in $\mathbf{D}^-$

Each word w_i appears $N^j(w_i)$ times in D^j

MLE: relative frequency estimation

- Labels: $L \sim \text{Bernoulli}(\pi)$ with $\pi = D^+/d$
- Words: $\mathbf{W}_i \,|+ \sim \text{Multinomial}(\boldsymbol{\theta}^+)$ with $\boldsymbol{\theta}_{\mathbf{i}}^+ = N^+(w_i)/N^+$
- Words: $\mathbf{W}_i \,|- \sim \text{Multinomial}(\boldsymbol{\theta}^-)$ with $\boldsymbol{\theta}_{\mathbf{i}}^- = N^-(w_i)/N^-$

Inference with MLE

The inference task:
Given a new document $\mathbf{W}_{i+1}$, what is its label L_{i+1}?

Recall: the word w_j occurs $N_{i+1}(w_j)$ times in $\mathbf{W}_{i+1}$.

$$\begin{aligned} P(L = +|\mathbf{W}_{i+1}) &\propto P(+)P(\mathbf{W}_{i+1}|+) \\ &= \pi \prod_{j=1}^{V} \theta_{+j}^{N_{i+1}(w_j)} \end{aligned}$$

Unsupervised MLE

The data is unlabeled:

We have a set $\mathbf{D}$ of D documents $\mathbf{W}_1 \ldots \mathbf{W}_d$ with N words

Each document $\mathbf{W}_\mathbf{i}$ has N^i words

Each word $w_1 \ldots w_i \ldots w_V$ appears $N^j(w_i)$ times in $\mathbf{W}_j$

EM algorithm: "expected relative frequency estimation"

Initialization: pick initial $\pi^{(0)}$, $\boldsymbol{\theta}^{+(0)}$, $\boldsymbol{\theta}^{-(0)}$

Iterate:

- Labels: $L \sim \text{Bernoulli}(\pi)$ with $\pi^{(t)} = \langle N_+ \rangle_{(t-1)} / \langle N \rangle_{(t-1)}$
- Words: $\mathbf{W}_i \mid + \sim \text{Multinomial}(\boldsymbol{\theta}^+)$ with $\theta_\mathbf{i}^{+\,(t)} = \langle N^+(w_i) \rangle_{(t-1)} / \langle W^+ \rangle_{(t-1)}$
- Words: $\mathbf{W}_i \mid - \sim \text{Multinomial}(\boldsymbol{\theta}^-)$ with $\theta_\mathbf{i}^{-\,(t)} = \langle N^-(w_i) \rangle_{(i-1)} / \langle W^- \rangle_{(i-1)}$

Maximum Likelihood estimation

With **complete** (= labeled) **data** $\mathbf{D} = \{ \langle \mathbf{X}_i , \mathbf{Z}_i \rangle \}$,
maximize the complete likelihood $p(\mathbf{X}, \mathbf{Z} \mid \theta)$:

$$\theta^* = \text{argmax}_{\theta} \prod_i p(\mathbf{X}_i , \mathbf{Z}_i \mid \theta)$$

or $$\theta^* = \text{argmax}_{\theta} \sum_i \ln(p(\mathbf{X}_i , \mathbf{Z}_i \mid \theta))$$

Maximum Likelihood estimation

With **incomplete** (= unlabeled) **data**, $\mathbf{D} = \{ \langle \mathbf{X}_i , ? \rangle \}$
maximize the incomplete (marginal) likelihood $p(\mathbf{X} \mid \theta)$:

$$\theta^* = \mathrm{argmax}_{\theta} \sum_i \ln(p(\mathbf{X}_i \mid \theta))$$
$$= \mathrm{argmax}_{\theta} \sum_i \ln\left(\sum_{\mathbf{Z}} p(\mathbf{X}_i , \mathbf{Z} \mid \theta)\, p(\mathbf{Z} \mid \mathbf{X}_i,\theta') \right)$$
$$= \mathrm{argmax}_{\theta} \sum_i \ln\left(\mathbf{E}_{\mathbf{Z}|\mathbf{X}_i,\theta'}[\, p(\mathbf{X}_i , \mathbf{Z} \mid \theta)] \right)$$

$p(\mathbf{Z} \mid \mathbf{X}, \theta)$: the posterior probability of $\mathbf{Z}$ ($\mathbf{X}$ = our data)
$\mathbf{E}_{\mathbf{Z}|\mathbf{X}_i,\theta}[\, p(\mathbf{X}_i, \mathbf{Z} \mid \theta)]$: the expectation of $p(\mathbf{X}, \mathbf{Z} \mid \theta)$ wrt. $p(\mathbf{Z} \mid \mathbf{X}, \theta)$

Find parameters θ^{new} that maximize the expected log-likelihood of the joint $p(\mathbf{Z},\mathbf{X} \mid \theta^{new})$ under $p(\mathbf{Z} \mid \mathbf{X}, \theta^{old})$
This requires an iterative approach

The EM algorithm

1. **Initialization:** Choose initial parameters θ^{old}

2. **Expectation step:** Compute $p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}})$
   (= posterior of the latent variables Z)

3. **Maximization step:** Compute θ^{new}
   θ^{new} maximizes the expected log-likelihood of the joint $p(\mathbf{Z},\mathbf{X} \mid \theta^{\text{new}})$ under $p(\mathbf{Z} \mid \mathbf{X}, \theta^{\text{old}})$:

$$\theta^{new} = \arg\max_{\theta} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

4. **Check for convergence.**
   Stop, or set $\theta^{old} := \theta^{new}$ and go to 2.

The EM algorithm

The classes we find may not correspond to the classes we would be interested in.

- Seed knowledge (e.g. a few positive and negative words) may help

We are not guaranteed to find a global optimum, and may get stuck in a local optimum.

- Initialization matters

In our example...

Initialization: Pick (random) $\pi_A,\ \pi_B = (1\text{-}\pi_A),\ \boldsymbol{\theta}_A\ ,\ \boldsymbol{\theta}_B$

E-step:

Set $N_A, N_B, N_A(w_1), \ldots, N_A(w_V), N_B(w_1), \ldots N_B(w_V) := 0$

For each document $\mathbf{W}_i$,

Set $\mathbf{L}_i = A$ with $P(\mathbf{L}_i = A \mid \mathbf{W}_i, \pi_A, \pi_B, \boldsymbol{\theta}_A, \boldsymbol{\theta}_B) \propto \pi_A \prod_j P(w_{ij} \mid \boldsymbol{\theta}_A)$

Set $\mathbf{L}_i = B$ with $P(\mathbf{L}_i = B \mid \mathbf{W}_i, \pi_A, \pi_B, \boldsymbol{\theta}_A, \boldsymbol{\theta}_B) \propto \pi_b \prod_j P(w_{ij} \mid \boldsymbol{\theta}_B)$

Update $N_A \mathrel{+}= P(\mathbf{L}_i = A \mid \mathbf{W}_i, \pi_A, \pi_B, \boldsymbol{\theta}_A, \boldsymbol{\theta}_B)$

$N_B \mathrel{+}= P(\mathbf{L}_i = B \mid \mathbf{W}_i, \pi_A, \pi_B, \boldsymbol{\theta}_A, \boldsymbol{\theta}_B)$

For all words w_{ij} in $\mathbf{W}_i$:

$N_A(w_{ij}) \mathrel{+}= P(\mathbf{L}_i = A \mid \mathbf{W}_i, \pi_A, \pi_B, \boldsymbol{\theta}_A, \boldsymbol{\theta}_B)$

$N_B(w_{ij}) \mathrel{+}= P(\mathbf{L}_i = B \mid \mathbf{W}_i, \pi_A, \pi_B, \boldsymbol{\theta}_A, \boldsymbol{\theta}_B)$

M-step:

$\pi_A := N_A/(N_A + N_B)$ $\quad\quad \pi_B := N_B/(N_A + N_B)$

$\theta_A(w_i) := N_A(w_i) / \sum_j (N_A(w_j))$ $\quad\quad \theta_B(w_i) := N_B(w_i) / \sum_j (N_B(w_j))$

The Bayesian model

The Bayesian model

The Bayesian model has priors $\mathrm{Dir}(\boldsymbol{\gamma})$ and $\mathrm{Beta}(\alpha,\beta)$ with hyperparameters $\boldsymbol{\gamma}=(\gamma_1, \ldots, \gamma_V)$ and α, β

It does not have specific $\boldsymbol{\theta}_L$ and π, but integrates them out:

$$L_i = \mathrm{argmax}_L \iint P(\mathbf{W}_i \mid \boldsymbol{\theta}_L)P(\boldsymbol{\theta}_L; \boldsymbol{\gamma}_L, \mathbf{D})\, P(L \mid \pi)P(\pi; \alpha,\beta,\mathbf{D})\, d\boldsymbol{\theta}_L d\pi$$

$$= \mathrm{argmax}_L \int P(\mathbf{W}_i \mid \boldsymbol{\theta}_L)P(\boldsymbol{\theta}_L; \boldsymbol{\gamma}_L, \mathbf{D})d\boldsymbol{\theta}_L \int P(L \mid \pi)P(\pi; \alpha,\beta,\mathbf{D})d\pi$$

$$= \mathrm{argmax}_L\, P(\mathbf{W}_i \mid \boldsymbol{\gamma}_L, \mathbf{D})\, P(L \mid \alpha,\beta,\mathbf{D})$$

$P(\mathbf{W}_i \mid \boldsymbol{\theta}_L)$ is a multinomial with parameter $\boldsymbol{\theta}_L = (\theta_1,\ldots, \theta_V)$,

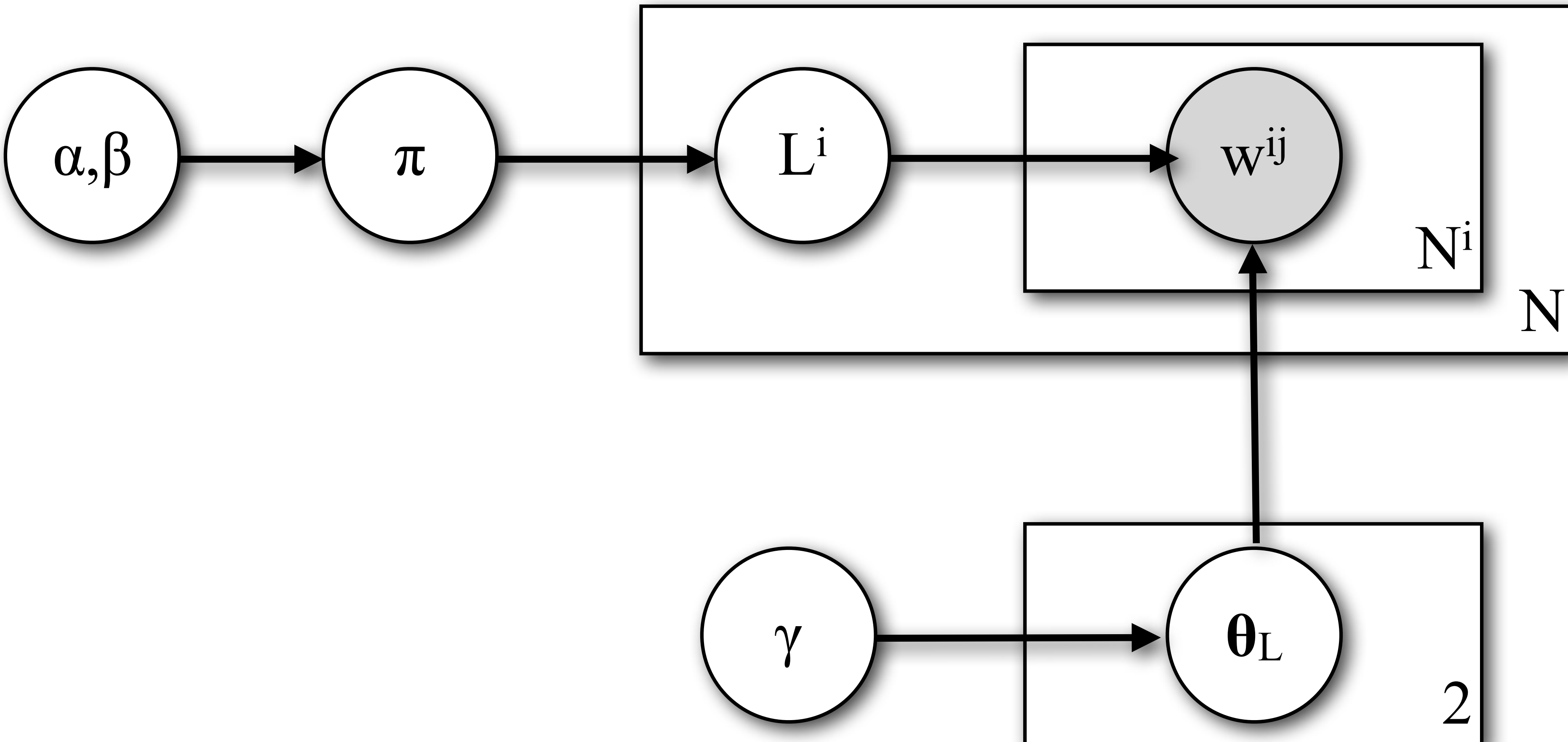
$P(\boldsymbol{\theta}_L; \boldsymbol{\gamma}_L)$ is a Dirichlet with hyperparameter $\boldsymbol{\gamma}_L = (\gamma_1,\ldots, \gamma_V)$

$$\boldsymbol{\theta}_L \sim \mathrm{Dirichlet}(\boldsymbol{\gamma}_L) \qquad \mathbf{W}_i \sim \mathrm{Multinomial}(\boldsymbol{\theta}_L)$$

$P(L \mid \pi)$ is a Bernoulli with parameter π, drawn from a Beta prior

$$\pi \sim \mathrm{Beta}(\alpha, \beta) \qquad L \sim \mathrm{Bernoulli}(\pi)$$

The Bayesian model



Bayesian: supervised

The data is labeled:

- We have a set $\mathbf{D}$ of D documents $\mathbf{W}_1 \ldots \mathbf{W}_D$ with N words
- Each document $W_\mathbf{i}$ has N^i words
- D^+ documents (subset $\mathbf{D}^+$) have a positive label and N^+ words
- D^- documents (subset $\mathbf{D}^-$) have a negative label and N^- words
- Each word w_i appears $N^+(w_i)$ times in $\mathbf{D}^+$, $N^-(w_i)$ times in $\mathbf{D}^-$
- Each word w_j appears $N^i(w_j)$ times in $\mathbf{W}_i$

Bayesian estimation

$$P(L = + \mid \mathbf{D}) = (D^+ + \alpha)/(D + \alpha + \beta)$$

$$P(w_i \mid +, \mathbf{D}) = (N^+(w_i) + \gamma_i)/(N^+(w_i) + \gamma_0)$$

$$P(\mathbf{W}_i \mid +, \mathbf{D}) = \prod_j P(w_j \mid +)^{Ni(wj)}$$

$$P(L_i = + \mid \mathbf{W}_i, \mathbf{D}) = [(D^+ + \alpha)/(D + \alpha + \beta)]\prod_j P(w_j \mid +)^{Ni(wj)}$$

Bayesian: unsupervised

We need to approximate an integral/expectation:

$$p(L_i = + \mid \mathbf{W}_i)$$
$$\propto \iint p(\mathbf{W}_i \mid +, \boldsymbol{\theta}_+)\, p(\boldsymbol{\theta}_+; \boldsymbol{\gamma}, \mathbf{D})\, p(L = + \mid \pi)\, p(\pi; \alpha, \beta, \mathbf{D})\, d\boldsymbol{\theta}_+\, d\pi$$
$$\propto \int p(\mathbf{W}_i \mid +, \boldsymbol{\theta}_+)\, p(\boldsymbol{\theta}_+; \boldsymbol{\gamma}, \mathbf{D})\, d\boldsymbol{\theta}_+ \int p(L = + \mid \pi)\, p(\pi; \alpha, \beta, \mathbf{D})\, d\pi$$
$$\propto p(\mathbf{W}_i \mid \boldsymbol{\gamma}, +, \mathbf{D})\; p(L_i = + \mid \alpha, \beta, \mathbf{D})$$

Approximating expectations

$$E[f(x)] = \int_0^1 f(x)p(x)dx$$

We can approximate the expectation of $f(x)$, $\langle f(x) \rangle = \int f(x)p(x)dx$, by sampling a finite number of points $x^{(1)}, \ldots, x^{(T)}$ according to $p(x)$, evaluating $f(x^{(i)})$ for each of them, and computing the average.

Markov Chain Monte Carlo

A multivariate distribution $p(\mathbf{x}) = p(x_1,\ldots,x_k)$ with discrete x_i has only a finite number of possible outcomes.

Markov Chain Monte Carlo methods construct a Markov chain whose states are the outcomes of $p(\mathbf{x})$.

The probability of visiting state $\mathbf{x_j}$ is $p(\mathbf{x_j})$

We sample from $p(\mathbf{x})$ by visiting a sequence of states from this Markov chain.

Gibbs sampling

Our states:

One label assignment $L_1, \ldots, L_N$ to each of our N documents

$\boldsymbol{x} = (L_1, \ldots, L_N)$

Our transitions:

We go from one label assignment $\mathbf{x} = (+,+,-,+,-\ldots+)$

to another $\mathbf{y} = (-,+,+,+,\ldots,+)$

Our intermediate steps:

We generate label Y_i conditioned on $Y_1 \ldots Y_{i-1}$ and $X_{i+1} \ldots X_N$

Call label assignment $Y_1 \ldots Y_{i-1}$, $X_{i+1} \ldots X_N$ $\boldsymbol{L}^{(-i)}$

We need to compute $\mathrm{P}(Y_i \mid \boldsymbol{D}, \boldsymbol{L}^{(-i)}, \alpha, \beta, \gamma)$

Gibbs sampling

We visit states according to transition probabilities $P(\mathbf{y}|\ \mathbf{x})$

We go from state $\mathbf{x} = (x_1,\dots,x_k)$ to state $\mathbf{y} = (y_1,\dots,y_k)$

We get from $\mathbf{x} = (x_1,\dots,x_k)$ to $\mathbf{y} = (y_1,\dots,y_k)$ in k steps:

$(x_1, x_2,\dots, x_i, \dots, x_{k-1}, x_k) = \mathbf{x} = \mathbf{x}^{(t)}$
$(y_1, x_2,\dots, x_i, \dots, x_{k-1}, x_k)$
$(y_1, y_2,\dots, x_i, \dots, x_{k-1}, x_k)$
$(y_1, y_2,\dots, x_i, \dots, x_{k-1}, x_k)$
$(y_1, y_2,\dots, y_i, \dots, x_{k-1}, x_k)$
$(y_1, y_2,\dots, y_i, \dots, x_{k-1}, x_k)$
$(y_1, y_2,\dots, y_i, \dots, y_{k-1}, x_k)$
$(y_1, y_2,\dots, y_i, \dots, y_{k-1}, y_k) = \mathbf{y} = \mathbf{x}^{(t+1)}$

Gibbs sampling

We will visit a sequence of states according to the transition probabilities $P(\mathbf{y} \mid \mathbf{x})$

That is, we will go from state $\mathbf{x} = (x_1, \ldots, x_k)$ to state $\mathbf{y} = (y_1,\ldots,y_k)$ with probability $P(\mathbf{y} \mid \mathbf{x})$

For $i = 1...k$:
pick a value for y_i by sampling from $P(Y_i \mid y_1,\ldots, y_{i-1}, x_{i+1},\ldots, x_k)$

$$P(Y_i = y_i \mid y_1,\ldots,y_{i-1}, x_{i+1},\ldots,x_k) = P(y_1,\ldots,y_{i-1}, y_i, x_{i+1},\ldots,x_k)/(y_1,\ldots,y_{i-1}, x_{i+1},\ldots, x_k)$$

Gibbs sampling

For us $p(\mathbf{x}) = p(\mathbf{D}, \mathbf{L}, \pi, \theta+, \theta-; \alpha, \beta, \boldsymbol{\gamma})$

$\pi, \theta+, \theta-$ are real-valued, but they disappear because we integrate them out:

$$P(L_j = + \mid \mathbf{L}^{(-\mathbf{j})}; \alpha, \beta) \quad = \quad \frac{\alpha + N_+^{(-j)}}{\alpha + \beta + N - 1}$$

$$P(w_k = y | D_+^{(-j)}; \boldsymbol{\gamma}) \quad = \quad \frac{N_{D_x^{(-j)}}(y) + \gamma_y}{\gamma_0 + N_{D_x^{(-j)}}}$$

Gibbs sampling

$$\underbrace{P(L_j = +|\mathbf{D}, \mathbf{L}^{(-\mathbf{j})}; \alpha, \beta, \boldsymbol{\gamma})}_{\text{prob. that } \mathrm{D_j} \text{ is pos. review}}$$

$$\propto \underbrace{P(\mathbf{W_j}|+, D_+^{(-j)}; \boldsymbol{\gamma})}_{\text{pos. review generates } \mathrm{D_j}} \underbrace{P(L_j = +|\mathbf{L}^{(-\mathbf{j})}; \alpha, \beta)}_{\text{prob. of pos. review}}$$

$$P(L_j = + \mid \mathbf{L}^{(-\mathbf{j})}; \alpha, \beta) = \frac{\alpha + N_+^{(-j)}}{\alpha + \beta + N - 1}$$

$$P(w_k = y|D_+^{(-j)}; \boldsymbol{\gamma}) = \frac{N_{D_x^{(-j)}}(y) + \gamma_y}{\gamma_0 + N_{D_x^{(-j)}}}$$

The Gibbs sampler

Initialize:

Define priors α, β, γ.

Assign initial labels $\mathbf{L}^{(0)}$ to documents

Iterate:

For each iteration $t = 1 \ldots T$:

For every document $\mathbf{W}_i$ (with current label $x = L_i^{(t-1)}$)

(Temporarily) remove its word counts $\mathbf{N}_i(w_j)$ from its class x:

$N_{x\backslash i}^{(t-1)}(w_j) = N_x^{(t-1)}(w_j) - N_i^{(t-1)}(w_j)$

(Temporarily) remove $\mathbf{W}_i$ from the documents in its class x:

$D_{x\backslash i}^{(t-1)} = D_x^{(t-1)} - 1$

Assign a new label $x' = L_i^{(t-1)}$ to $\mathbf{W}_i$ with

$P(\, L \mid \mathbf{W}_i \,,\, L_0^{(t)} \ldots L_{i-1}^{(t)} \; L_{i+1}^{(t-1)} \ldots L_D^{(t-1)} ;\, \alpha, \beta, \gamma)$

Add $\mathbf{W}_i$ to the documents in class x'

Add its word counts $\mathbf{N}_i(w_j)$ to the word counts for class x'

Final estimate:

Use (some of the) snapshots $\mathbf{L}^{(1)} \ldots \mathbf{L}^{(T)}$ to estimate $P(+)$, $P(w_i \mid +)$, $P(w_i \mid -)$

Estimation

- Labels: $L \sim Bernoulli(\pi)$ Words: $\boldsymbol{W}_i \,|L \sim Multinomial(\boldsymbol{\theta}^L)$

	Supervised	Unsupervised
Freq.	**Relative frequency estimation** - Labels: $\pi = D^+/d$ - Words: $\boldsymbol{\theta_i}^+ = N^+(w_i)/N^+$	**Expectation Maximization:** At each iteration t: - Labels: $\pi^{(t)} = E[D]_{(t-1)}/d$ - Words: $\boldsymbol{\theta_i}^+ = E[N^+(w_i)]_{(t-1)}/E[N^+(w_i)]_{(t-1)}$
Bayes	**With priors:** - Labels: $\pi = (D^+ + \alpha)/(D + \alpha + \beta)$ - Words: $\boldsymbol{\theta_i}^+ = (N^+(w_i) + \gamma_i)/(N^+(w) + \gamma_0)$	**Gibbs sampling:** For each ministep i at each iteration t: - Labels: $\pi_i = (D^{+(-i)} + \alpha)/(D\text{-}1 + \alpha + \beta)$ - Words: $\boldsymbol{\theta_i}^+ = (N^{+\,(-i)}(w_i) + \gamma_i)/(N^{+\,(-i)}(w) + \gamma_0)$