Machine learning for NLP: Sampling Methods

Chenghua Lin
Computing Science
University of Aberdeen

Overall Schedule

- MCMC
- Gibbs sampling
- Deriving Gibbs sampler for LDA

Some recaps on ML and MAP

- ML does NOT allow us to inject our prior beliefs about the likely values for θ in the estimation calculations
- MAP allows for the fact that the parameter vector θ can take values from a distribution that expresses our prior beliefs regarding the parameters.
- Both ML and MAP return only single and specific values for the parameter θ.

Bayesian Estimation

- Bayesian estimation, by contrast, calculates fully the posterior distribution $p(\theta|x)$
- The variance that we can calculate for the parameter θ from its posterior distribution allows us to express our confidence in any specific value we may use as an estimate.
- If the variance is too large, we may declare that there does not exist a good estimate for θ. – That is it will tell you "I don't know.."

Bayesian Estimation

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

where $p(x) = \int_\theta \; p(x|\theta) \cdot p(\vartheta) d\theta$

- Now the denominator, known as the probability of evidence, is related to the other probabilities
- The denominator can no longer be ignored, and is yet often intractable …

Bayesian Estimation

$$p(x) = \int_{\theta} p(x|\theta) \cdot p(\vartheta) d\theta$$

- Integrals that involve probability density functions in the integrands are ideal for solution by Monte Carlo methods.

Monte-Carlo integration

- Suppose we wish to compute a complex integral:

$$\int_a^b h(x)dx$$

- Decomposing $h(x)$ into a production of a function $f(x)$ and a probability density function $p(x)$ yields

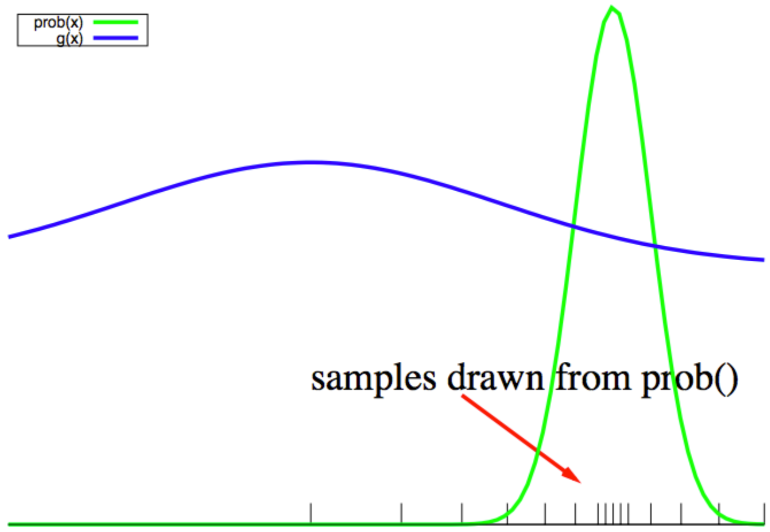
$$\int_a^b h(x)\,dx = \int_a^b f(x)\,p(x)\,dx = E_{p(x)}[\,f(x)\,]$$

- If we draw a large number $x_1, \ldots, x_n$ of random variables from the density $p(x)$, then

Monte-Carlo integration

$$\int_a^b h(x)\,dx = E_{p(x)}[\,f(x)\,] \simeq \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

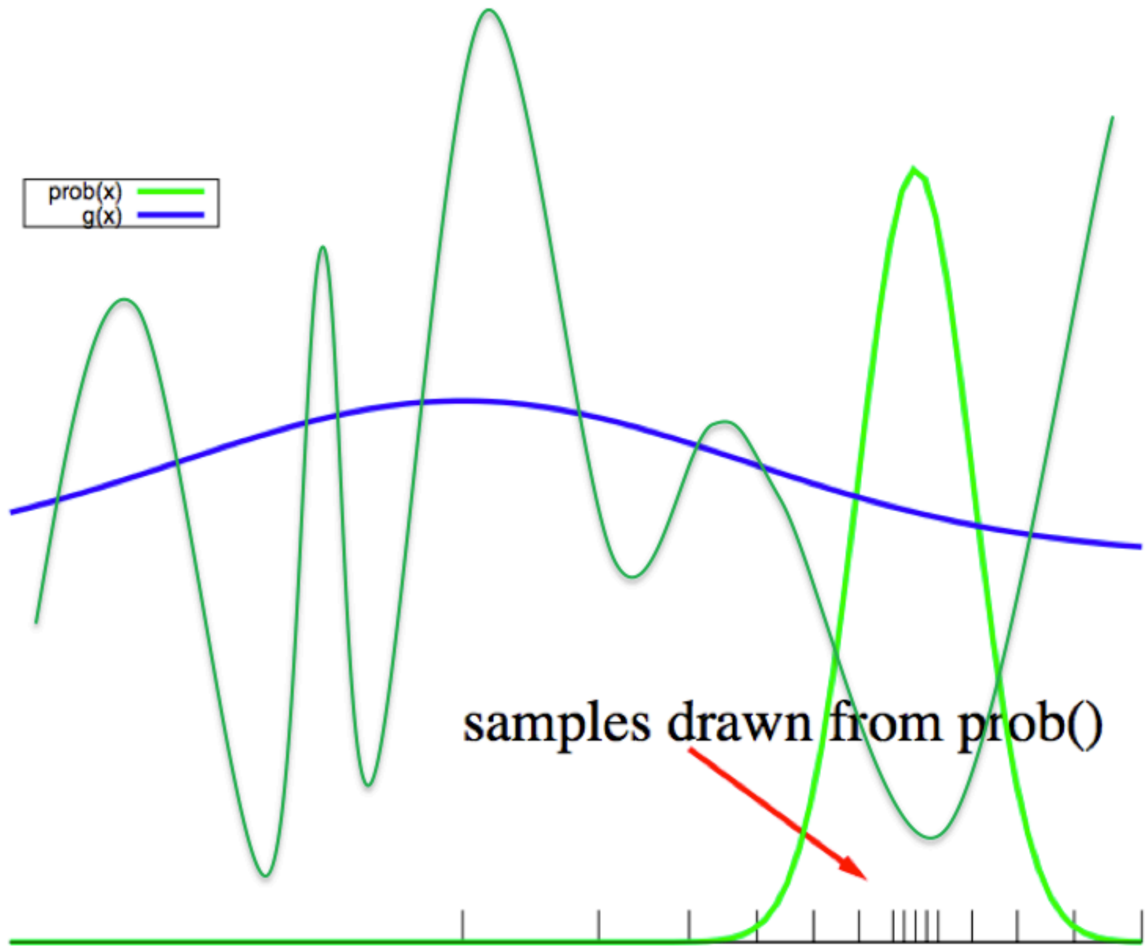
The easy case



- If $p(x)$ is simple, e.g., a uniform or a Gaussian density
- We can easily draw N samples x_i from $p(x)$, leading to unbiased estimate of integral
- This is the standard Monte-Carlo approach

$$\int_x p(x)f(x)dx = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

In reality …



How about when $p(x)$ is a complicated probability density function?

- The standard Monte-Carlo approach won't work
- Simply because it is non-trivial to sample complicated density functions algorithmically.

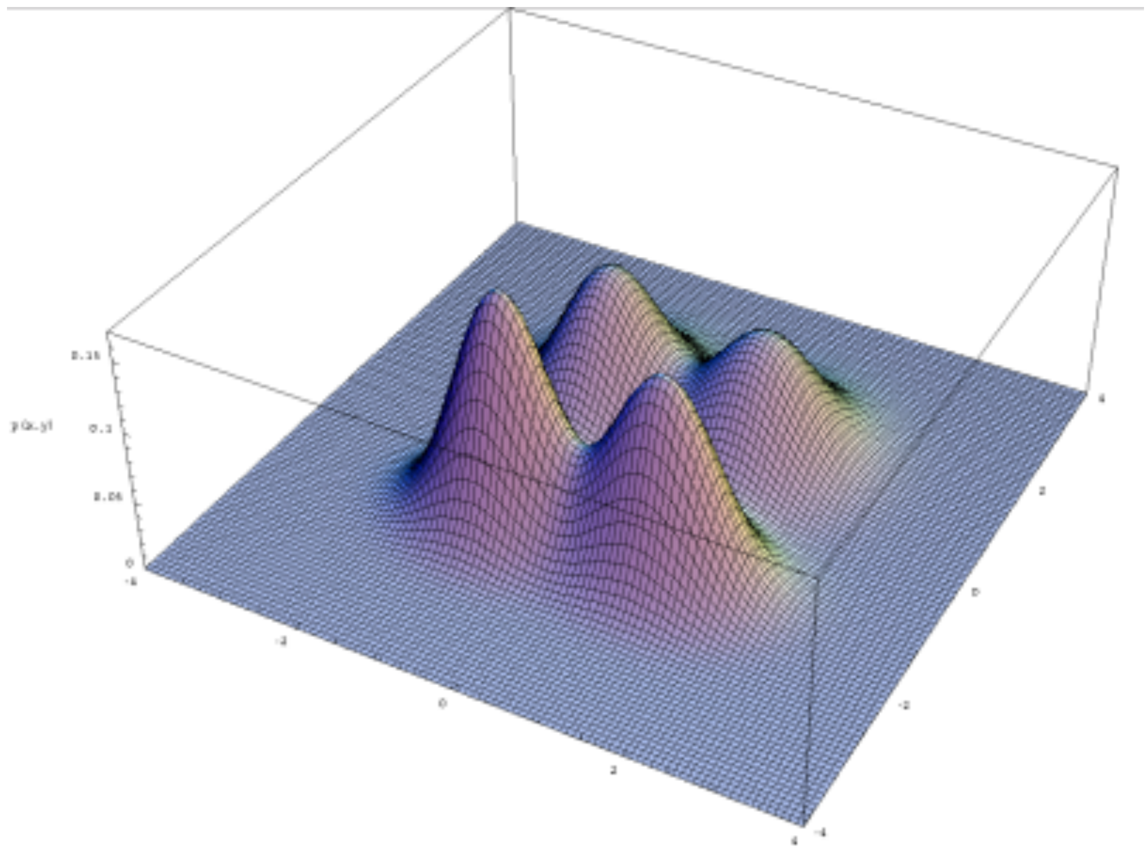
Markov-Chain Monte-Carlo (MCMC)

- Modern approaches for drawing samples from an arbitrary probability distribution $p(x)$ for the purpose of Monte-Carlo integration are based on MCMC
- An approach that draws samples from a distribution for the purpose of Monte-Carlo integration of complex integrands is commonly referred as the MCMC sampler.
- In MCMC, each sample chosen depends on just the sample selected previously, and a sequence of such samples forms a **Markov chain**.
- The resulting Markov chain has the desired distribution

MCMC vs. Gibbs Sampling

- MCMC
  - Suppose we have a vector variable X of an arbitrary number of dimensions, i.e. $X = (x_1, .., x_n)^T$
  - Assume that we do K MCMC samplings, the MCMC sampler directly gives us a sequence of samples in the *n*-dimensional space spanned by X, i.e. $X_1, X_2, \ldots, X_K$
- Gibbs Sampling
  - Gibbs sampler is a special case of the MCMC sampler.
  - Based on the observation that even when the joint distribution $p(X)$ is multimodal, the univariate conditional distribution for each $p(x_i)$ when all the other variables are held constant is likely to be approximable by a relatively **easy unimodal distribution**, such as **uniform** or **normal**.
  - Samples each dimension of X separately through the univariate conditional distribution along that dimension against the rest

Multimodal distribution



Gibbs Sampling

- Some notations:
  - As in the previous slide, we make the individual components of X explicit by writing $X = (x_1, .., x_n)^T$.
  - We will also write $X^{(-i)} = (x_1, .. x_{i-1}, x_{i+1}, \dots, x_n)^T$
  - Let's now focus on the *N* univariate conditional distributions: $p(x_i | X^{(-i)})$ for $i = 1, \dots, n$.
  - Keep in mind the fact that a conditional distribution for the component scalar variable x_i makes sense only when the other $n - 1$ variables in $X^{(-i)}$ are given constant values.

Gibbs sampling Procedures

Task:

– Sampling individual scalar variables

Steps:

1. Initialize $\{x_i : i = 1, \dots, N\}$
2. For $k = 1, \dots, K$:
   - Sample $x_1^{k+1} \sim p\left(x_2^k, x_3^k, \dots, x_N^k\right)$.
   - Sample $x_2^{k+1} \sim p\left(x_1^{k+1}, x_3^k, \dots, x_N^k\right)$.

     $\vdots$
   - Sample $x_j^{k+1} \sim p\left(x_1^{k+1}, \dots, x_{j-1}^{k+1}, x_{j+1}^k \dots, x_N^k\right)$

     $\vdots$
   - Sample $x_N^{k+1} \sim p\left(x_1^{k+1}, x_2^{k+1}, \dots, x_{N-1}^{k+1}\right)$

Gibbs sampling Procedures

Steps:

1. Initialize $\{x_i : i = 1, \dots, N\}$
2. For $k = 1, \dots, K$:
   - Sample $x_1^{k+1} \sim p\left(x_2^k, x_3^k, \dots, x_N^k\right)$.
   - Sample $x_2^{k+1} \sim p\left(x_1^{k+1}, x_3^k, \dots, x_N^k\right)$.
   - $\vdots$
   - Sample $x_j^{k+1} \sim p\left(x_1^{k+1}, \dots, x_{j-1}^{k+1}, x_{j+1}^k \dots, x_N^k\right)$
   - $\vdots$
   - Sample $x_N^{k+1} \sim p\left(x_1^{k+1}, x_2^{k+1}, \dots, x_{N-1}^{k+1}\right)$

- In this manner, we complete one "scan" through all the *N* dimensions of X.
- In the next scan, we now use the previously calculated sample values for the conditioning variables and proceed in exactly the same manner as above.
- After ***K* such scans** through the component variables, we end up with *K* sampling points for vector variable X.

A basic Gibbs sampling example

- Suppose we want to simulate from joint distribution

$$\pi(x_1, x_2): \quad \begin{array}{cc|ccc} & & & X_2 & \\ & & 1 & 2 & 3 \\ \hline & 1 & .1 & 0 & .1 \\ X_1 & 2 & .1 & .2 & .1 \\ & 3 & .1 & .1 & .2 \end{array}$$

A basic Gibbs sampling example

$\pi_1(x_1 \mid x_2)$:

		X_2 = 1	2	3
X_1	1	1/3	0	1/4
	2	1/3	2/3	1/4
	3	1/3	1/3	1/2

$\pi_2(x_2 \mid x_1)$:

		X_2 = 1	2	3
X_1	1	1/2	0	1/2
	2	1/4	1/2	1/4
	3	1/4	1/4	1/2

Assume that

- arbitrarily we start at $X^{(1)} = \binom{1}{2}$
- our uniform random numbers are *.70, .41, .28, .48, .30, .47*

- We begin by changing X_1 using π_1 when $X_2 = 2$, i.e. we use the second column of π_1:

$$\pi_1(x_1 \,|\, x_2 = 2): \quad \begin{array}{cc|c} & & X_2 \\ & & 2 \\ \hline & 1 & 0 \\ X_1 & 2 & 2/3 \\ & 3 & 1/3 \end{array}$$

  Since $2/3 < .7 \leq 2/3 + 1/3$, we set $X_1^{(2)} = 3$.

- Now we change X_2 using π_2 when $X_1 = 3$, i.e. we use the third row of π_2:

$$\pi_2(x_2 \,|\, x_1 = 3): \quad \begin{array}{cc|ccc} & & & X_2 & \\ & & 1 & 2 & 3 \\ \hline X_1 & 3 & 1/4 & 1/4 & 1/2 \end{array}$$

  Since $1/4 < .41 \leq 1/4 + 1/4$, we set $X_2^{(2)} = 2$.

- Putting these two stages together, we now have $X^{(2)} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$

- Since $X_2^{(2)} = 2$, we again use the second column of π_1:

$\pi_1(x_1 \,|\, x_2 = 2)$:

		X_2
		2
	1	0
X_1	2	2/3
	3	1/3

  Since $.28 \le 2/3$, we set $X_1^{(3)} = 2$.

- This time we need the second row of π_2

$\pi_2(x_2 \,|\, x_1)$:

		X_2		
		1	2	3
X_1	2	1/4	1/2	1/4

  Since $1/4 < .48 \le 1/4 + 1/2$, $X_2^{(3)} = 2$.

- $X^{(3)} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$

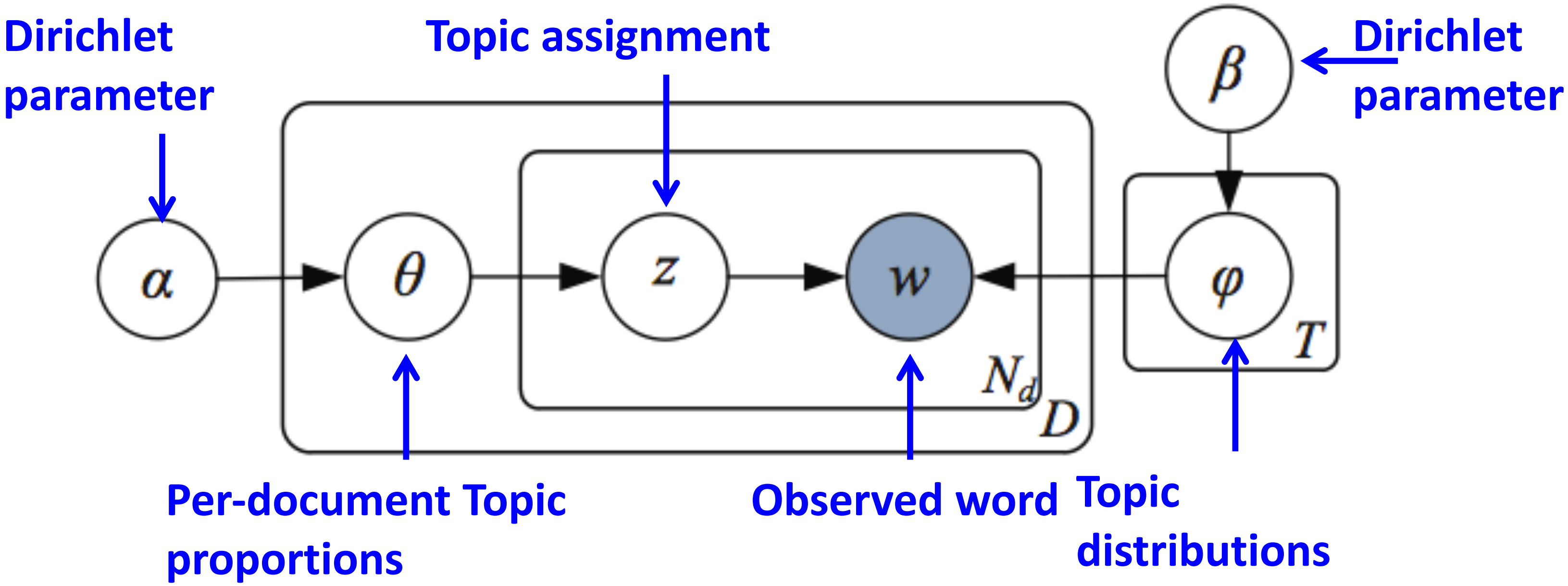
A basic Gibbs sampling example

- Again use second column of π_1 and .30 gives $X_1^{(4)} = 2$.
- Second row of π_2 with .47 gives $X_2^{(4)} = 2$
- $X^{(4)} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$

- So on and so forth, we run Gibbs samplings until the Markov chain reaches the stationary status.

Deriving a Gibbs sampler for LDA

LDA Posterior



- θ: per-document topic proportion
- φ: per-corpus topic word distribution
- Z: per-word topic assignment

Posterior

- The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document:

$$P(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\mathbf{w} | \boldsymbol{\alpha}, \boldsymbol{\beta})}$$

Intractable Posterior

$$P(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\mathbf{w} | \boldsymbol{\alpha}, \boldsymbol{\beta})}$$

$$p(\boldsymbol{w}|\alpha, \beta) = \iint p(\theta_d|\alpha) \cdot p(\varphi|\beta) \cdot \prod_{n=1}^{N_d} p(w_{d,n}|\theta_d, \varphi) d\varphi d\theta_d$$

- The integral in this expression is intractable due to the coupled parameters θ and φ, and is thus usually estimated by using
  - MCMC approaches, e.g. Gibbs Sampling
  - Variational Bayes
  - Expectation propagation

Gibbs Sampling

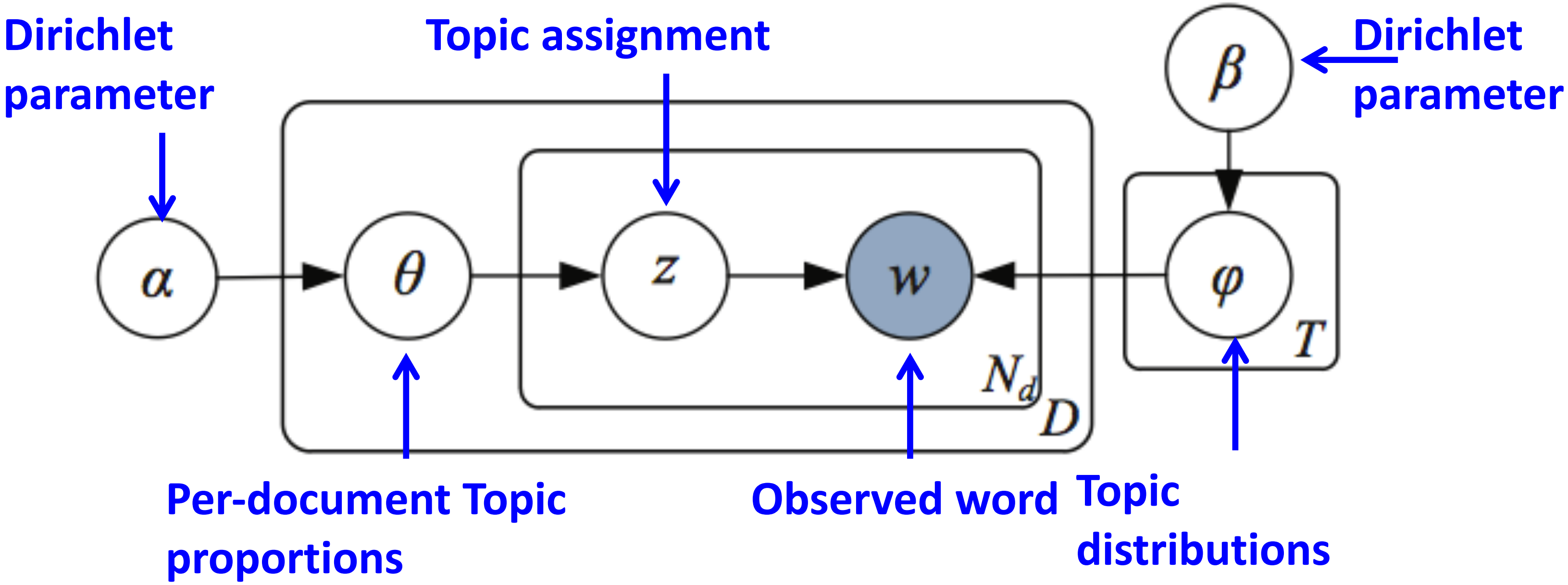
- Do not explicitly representing θ or φ as parameters to be estimated.
- Instead considering the joint distribution over the assignments of words to topics, $p(\boldsymbol{z}, \boldsymbol{w})$
- We then obtain estimates of θ and φ by examining this joint distribution.
- Evaluating $p(\boldsymbol{z}, \boldsymbol{w})$ using Gibbs sampling

Gibbs Sampler for LDA

- Recall that Gibbs sampling operates on a univariate conditional distribution $p(x_i|X^{(-i)})$ for $i = 1, \ldots, n$
- The first step for depriving the Gibbs sampler for LDA is to work out the target conditional distribution of interest.

$$P(z_t|\mathbf{z}^{\neg t}, \mathbf{w})$$

LDA Posterior



- θ: per-document topic proportion
- φ: per-corpus topic word distribution
- Z: per-word topic assignment

Gibbs Sampler for LDA

$$\boxed{\begin{aligned} P(\mathbf{w},\mathbf{z}) &= P(\mathbf{w}|\mathbf{z})P(\mathbf{z}) \\ &= \int P(\mathbf{w}|\mathbf{z},\mathbf{\Phi})P(\mathbf{\Phi}|\boldsymbol{\beta})\,d\mathbf{\Phi}\cdot\int P(\mathbf{z}|\mathbf{\Theta})P(\mathbf{\Theta}|\boldsymbol{\alpha})\,d\mathbf{\Theta}. \end{aligned}}$$

$$P(\mathbf{w}|\mathbf{z},\mathbf{\Phi}) = \prod_{i=1}^{W} P(w_i|z_i) = \prod_{k=1}^{K}\prod_{\forall i:\, z_i=k} P(w_i=t|z_i=k) = \prod_{k=1}^{K}\prod_{t=1}^{V}\varphi_{k,t}^{n_k^{(t)}},$$

$$P(\mathbf{\Phi}|\boldsymbol{\beta}) = \frac{\Gamma(\sum_{t=1}^{V}\beta_t)}{\prod_{t=1}^{V}\Gamma(\beta_t)}\prod_{t=1}^{V}\varphi_{k,t}^{\beta_t-1}\,d\varphi_k,$$

$$P(\mathbf{z}|\mathbf{\Theta}) = \prod_{i=1}^{W} P(z_i|d_i) = \prod_{m=1}^{M}\prod_{\forall i:\, d_i=m} P(z_i=k|d_i=m) = \prod_{m=1}^{M}\prod_{k=1}^{K}\theta_m^{n_m^{(k)}},$$

$$P(\mathbf{\Theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{m,k}^{\alpha_k-1}\,d\theta_m.$$

Per-document topic proportion

$$\boxed{\begin{aligned} P(\mathbf{w},\mathbf{z}) &= P(\mathbf{w}|\mathbf{z})P(\mathbf{z}) \\ &= \int P(\mathbf{w}|\mathbf{z},\mathbf{\Phi})P(\mathbf{\Phi}|\boldsymbol{\beta})\,d\mathbf{\Phi} \cdot \int P(\mathbf{z}|\mathbf{\Theta})P(\mathbf{\Theta}|\boldsymbol{\alpha})\,d\mathbf{\Theta}. \end{aligned}}$$

$$\begin{aligned} P(\mathbf{z}) &= \int P(\mathbf{z}|\mathbf{\Theta})P(\mathbf{\Theta}|\boldsymbol{\alpha})\,d\mathbf{\Theta} \\ &= \int \prod_{d=1}^{D}\prod_{j=1}^{T}\theta_{d,j}^{N_{d,j}} \frac{\Gamma(\sum_{j=1}^{T}\alpha_j)}{\prod_{j=1}^{T}\Gamma(\alpha_j)} \prod_{j=1}^{T}\theta_{d,j}^{\alpha_j-1}\,d\boldsymbol{\theta}_d \\ &= \prod_{d=1}^{D} \frac{\Gamma(\sum_{j=1}^{T}\alpha_j)}{\prod_{j=1}^{T}\Gamma(\alpha_j)} \frac{\prod_{j=1}^{T}\Gamma(N_{d,j}+\alpha_j)}{\Gamma(\sum_{j=1}^{T}N_{d,j}+\alpha_j)} \\ &= \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_d \frac{\prod_j \Gamma(N_{d,j}+\alpha)}{\Gamma(N_d+T\alpha)}, \end{aligned}$$

Per-corpus word probability

$$\boxed{\begin{aligned} P(\mathbf{w},\mathbf{z}) &= P(\mathbf{w}|\mathbf{z})P(\mathbf{z}) \\ &= \int P(\mathbf{w}|\mathbf{z},\mathbf{\Phi})P(\mathbf{\Phi}|\boldsymbol{\beta})\,d\mathbf{\Phi}\cdot\int P(\mathbf{z}|\mathbf{\Theta})P(\mathbf{\Theta}|\boldsymbol{\alpha})\,d\mathbf{\Theta}. \end{aligned}}$$

$$\begin{aligned} P(\mathbf{w}|\mathbf{z}) &= \int P(\mathbf{w}|\mathbf{z},\mathbf{\Phi})P(\mathbf{\Phi}|\boldsymbol{\beta})\,d\mathbf{\Phi} \\ &= \int \prod_{j=1}^{T}\prod_{i=1}^{V}\varphi_{j,i}^{N_{j,i}}\,\frac{\Gamma(\sum_{i=1}^{V}\beta_{j,i})}{\prod_{i=1}^{V}\Gamma(\beta_{j,i})}\prod_{i=1}^{V}\varphi_{j,i}^{\beta_{j,i}-1}\,d\boldsymbol{\varphi}_j \\ &= \prod_{j=1}^{T}\frac{\Gamma(\sum_{i=1}^{V}\beta_{j,i})}{\prod_{i=1}^{V}\Gamma(\beta_{j,i})}\,\frac{\prod_{i=1}^{V}\Gamma(N_{j,i}+\beta_{j,i})}{\Gamma(\sum_{i=1}^{V}N_{j,i}+\beta_{j,i})} \\ &= \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^T\prod_j\frac{\prod_i\Gamma(N_{j,i}+\beta)}{\Gamma(N_j+V\beta)}, \end{aligned}$$

The conditional distribution for the Gibbs Sampler

By making use of the probability of Gamma function:

$$\Gamma(n) = (n-1)!$$

We finally yield:

$$\begin{aligned}
P(z_t = j | \mathbf{w}, \mathbf{z}^{\neg t}) &= \frac{P(\mathbf{w}, \mathbf{z})}{P(\mathbf{w}, \mathbf{z}^{\neg t})} = \frac{P(\mathbf{w}|\mathbf{z})}{P(\mathbf{w}^{\neg t}|\mathbf{z}^{\neg t})P(w_t)} \cdot \frac{P(\mathbf{z})}{P(\mathbf{z}^{\neg t})} \\
&\propto \frac{\Gamma(N_{j,i} + \beta)\Gamma(N_j^{\neg t} + V\beta)}{\Gamma(N_{j,i}^{\neg t} + \beta)\Gamma(N_j + V\beta)} \cdot \frac{\Gamma(N_{d,j} + \alpha)\Gamma(N_d^{\neg t} + T\alpha)}{\Gamma(N_{d,j}^{\neg t} + \alpha)\Gamma(N_d + T\alpha)} \\
&\propto \frac{N_{j,i}^{\neg t} + \beta}{N_j^{-t} + V\beta} \cdot \frac{N_{d,j}^{\neg t} + \alpha}{N_d^{\neg t} + T\alpha}.
\end{aligned}$$

Summary

What you should know

- Monte-Carlo integration
- MCMC
- Gibbs sampling – a special case of MCMC