Reasoning (IV)

Mingsheng Long

Tsinghua University

Spring 2023

Outline

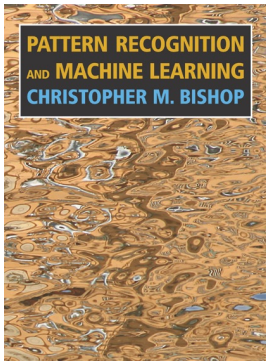
- **Sampling Method**
  - **Monte Carlo**
  - Markov Chain Monte Carlo
    - Metropolis-Hastings Algorithm
    - Gibbs Sampling
- Sampling Method for LDA
  - Bayesian Estimation
    - Dirichlet-Multinomial Conjugate
  - Gibbs Sampling for LDA

ELBO and EM Algorithm

- Evidence Lower Bound (ELBO):

$$\mathcal{L}(q,\theta) = \sum_z q(z)\log\left(\frac{p(x,z|\theta)}{q(z)}\right)$$

$$= -\mathrm{KL}[q(z)||p(z|x,\theta)] + \log p(x|\theta)$$



PRML
Chapter 11

- EM Algorithm:

1. Choose initial $\theta^{\mathbf{old}}$
2. E-step: Let $q^* = \underset{q}{\mathrm{argmax}}\,\mathcal{L}(q,\theta^{\mathbf{old}})$
3. M-step: Let $\theta^{\mathbf{new}} = \underset{\theta}{\mathrm{argmax}}\,\mathcal{L}(q^*,\theta)$
4. Go to step 2, until converged.

Minimize

Optimal choice for E-step
$q^*(z) = p(z|x,\theta^{\mathbf{old}})$

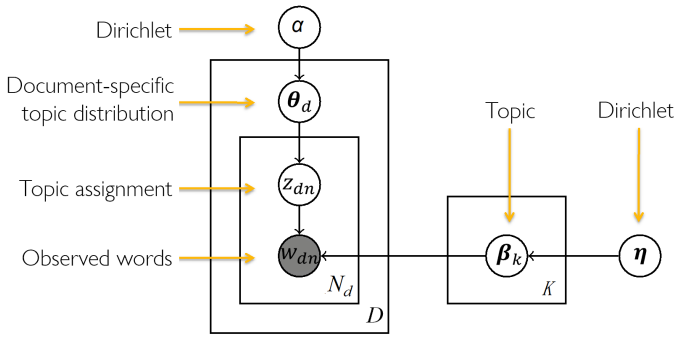
Intractable E-Step

- We need to solve: Optimal choice for E-step $q^*(z) = p(z|x, \theta^{\text{old}})$

In LDA: $p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta} \mid \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \dfrac{p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}$

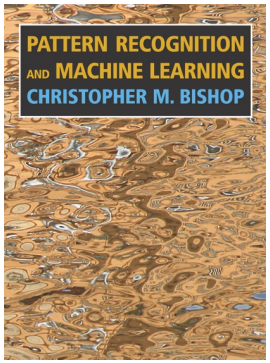
- But the denominator is intractable!

$$p(\mathbf{w}|\alpha, \eta) = \int \int \sum_{\mathbf{z}} p(\boldsymbol{\theta}, \mathbf{z}, \boldsymbol{\beta}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\eta})\, d\boldsymbol{\theta}\, d\boldsymbol{\beta}$$



- We need approximation here!
  - There are two ways to solve E-step approximately:
    - Variational Inference (last lecture)
    - Sampling – Markov Chain Monte Carlo (this lecture)

Sampling Method



PRML
Chapter 11

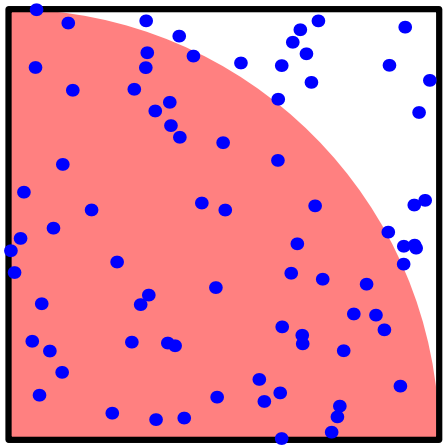
- We want to estimate the integral on some distribution:

$$\int_{x\in\mathcal{X}} f(x)p(x)\mathrm{d}x = \mathbb{E}_p f(x)$$

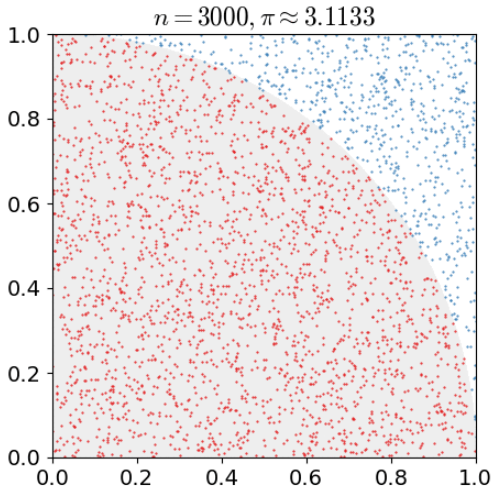
- Sampling method:
  - Draw an i.i.d. sample set from distribution p: $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$
  - In practice: $\{\boldsymbol{x}_{1:n}\}$ not independent, effective sample size is small.
  - We can only generate uniform pseudo-random numbers in $U[0,1]$.
- Monte Carlo Estimator: $\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)$
  - Unbiased: $\mathbb{E}_{\mathcal{D}\sim p^n} \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{x}_i\sim p} f(\boldsymbol{x}_i) = \mathbb{E}_p f(x)$
  - Variance: $\mathrm{Var}_{\mathcal{D}\sim p^n}\left[\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}_p[f(\boldsymbol{x}_i)] = \frac{1}{n}\mathrm{Var}_p[f]$

Monte Carlo Method

- Many problems can be changed into integral form.
- Example: the computation of π.
  - View π as a quarter of area of a unit circle.



$$\pi = 4 \iint \mathbb{I}(x^2 + y^2 < 1) P(x, y) \mathrm{d}x \mathrm{d}y$$



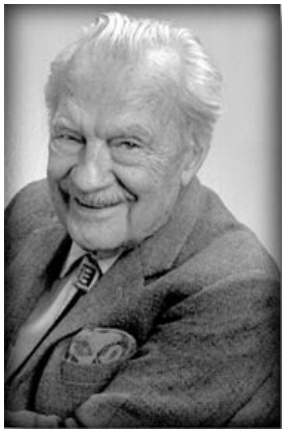
  - Consider a uniform distribution on $U[0,1]^2$

$$P(x, y) = \begin{cases} 1 & 0 < x < 1 \text{ and } 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Monte Carlo Method

Stanislaw Ulam

John von Neumann



Nicholas Metropolis

- For being secret, this work required a code name.
  - Monte Carlo is a city in Monaco.

- "Monte Carlo is an extremely bad method; it should be used only when all alternative methods are worse." — Alan Sokal, 1996

Monte Carlo EM Algorithm

- Evidence lower-bound (ELBO):

$$\mathcal{L}(q,\theta) = \sum_z q(z)\log\left(\frac{p(x,z|\theta)}{q(z)}\right) = \mathbb{E}_{z\sim q}\log\left(\frac{p(x,z|\theta)}{q(z)}\right)$$

- EM Algorithm:
  - E-step: $q^* = \underset{q}{\operatorname{argmax}}\,\mathcal{L}\left(q,\theta^{\text{old}}\right) = p\left(z|x,\theta^{\text{old}}\right)$ Sampling form it!
  - M-Step: $\theta^{\text{new}} = \underset{\theta}{\operatorname{argmax}}\,\mathcal{L}(q^*,\theta) = \underset{\theta}{\operatorname{argmax}}\,\mathbb{E}_{z\sim q}\log\left(\frac{p(x,z|\theta)}{q(z)}\right)$

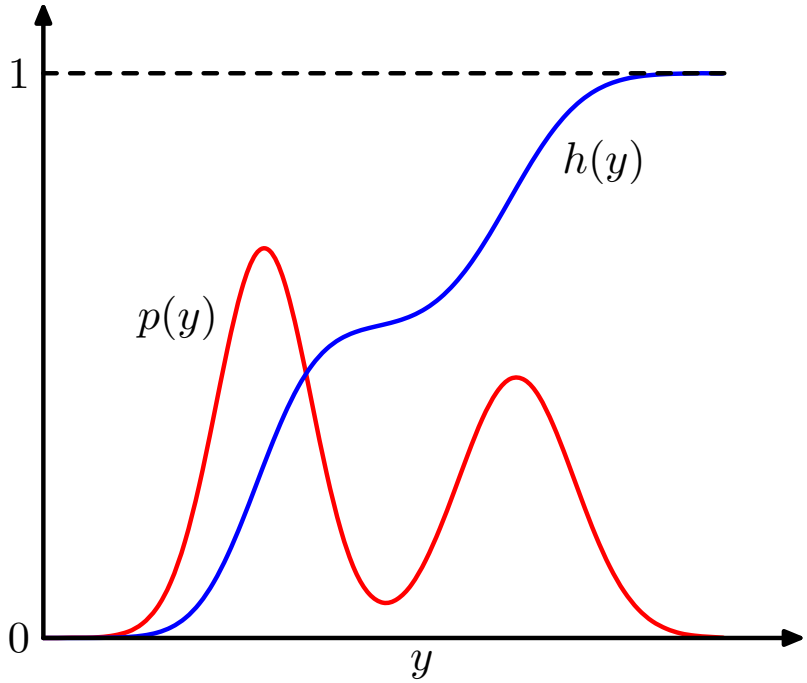
$$= \underset{\theta}{\operatorname{argmax}}\,\mathbb{E}_{z\sim q}\log p(x,z|\theta) \approx \underset{\theta}{\operatorname{argmax}}\frac{1}{T}\sum_{i=1}^{T}\log p(x,z_i|\theta)$$

- We don't need to explicitly compute q^* but sample $z_1,\ldots,z_T$ from q^*!

Sampling from Distribution

- **Key problem:** How to sample from some complex distribution p?
- Suppose we know the sampling method of some simple distribution.
  - We convert samples to target distributions p from the simple one.
- For example, if we can sample from
  - Uniform distribution on $[0,1]$.
- For a target p with explicit CDF:
  - $h(x) = \int_{-\infty}^{x} p(z)\,\mathrm{d}z$

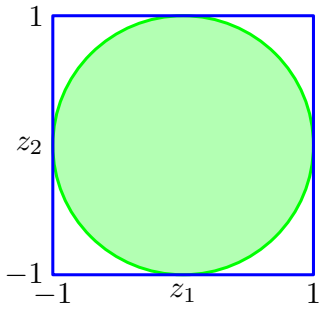
- First, sample $u \sim U[0,1]$.
- Second, compute $x(u) = h^{-1}(u)$



Example: Sampling from Gaussian

- **Box-Muller method:** generate samples from Gaussian distribution
  - Generate pairs of uniform random numbers $z_1, z_2 \in (-1,1)$
  - Discard each pair unless it satisfies $z_1^2 + z_2^2 \leq 1$
    - A uniform distribution inside the unit circle with $p(z_1, z_2) = \frac{1}{\pi}$
  - Define $r^2 = z_1^2 + z_2^2$, and

$$y_1 = z_1 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}, y_2 = z_2 \left(\frac{-2\ln r^2}{r^2}\right)^{1/2}$$



- Then $p(y_1, y_2) = p(z_1, z_2)\left|\frac{\partial(z_1,z_2)}{\partial(y_1,y_2)}\right|$

$$= \left[\frac{1}{\sqrt{2\pi}}\exp(-y_1^2/2)\right]\left[\frac{1}{\sqrt{2\pi}}\exp(-y_2^2/2)\right]$$

Rejection Sampling

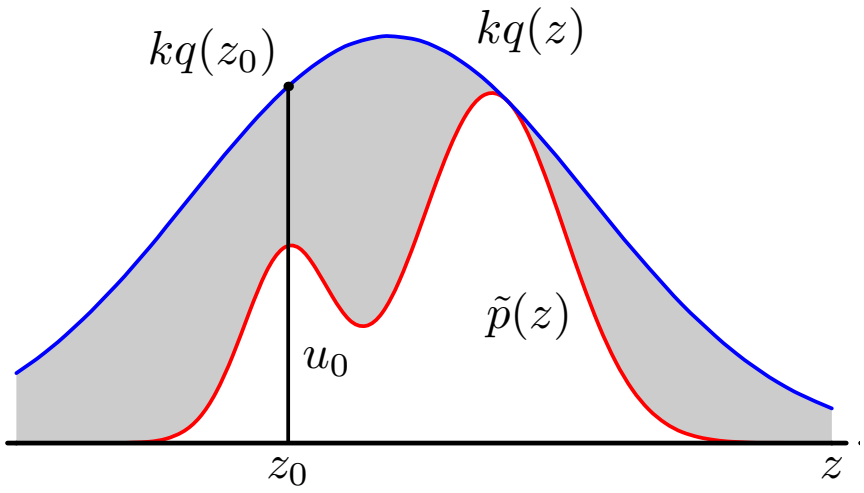
When unable to compute and invert $h(x)$.

→ Sampling based on $p(x)$.

Easy to evaluate $p(x) = \frac{\tilde{p}(x)}{Z}$

unknown

- Find a proposal distribution q:
  - We can sample from it.
  - $\exists k, k \cdot q(x) \geq p(x)$



- Rejection Sampling:
  - Draw $x \sim q, u \sim U[0, k \cdot q(x)]$
  - Reject x if $u > p(x)$.

- $p(x|\text{accept}) \propto q(x)p(\text{accept}|x) = q(x)\left(\frac{p(x)}{k \cdot q(x)}\right) = \frac{p(x)}{k}$

- The acceptance rate decreases exponentially with dimensionality of x.

Outline

- **Sampling Method**
  - Monte Carlo
  - **Markov Chain Monte Carlo**
    - Metropolis-Hastings Algorithm
    - Gibbs Sampling
- Sampling Method for LDA
  - Bayesian Estimation
    - Dirichlet-Multinomial Conjugate
  - Gibbs Sampling for LDA

Monte Carlo for LDA

- We want to compute the following terms in E-step of LDA:

$$p(z \mid w, \theta, \beta) = \frac{p(z, w \mid \beta, \theta)}{p(w \mid \beta, \theta)}$$

Easier to compute!

Intractable! No explicit PDF.

- The problem turns into a canonical problem:

  - If we want to sample from a distribution:

$$p(x = j) = \pi_j \propto b_j$$

  - Where b_j is computable, the value of $\sum_{j=1}^{m} b_j = B$ is unknown.

- Question: How to sample from distribution $p(x)$ for Monte Carlo?

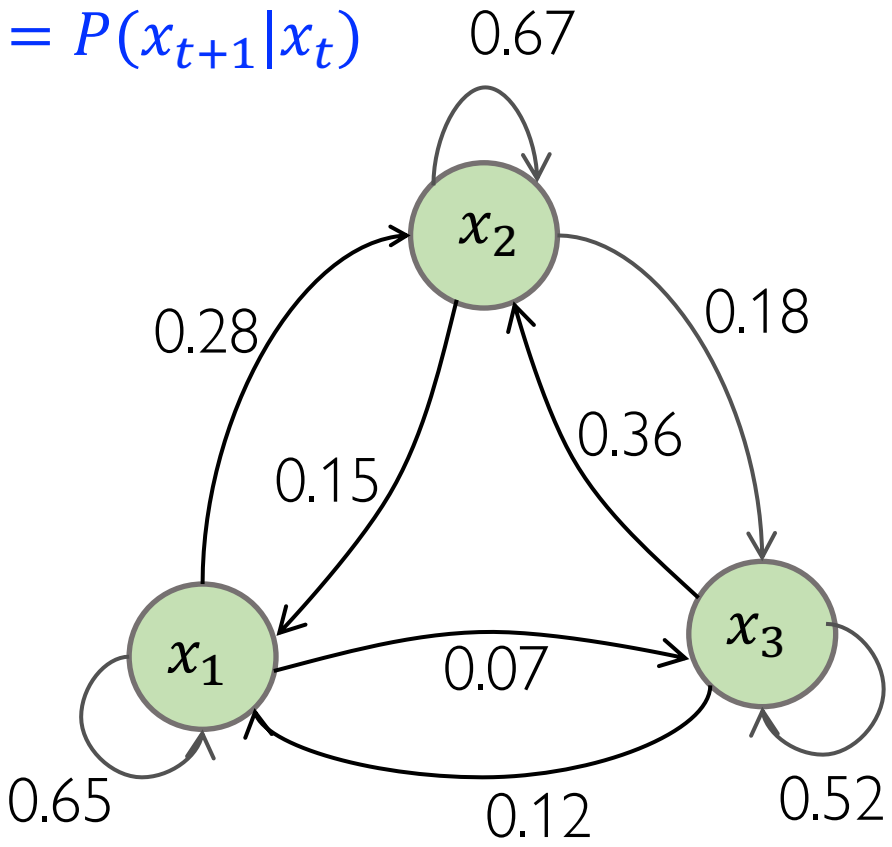
Markov Chain

- Markov Chain:

$$P(x_{t+1}|x_1, \dots, x_t) = P(x_{t+1}|x_t)$$

- Assume that the state space is finite:
  - $\mathcal{X} = \{1, \dots, k\}$
- Transition matrix $\boldsymbol{P}$:
  - $P_{ij} = P(x_{t+1} = j | x_t = i)$
- For example:

$$\boldsymbol{P} = \begin{bmatrix} 0.65 & 0.28 & 0.07 \\ 0.15 & 0.67 & 0.18 \\ 0.12 & 0.36 & 0.52 \end{bmatrix}$$



- Describe the probability of transition from one state to other states.

Convergence of Markov Chain

- We use state distribution to represent probability of state at time $\boldsymbol{t}$:

$$\boldsymbol{\pi}^t = (\pi_1^t \quad \dots \quad \pi_k^t)$$

  - **One transition**: $\boldsymbol{\pi}^{t+1} = \boldsymbol{\pi}^t \boldsymbol{P}$

$$p(x_{t+1} = i) = \sum_{j=1}^{k} p(x_t = j) P_{ji}$$

- An example:
  - If $\boldsymbol{\pi}^0 = (0.21 \quad 0.68 \quad 0.11)$ and transit for T times:

$$\boldsymbol{\pi}^T = \boldsymbol{\pi}^0 (\boldsymbol{P})^T$$

  - The result will be $\boldsymbol{\pi}^{10} = (0.286 \quad 0.489 \quad 0.225)$
  - If we change the **initial state**, the result will remain unchanged.
- Observation: $\boldsymbol{\pi}^{10} \boldsymbol{P} = \boldsymbol{\pi}^{10}$.

Convergence of Markov Chain

- **Ergodicity:** A Markov chain has a unique stationary distribution if:
  - From all initial states $\boldsymbol{\pi}^0$, the chain will converge to $\boldsymbol{\pi}$:

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$$

    - After a sufficient number of transitions.
  - Each transition is equivalent to sampling from distribution $\boldsymbol{\pi}$.

- **Detailed balance condition**
  - Markov chain is ergodic if and only if: $\pi_i P_{ij} = \pi_j P_{ji}$

$$\sum_{i=1}^{+\infty} \pi_i P_{ij} = \sum_{i=1}^{+\infty} \pi_j P_{ji} = \pi_j \sum_{i=1}^{+\infty} P_{ji} = \pi_j$$

Sampling with Markov Chain

- The problem turns into a canonical problem:

  - If we want to sample from a distribution:

$$p(x = j) = \pi_j \propto b_j$$

  - Where b_j is computable, the value of $\sum_{j=1}^{m} b_j = B$ is unknown.

- Question: How to sample from distribution $p(x)$ for Monte Carlo?

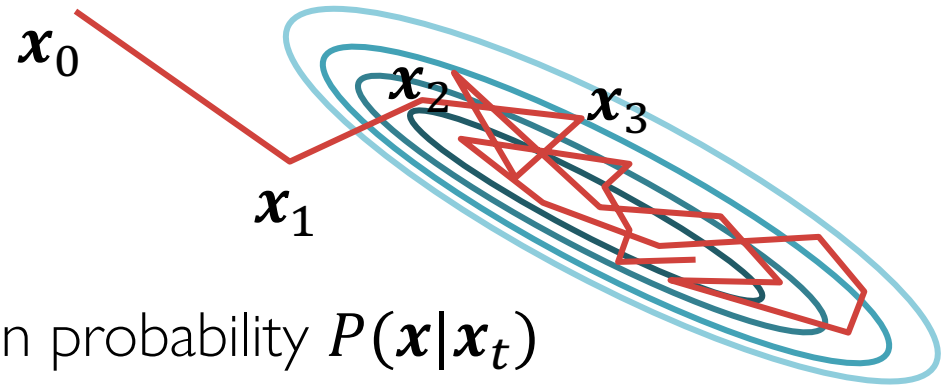
- Answer: Find some $\boldsymbol{P}$ such that the stationary distribution is $\pi_j \propto b_j$:

  - We use random samples transiting with $\boldsymbol{P}$:
  - After several transitions, the state distribution will converge to $\boldsymbol{\pi}$.
  - The transiting samples after convergence will follow distribution $\boldsymbol{\pi}$.

Markov Chain Monte Carlo

- Given a Markov chain with known transition matrix $\boldsymbol{P}$ (unknown!).
  - Assume that detailed balance condition is hold: $\pi_i P_{ij} = \pi_j P_{ji}$

- Markov chain sampling:
  - Start from random $\boldsymbol{x_0}$
  - For $t = 0, \ldots, T-1$
    - Sample $\boldsymbol{x_{t+1}}$ from transition probability $P(\boldsymbol{x}|\boldsymbol{x_t})$
    - Assume convergence at time T_0

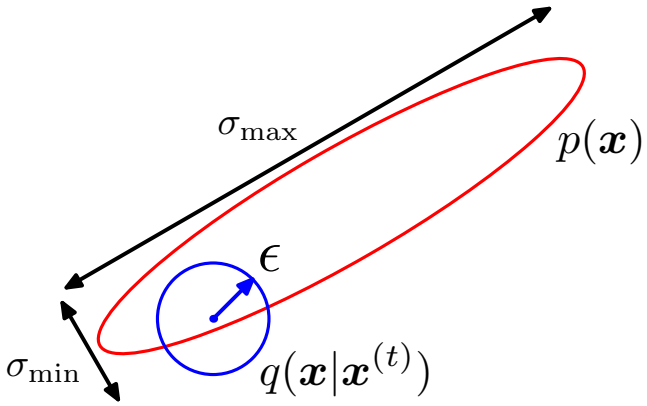


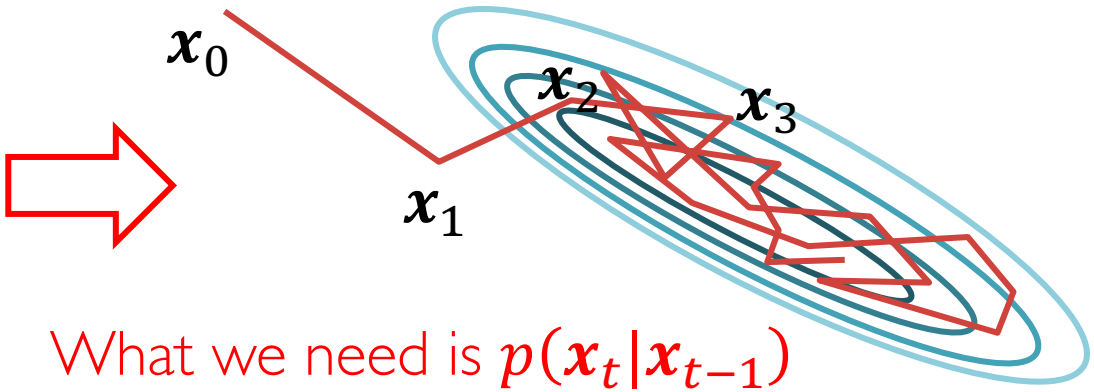
- Then the sample set $\{\boldsymbol{x}_{T_0}, \ldots, \boldsymbol{x}_T\}$ follows stationary distribution $\boldsymbol{\pi}$.

- Monte Carlo with Markov Chain: $\frac{1}{T-T_0}\sum_{t=T_0}^{T-1} f(\boldsymbol{x_t}) \to \mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{\pi}} f(\boldsymbol{x})$

Markov Chain Monte Carlo

- We do not know transition matrix $\boldsymbol{P}$ in practice. How to sample?
- We first choose a proposal distribution $\boldsymbol{q}$ as transition probability:
  - Usually set as Gaussian: $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\epsilon}^2)$.
    - This distribution is symmetric: $q(\boldsymbol{x}|\boldsymbol{y}) = q(\boldsymbol{y}|\boldsymbol{x})$
  - Forms Markov Chain with transition matrix $\boldsymbol{Q}$: $Q_{ij} = \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{x}_i, \boldsymbol{\epsilon}^2)$





What we need is $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$

Markov Chain Monte Carlo

- Transition matrix by proposal distribution $\boldsymbol{Q}$: $Q_{ij} = \mathcal{N}\left(\boldsymbol{x}_j | \boldsymbol{x}_i, \boldsymbol{\epsilon}^2\right)$
  - Detailed balance condition may be violated:

$$\pi_i Q_{ij} \neq \pi_j Q_{ji}$$

- We add acceptance rate α_{ij} and α_{ji} to keep it balance:

$$\pi_i Q_{ij} \alpha_{ij} = \pi_j Q_{ji} \alpha_{ji}$$

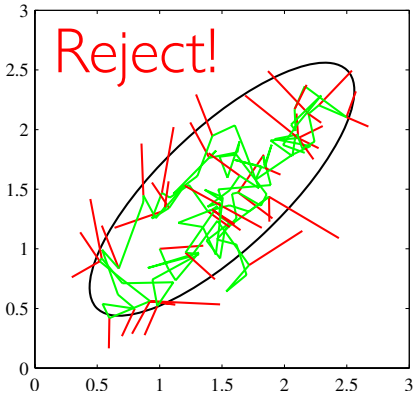
  - Let new transition matrix $P_{ij} = Q_{ij}\alpha_{ij}$ and $P_{ji} = Q_{ji}\alpha_{ji}$
  - Then we have $\pi_i P_{ij} = \pi_j P_{ji}$.
  - Detailed balance condition holds! We can use it to sample.

- Given transition matrix $\boldsymbol{Q}$, how to make Markov chain transit with $\boldsymbol{P}$?

Markov Chain Monte Carlo

- Given transition matrix $\boldsymbol{Q}$, how to make Markov chain transit with $\boldsymbol{P}$?

- Rejection Sampling!
  - Sample $\boldsymbol{x}_* \sim q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$.
    - Suppose $\boldsymbol{x}_{t-1} = \boldsymbol{x}_i, \boldsymbol{x}_* = \boldsymbol{x}_j$.
  - Sample $u \sim U[0,1]$.
  - If $u > \alpha_{ij}$, reject the transition: $\boldsymbol{x}_t = \boldsymbol{x}_{t-1}$.
  - Else, accept: $\boldsymbol{x}_t = \boldsymbol{x}_*$.



- Assume convergence at time T_0
- Then the sample set $\{\boldsymbol{x}_{T_0}, \ldots, \boldsymbol{x}_T\}$ follows stationary distribution $\boldsymbol{\pi}$.

Outline

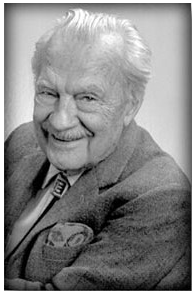
- **Sampling Method**
  - Monte Carlo
  - Markov Chain Monte Carlo
    - **Metropolis-Hastings Algorithm**
    - Gibbs Sampling
- Sampling Method for LDA
  - Bayesian Estimation
    - Dirichlet-Multinomial Conjugate
  - Gibbs Sampling for LDA

Metropolis-Hastings Algorithm

- Acceptance rate α_{ij}: the larger, the more efficient.
- Recall the detailed balance condition: $\pi_i Q_{ij} \alpha_{ij} = \pi_j Q_{ji} \alpha_{ji}$
- A natural choice: $\alpha_{ij} = \pi_j Q_{ji}$ and $\alpha_{ji} = \pi_i Q_{ij}$
  - Then we have: $\pi_i Q_{ij} \alpha_{ij} = \pi_i Q_{ij} \pi_j Q_{ji} = \pi_j Q_{ji} \alpha_{ji}$.
- Two problems:
  - First, π_j is not computable. Second, $\pi_j Q_{ji}$ is too small, very slow.
- Final choice: $\alpha_{ij} = \min(1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}})$. Big enough to accept more.
  - Compute α_{ij} only needs $\frac{\pi_j}{\pi_i} = \frac{b_j}{b_i}$. No need of intractable B.
  - $\pi_i Q_{ij} \alpha_{ij} = \min(\pi_i Q_{ij}, \pi_j Q_{ji}) = \pi_j Q_{ji} \alpha_{ji}$. Detailed balance holds.

Metropolis-Hastings Algorithm

- Metropolis-Hastings Algorithm
  - Sample $\boldsymbol{x}_* \sim q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$.
    - Suppose $\boldsymbol{x}_{t-1} = \boldsymbol{x}_i, \boldsymbol{x}_* = \boldsymbol{x}_j$.
  - Sample $u \sim U[0,1]$.
  - If $u > \alpha_{ij} = \min\left(1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\right)$, reject the transition: $\boldsymbol{x}_t = \boldsymbol{x}_{t-1}$.
  - Else, accept: $\boldsymbol{x}_t = \boldsymbol{x}_*$.



Nicholas Metropolis

Wilfred Keith Hastings

- Solves the problem of sampling from any distribution for Monte Carlo!
- Chosen as one of the top 10 algorithms in the 20th century.

Outline

- **Sampling Method**
  - Monte Carlo
  - Markov Chain Monte Carlo
    - Metropolis-Hastings Algorithm
    - **Gibbs Sampling**
- Sampling Method for LDA
  - Bayesian Estimation
    - Dirichlet-Multinomial Conjugate
  - Gibbs Sampling for LDA

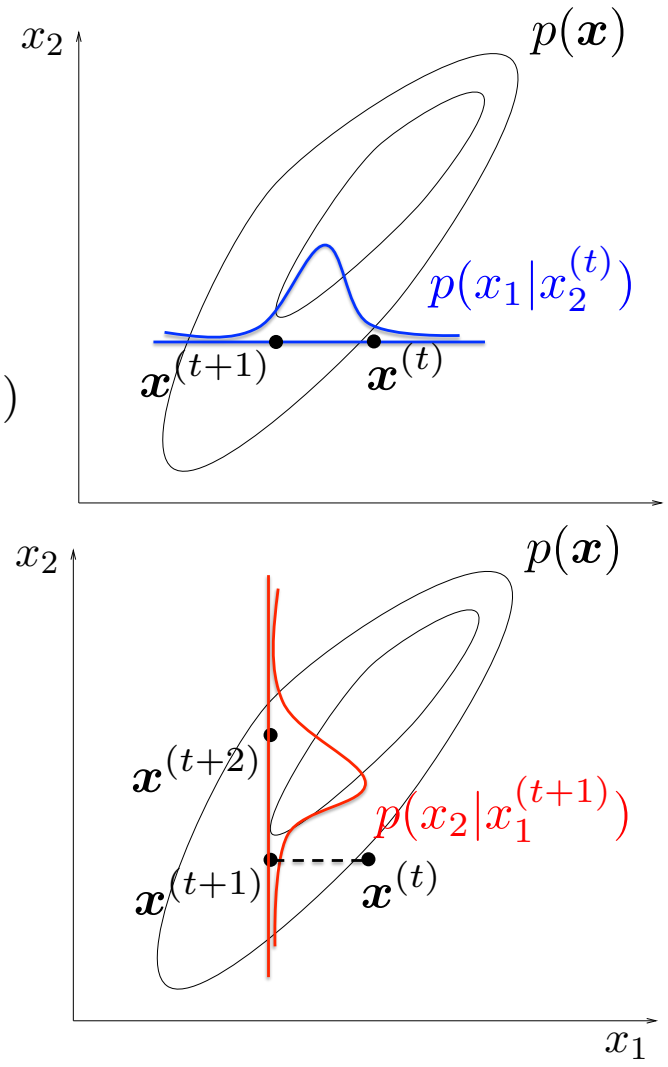
Gibbs Sampling

- Problems with Metropolis-Hastings Algorithm
  - Computing $\alpha_{ij} = \min\left(1, \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}\right)$ is inefficient in high-dimensions
  - $\alpha_{ij} \leq 1$, so rejection rate may high.
    - Can we not reject?
  - Sometimes it is impossible to compute the joint distribution in high-dimensions
    - Can we sample only based on the conditional distributions between different dimensions?
- The answer: Gibbs sampling. The algorithm for the big data era!

Gibbs Sampling

1. $x_1^{(t+1)} \sim p(x_1 | x_2^{(t)}, x_3^{(t)}, \cdots, x_n^{(t)})$
2. $x_2^{(t+1)} \sim p(x_2 | x_1^{(t+1)}, x_3^{(t)}, \cdots, x_n^{(t)})$
3. $\cdots$
4. $x_j^{(t+1)} \sim p(x_j | x_1^{(t+1)}, \cdots, x_{j-1}^{(t+1)}, x_{j+1}^{(t)}, \cdots, x_n^{(t)})$
5. $\cdots$
6. $x_n^{(t+1)} \sim p(x_n | x_1^{(t+1)}, x_2^t, \cdots, x_{n-1}^{(t+1)})$

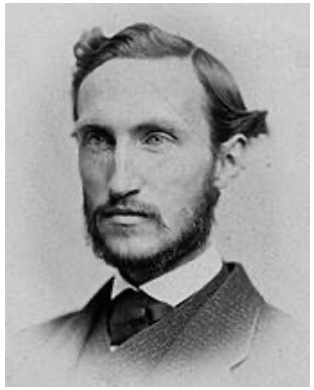
Sample Fix the rest $n-1$ var.



- Iterate across dimensions until convergence.
- One of the most beautiful algorithms in 20th century.

Gibbs Sampling

- Gibbs Sampling is a special case of Metropolis-Hastings Algorithm
  - The proposal distributions are posterior conditionals (computable!)
  - Accept all samples $\alpha_{ij} = 1$
- Proof:
  - We sample from the conditional $p(x_k^*|\boldsymbol{x}_{\neg k})$
    - Note that in each iteration we fix $\boldsymbol{x}_{\neg k}^* = \boldsymbol{x}_{\neg k}$
  - Transition probability $q(\boldsymbol{x}^*|\boldsymbol{x}) = p(x_k^*|\boldsymbol{x}_{\neg k})$
    - Joint distribution $p(\boldsymbol{x}) = p(\boldsymbol{x}_{\neg k})p(x_k|\boldsymbol{x}_{\neg k})$
  - Accept: $\alpha_k = \frac{\pi_j Q_{ji}}{\pi_i Q_{ij}} = \frac{p(\boldsymbol{x}^*)q(\boldsymbol{x}|\boldsymbol{x}^*)}{p(\boldsymbol{x})q(\boldsymbol{x}^*|\boldsymbol{x})} = \frac{p(\boldsymbol{x}_{\neg k}^*)p(x_k^*|\boldsymbol{x}_{\neg k}^*)p(x_k|\boldsymbol{x}_{\neg k}^*)}{p(\boldsymbol{x}_{\neg k})p(x_k|\boldsymbol{x}_{\neg k})p(x_k^*|\boldsymbol{x}_{\neg k})} = 1$



Josiah W. Gibbs

Outline

- Sampling Method
  - Monte Carlo
  - Markov Chain Monte Carlo
    - Metropolis-Hastings Algorithm
    - Gibbs Sampling
- **Sampling Method for LDA**
  - **Bayesian Estimation**
    - Dirichlet-Multinomial Conjugate
  - Gibbs Sampling for LDA

Bayes Rule

- Bayes rule:

Observed Data Parameter

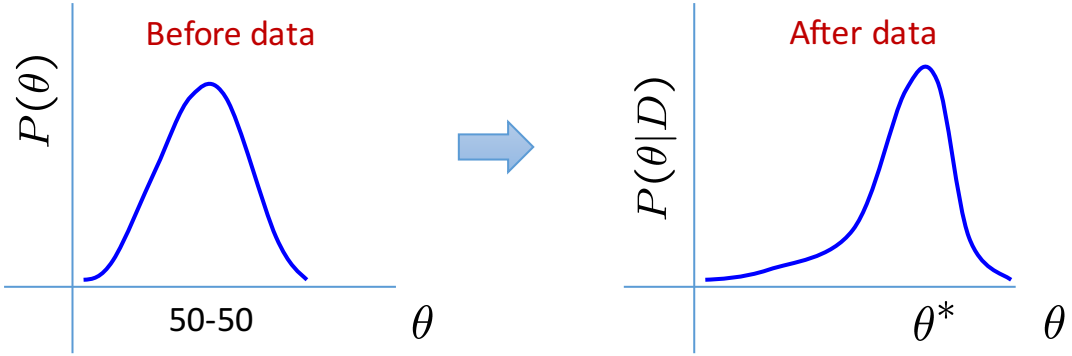
$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- Or equivalently:

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

Posterior Likelihood Prior

Thomas Bayes

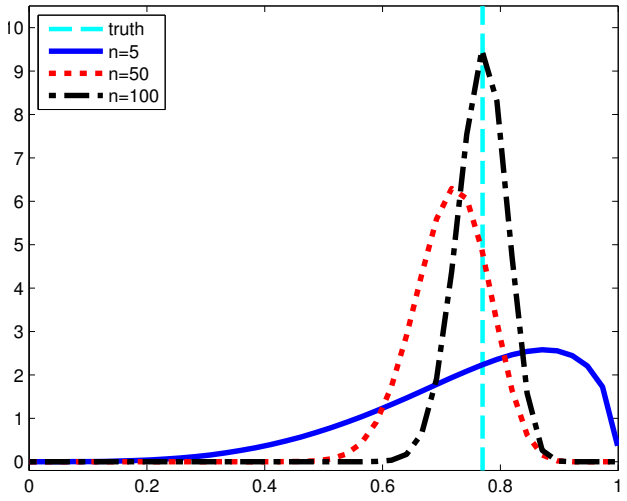


Maximum A Posteriori Estimation

$$\hat{\theta} = \arg\max_{\theta} \log p(\theta|D) = \arg\max_{\theta} \{\log p(D|\theta) + \log p(\theta)\}$$

Posterior — Likelihood — Prior

- MAP: Maximum a posteriori estimation of parameters θ
- We can view MLE as MAP
  - with a uniform prior distribution.
- As amount of data becomes large, posterior variance becomes small, and MAP behaves like other estimators such as MLE.

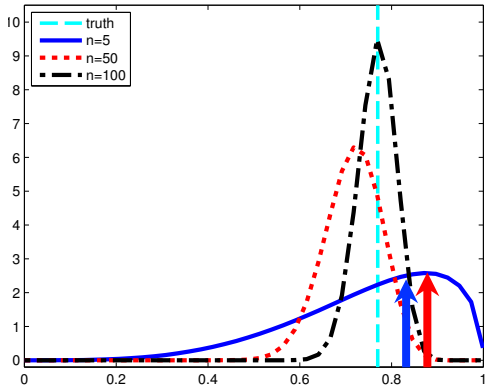


Bayes Estimator

$$\hat{\theta} = \mathbb{E}_{\theta \sim p(\theta|D)}[\theta] = \int \theta \cdot p(\theta|D) \mathrm{d}\theta$$

Mean — Posterior

- Bayes Estimator of parameters θ
  - Posterior $p(\theta|D)$ is a distribution of parameter.
  - The mean of this distribution is a stable solution.
- The solutions to MAP and Bayes Estimator are not distant when distribution is unimodal.
  - Choose one by mathematical convenience.

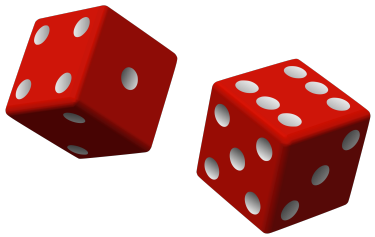


Outline

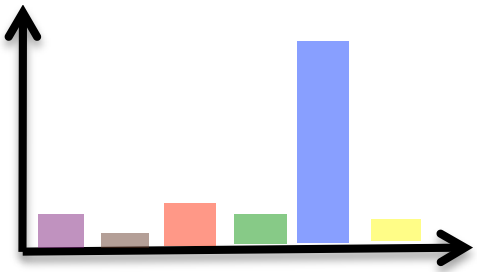
- Sampling Method
  - Monte Carlo
  - Markov Chain Monte Carlo
    - Metropolis-Hastings Algorithm
    - Gibbs Sampling
- **Sampling Method for LDA**
  - Bayesian Estimation
    - **Dirichlet-Multinomial Conjugate**
  - Gibbs Sampling for LDA

Multinomial Distribution

- Multinomial Distribution models the probability of counts of each side for rolling a k-sided dice n times.
  - Each side equips with a probability θ_i.
    - $\sum_{i=1}^{k} \theta_i = 1$.
  - Write by $\mathbf{Mult}(n, \boldsymbol{\theta})$
  - $p(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! \dots n_k!} \theta_1^{n_1} \theta_2^{n_2} \dots \theta_k^{n_k}$
    - $\sum_{i=1}^{k} n_i = n$
- When $n = 1$, it is Categorical Distribution, $\mathbf{Cat}(\boldsymbol{\theta}) = \mathbf{Mult}(1, \boldsymbol{\theta})$



Flipping dice:
$\sum_{l=1}^{6} \theta_l = 1$

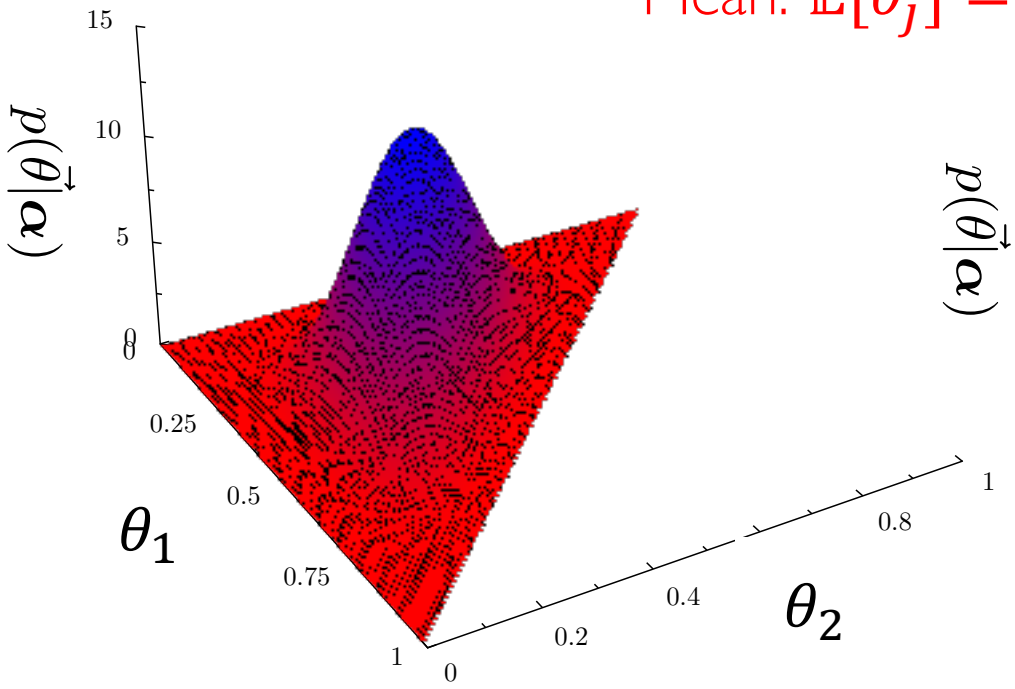


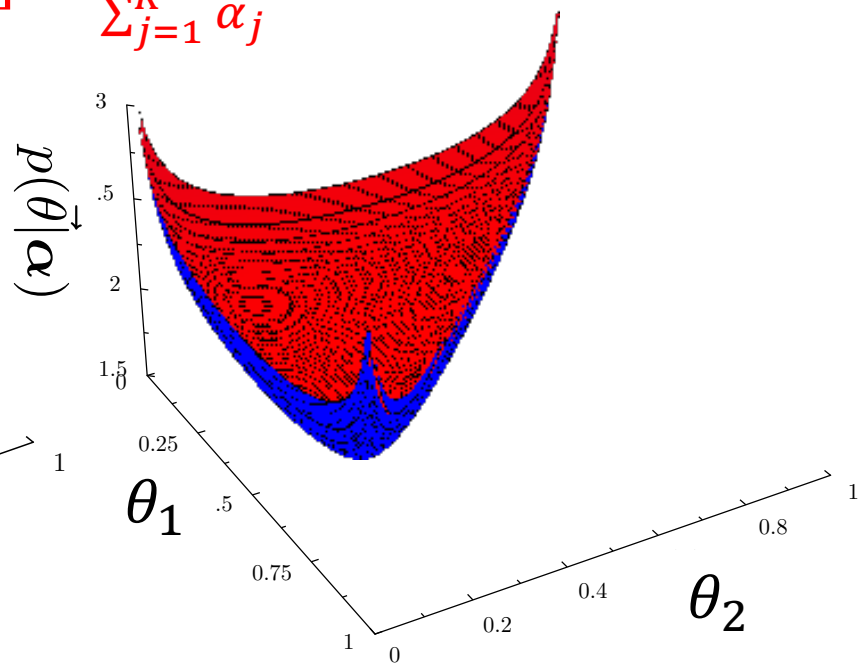
Dirichlet-Multinomial Model

- Dirichlet Distribution: Multi-dimensional version of Beta Distribution

$$p(\vec{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} \qquad \text{Where } B(\alpha) = \frac{\Pi_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$$

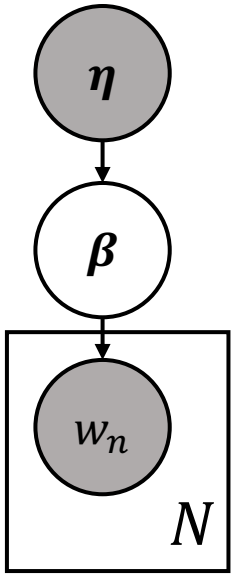
Mean: $\mathbb{E}[\theta_j] = \frac{\alpha_j}{\sum_{j=1}^{K}\alpha_j}$





Dirichlet-Multinomial Model

- Consider a simple document generation setting.
  - $\boldsymbol{\beta} \sim \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta})$, dimension of $\boldsymbol{\beta}$ is V and $w \sim \mathrm{Mult}(1, \boldsymbol{\beta})$.
- We can observe N words.
  - The count of term i is n_i. $(1 \le i \le V, \sum_{i=1}^{V} n_i = N)$



the	he	she
w_1	w_2	w_3

the	and	the
w_4	w_5	w_6

the	she	she	is
w_7	w_8	w_9	w_{10}

No.	Term	Count
1	the	4
2	he	1
3	she	3
4	and	1
5	…	

Posterior Distribution

- Compute the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\mathcal{W}})$
  - Where $\boldsymbol{\mathcal{W}}$ is the whole corpus.
- So that is to compute:

$$p(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\mathcal{W}}) = \frac{p(\boldsymbol{\mathcal{W}}|\boldsymbol{\beta}, \boldsymbol{\eta})p(\boldsymbol{\beta}|\boldsymbol{\eta})}{p(\boldsymbol{\mathcal{W}}|\boldsymbol{\eta})} = \frac{p(\boldsymbol{\mathcal{W}}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\eta})}{p(\boldsymbol{\mathcal{W}}|\boldsymbol{\eta})}$$

$$\propto p(\boldsymbol{\mathcal{W}}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{\eta}) = \prod_{i=1}^{N} p(w_i|\boldsymbol{\beta})\, p(\boldsymbol{\beta}|\boldsymbol{\eta}) = \prod_{i=1}^{N} \beta_{w_i} \frac{1}{B(\boldsymbol{\eta})} \prod_{j=1}^{V} \beta_j^{\eta_j - 1}$$

$$\prod_{i=1}^{N} \beta_{w_i} = \prod_{j=1}^{V} \beta_j^{\sum_{i=1}^{N} \mathbf{1}\{w_i = j\}} = \prod_{j=1}^{V} \beta_j^{n_j}$$

n_j: Count of j^{th} term

Posterior Distribution

- Finally, we have:

$$p(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\mathcal{W}}) \propto \prod_{i=1}^{N} \beta_{w_i} \frac{1}{B(\boldsymbol{\eta})} \prod_{j=1}^{V} \beta_j^{\eta_j - 1} = \prod_{j=1}^{V} \beta_j^{n_j} \frac{1}{B(\boldsymbol{\eta})} \prod_{j=1}^{V} \beta_j^{\eta_j - 1}$$

$$\propto \prod_{j=1}^{V} \beta_j^{\eta_j + n_j - 1}$$

- Recall that $\mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta}) = \frac{1}{B(\boldsymbol{\eta})} \prod_{j=1}^{V} \beta_j^{\eta_j - 1} \propto \prod_{j=1}^{V} \beta_j^{\eta_j - 1}$
- We find that $p(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\mathcal{W}})$ is also a Dirichlet Distribution:

$$p(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\mathcal{W}}) = \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n})$$

Conjugate!

Dirichlet-Multinomial Model

- The Dirichlet is conjugate to the Multinomial:

$\boldsymbol{\beta} \sim \mathrm{Dir}(\boldsymbol{\eta})$ [draw distribution over words]

For each word $n \in \{1, \ldots, N\}$

$w_n \sim \mathrm{Mult}(1, \boldsymbol{\beta})$ [draw word]

- The prior distribution of $\boldsymbol{\beta}$ is a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta})$.
- Observe some samples from Multinomial parameterized by $\boldsymbol{\beta}$.
- Then the posterior is also a Dirichlet distribution:

$$p(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{w}) \sim \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n})$$

  - The difference depends on term counts $\boldsymbol{n}$.

Conjugate!

- Will see conjugate distribution brings great convenience to inference.

MAP vs. Bayes Estimator

- Now we get our estimation $p(\boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{\mathcal{W}}) \sim \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n})$.
- Maximum a Posterior:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{MAP}} = \mathrm{argmax}_{\boldsymbol{\beta}} \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n})$$

$$(\widehat{\boldsymbol{\beta}}_{\mathrm{MAP}})_i = \frac{\eta_i + n_i - 1}{\sum_{j=1}^{V}(\eta_j + n_j - 1)} = \frac{\eta_i + n_i - 1}{\sum_{j=1}^{V}\eta_j + N - V}$$

- Bayes Estimator:

May<0

$$\widehat{\boldsymbol{\beta}}_{\mathrm{Bayes}} = \int \boldsymbol{\beta} \cdot \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n}) d\boldsymbol{\beta}$$

$$(\widehat{\boldsymbol{\beta}}_{\mathrm{Bayes}})_i = \frac{\eta_i + n_i}{\sum_{j=1}^{V}(\eta_j + n_j)} = \frac{\eta_i + n_i}{\sum_{j=1}^{V}\eta_j + N}$$

Outline

- Sampling Method
  - Monte Carlo
  - Markov Chain Monte Carlo
    - Metropolis-Hastings Algorithm
    - Gibbs Sampling
- **Sampling Method for LDA**
  - Bayesian Estimation
    - Dirichlet-Multinomial Conjugate
  - **Gibbs Sampling for LDA**

LDA as Latent Variable Model

- Two sets of random variables: z and x
  - z consists of unobserved hidden variables.
  - x consists of observed variables.
- Joint probability model parameterized by $\theta \in \Theta$:

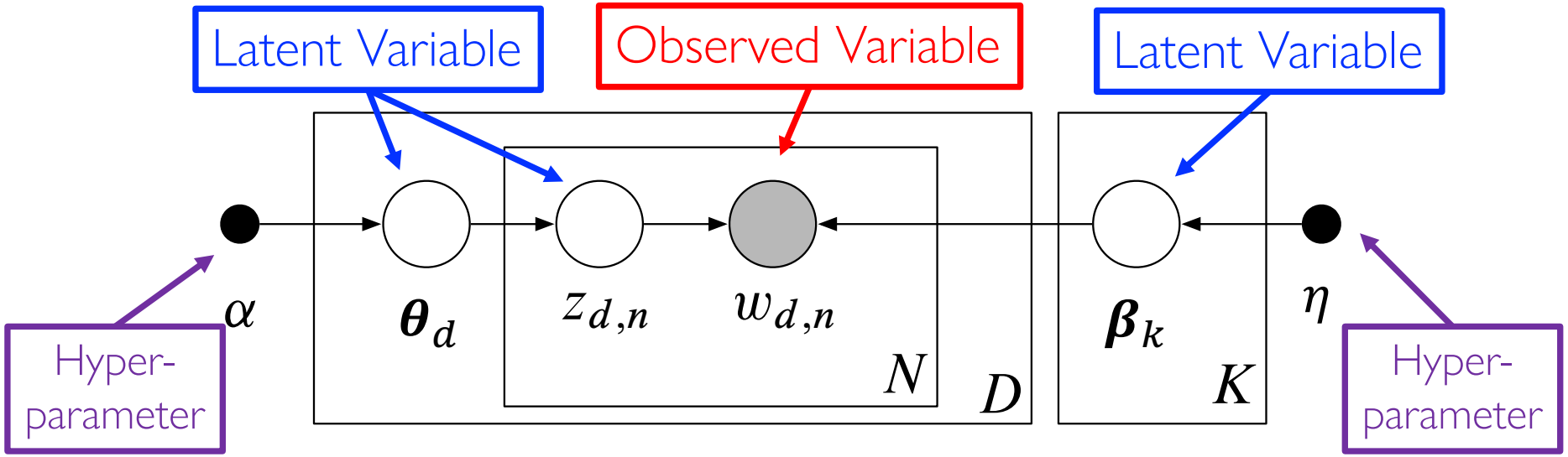
$$p(x, z|\theta)$$


Colloquium

Finding scientific topics

Thomas L. Griffiths*†‡ and Mark Steyvers§

Griffiths *et al.* 2004



Gibbs Sampling for LDA

- We need to solve:

$$\text{In LDA: } p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\beta} \mid \boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}$$

  - But the denominator is intractable!

- With Gibbs Sampling, we can sample from it!

- Shall we sample each vector in order?

  - $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_D, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K, z_{11}, z_{12}, \ldots, z_{dn}$.

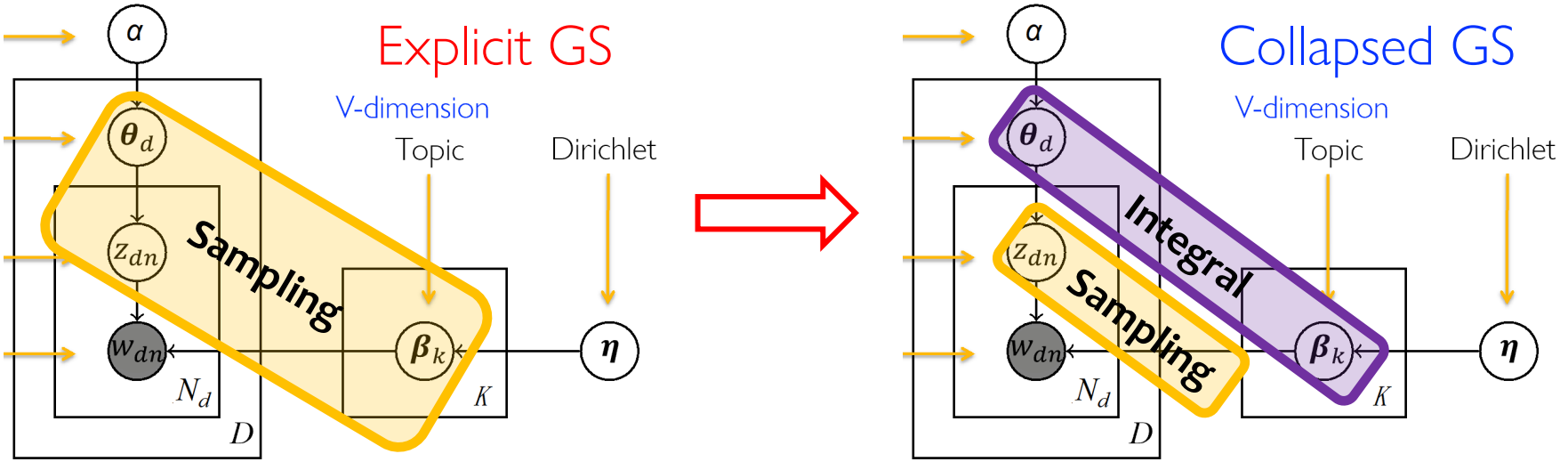
- No! We only sample $\theta_{11}, \theta_{12}, \ldots, \theta_{1K}$ in order:

  - Each dimension cannot be easily disentangled in Dirichlet variable!

Gregor Heinrich. Parameter estimation for text analysis. Tech. Report.

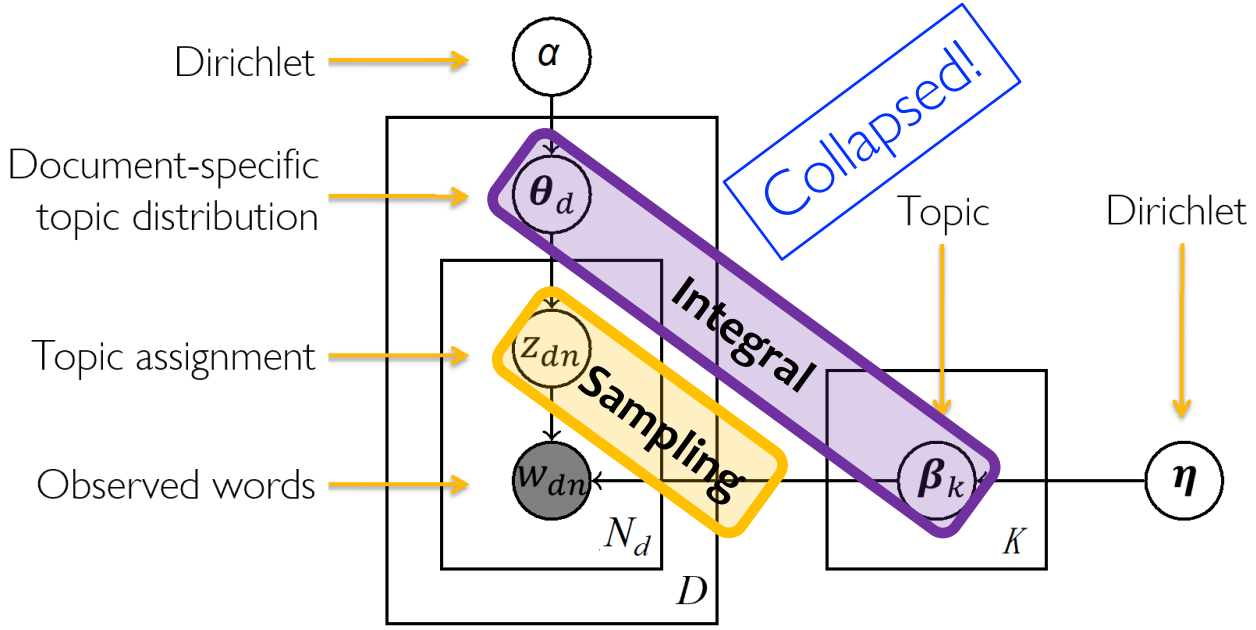
Gibbs Sampling for LDA

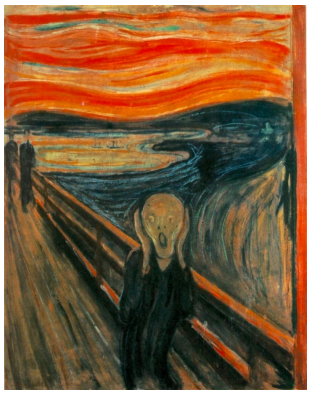
- Sampling from a high-dimension distribution is very difficult.
- We need an approximation here:
  - Do not sample, but just compute mean of Posterior distribution.
- The the question changes to: If we know all z_{dn} and w_{dn} (sampling!)
  - What are $p(\theta_d|z, w, \beta, \alpha, \eta)$, $p(\beta_k|z, w, \theta, \alpha, \eta)$ and their mean.



Collapsed Gibbs Sampling for LDA

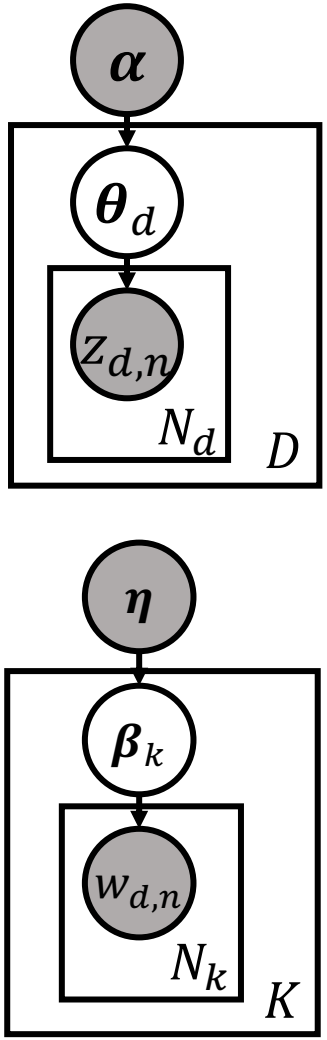
- Iterate between sampling and integral:
  1. Sample latent topic $z_{11}, z_{12}, \ldots, z_{DN_d}$: $z_{ij} \sim p(z|\boldsymbol{z}^{-ij}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\beta})$
  2. Compute mean of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_D$: $\widehat{\boldsymbol{\theta}}_d = \mathbb{E}_{\boldsymbol{\theta}_d \sim p(\boldsymbol{\theta}_d|\boldsymbol{z}, \boldsymbol{\alpha})}[\boldsymbol{\theta}_d]$
  3. Compute mean of $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K$: $\widehat{\boldsymbol{\beta}}_k = \mathbb{E}_{\boldsymbol{\beta}_k \sim p(\boldsymbol{\beta}_k|\boldsymbol{z}, \boldsymbol{w}, \boldsymbol{\eta})}[\boldsymbol{\beta}_k]$





Terminology

- Component $\boldsymbol{\alpha}$:
  - N_d ← count of all words in document d
  - t_{dk} ← count of words in topic k in document d
  - Count vector: $\boldsymbol{t}_d = (t_{d1}, \ldots, t_{dK})$
  - Conjugate in LDA: $p(\boldsymbol{\theta}_d = \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{z}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha} + \boldsymbol{t}_d)$
- Component $\boldsymbol{\eta}$:
  - N_k ← count of all words in topic k
  - n_{kv} ← count of words of term v in topic k
  - Count vector: $\boldsymbol{n}_k = (n_{k1}, \ldots, n_{kV})$
  - Conjugate in LDA: $p(\boldsymbol{\beta}_k = \boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{w}, \boldsymbol{z}) = \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n}_k)$



Sampling from $p(z_i = k|\vec{\boldsymbol{z}}_{\neg i}, \vec{\boldsymbol{w}})$

- Gibbs sampling on latent topic from posterior: $p(z_i = k|\vec{\boldsymbol{z}}_{\neg i}, \vec{\boldsymbol{w}})$
- **Problem:** $\boldsymbol{w_i}$ is in the condition.
- We eliminate it with the Bayesian rule:

$$p\left(z_i = k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}\right) = p\left(z_i = k|w_i = v, \overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}\right) = \frac{p\left(z_i = k, w_i = v|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right)}{p\left(w_i = v|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right)}$$

  - The denominator is irrelevant to $\boldsymbol{w_i}$:

$$p\left(w_i = v|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right)$$

- We have:

$$p\left(z_i = k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}\right) \propto p\left(z_i = k, w_i = v|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right)$$

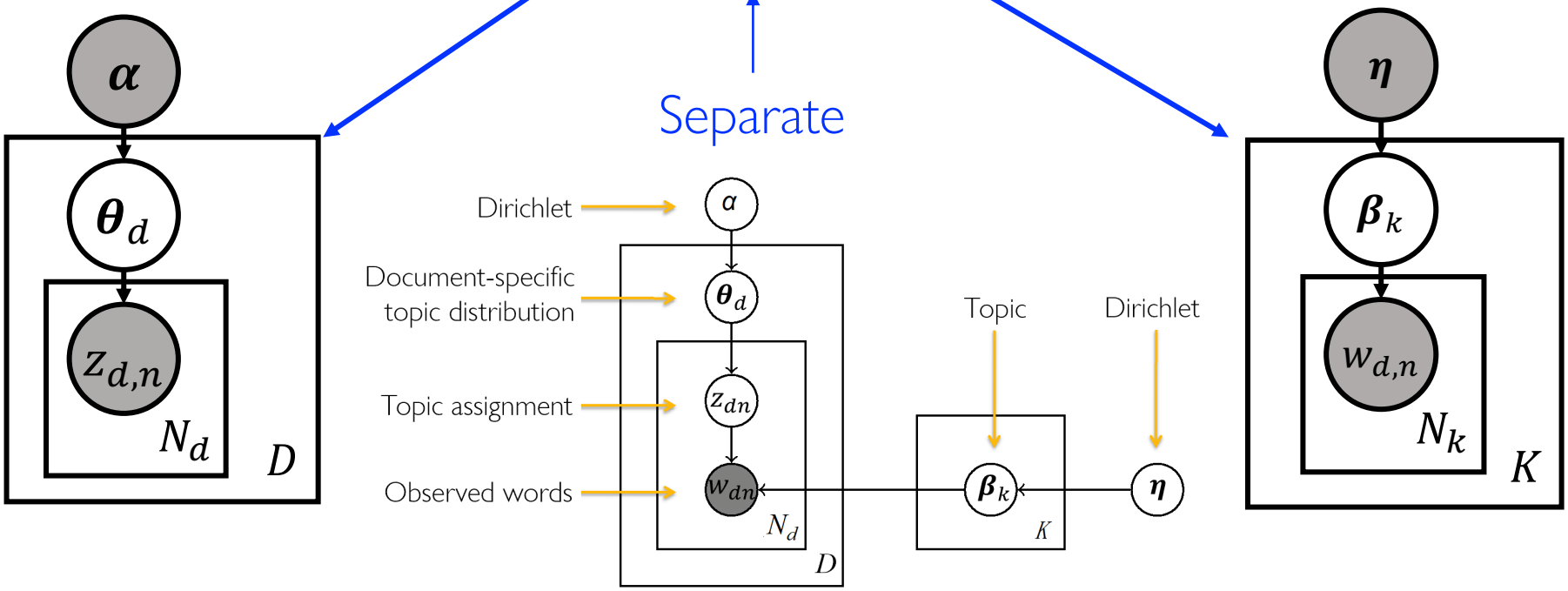
  - There is no $\boldsymbol{w_i}$ in the conditional distribution. Easy to sample!

Sampling from $p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\boldsymbol{w}})$

$$p\left(z_i = k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}\right) \propto p\left(z_i = k, w_i = v|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right)$$

$$= \int p\left(z_i = k, w_i = v, \vec{\boldsymbol{\theta}}_d, \vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k$$

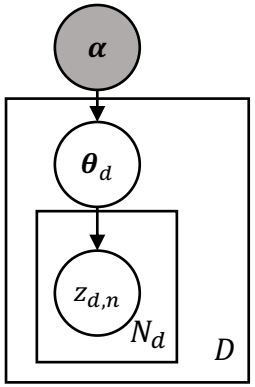
$$= \int p\left(z_i = k, \vec{\boldsymbol{\theta}}_d|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \cdot p\left(w_i = v, \vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k$$



Gregor Heinrich. Parameter estimation for text analysis.

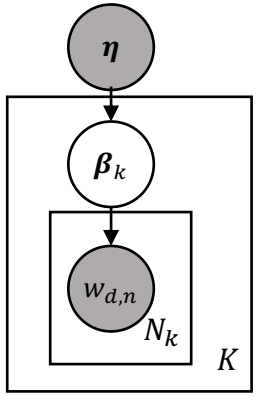
Sampling from $p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\boldsymbol{w}})$

$$
\begin{aligned}
p\left(z_i = k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}\right) &\propto p\left(z_i = k, w_i = v|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \\
&= \int p\left(z_i = k, w_i = v, \vec{\boldsymbol{\theta}}_d, \vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k, \vec{\boldsymbol{\theta}}_d|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \cdot p\left(w_i = v, \vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k|\vec{\boldsymbol{\theta}}_d\right) p\left(\vec{\boldsymbol{\theta}}_d|\overrightarrow{\mathbf{z}}_{\rightarrow i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \cdot p\left(w_i = v|\vec{\boldsymbol{\beta}}_k\right) p\left(\vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k|\vec{\boldsymbol{\theta}}_d\right) \mathrm{Dir}\left(\vec{\boldsymbol{\theta}}_d|\vec{\boldsymbol{t}}_{d,\neg i} + \vec{\boldsymbol{\alpha}}\right) d\vec{\boldsymbol{\theta}}_d \cdot \int p\left(w_i = v|\vec{\boldsymbol{\beta}}_k\right) \mathrm{Dir}\left(\vec{\boldsymbol{\beta}}_k|\vec{\boldsymbol{n}}_{k,\neg i} + \vec{\boldsymbol{\eta}}\right) d\vec{\boldsymbol{\beta}}_k
\end{aligned}
$$



$$\longrightarrow \quad p(\boldsymbol{\theta}_d = \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{z}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha} + \boldsymbol{t}_d)$$

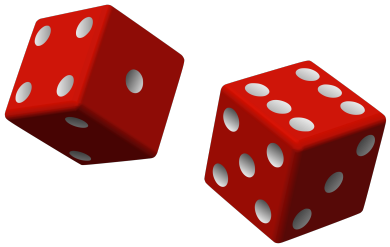
$$p(\boldsymbol{\beta}_k = \boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{w}, \boldsymbol{z}) = \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n}_k) \quad \longleftarrow$$

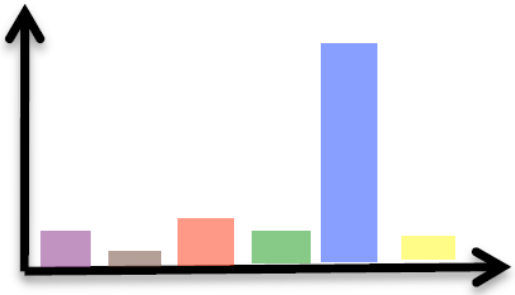


Gregor Heinrich. Parameter estimation for text analysis.

Sampling from $p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\boldsymbol{w}})$

$$
\begin{aligned}
p\left(z_i = k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}\right) &\propto p\left(z_i = k, w_i = v|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \\
&= \int p\left(z_i = k, w_i = v, \vec{\boldsymbol{\theta}}_d, \vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k, \vec{\boldsymbol{\theta}}_d|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \cdot p\left(w_i = v, \vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k|\vec{\boldsymbol{\theta}}_d\right) p\left(\vec{\boldsymbol{\theta}}_d|\overrightarrow{\mathbf{z}}_{\rightarrow i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \cdot p\left(w_i = v|\vec{\boldsymbol{\beta}}_k\right) p\left(\vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k|\vec{\boldsymbol{\theta}}_d\right) \mathrm{Dir}\left(\vec{\boldsymbol{\theta}}_d|\vec{\boldsymbol{t}}_{d,\neg i} + \vec{\boldsymbol{\alpha}}\right) d\vec{\boldsymbol{\theta}}_d \cdot \int p\left(w_i = v|\vec{\boldsymbol{\beta}}_k\right) \mathrm{Dir}\left(\vec{\boldsymbol{\beta}}_k|\vec{\boldsymbol{n}}_{k,\neg i} + \vec{\boldsymbol{\eta}}\right) d\vec{\boldsymbol{\beta}}_k \\
&= \int \theta_{dk} \mathrm{Dir}\left(\vec{\boldsymbol{\theta}}_d|\vec{\boldsymbol{t}}_{d,\neg i} + \vec{\boldsymbol{\alpha}}\right) d\vec{\boldsymbol{\theta}}_d \cdot \int \beta_{kv} \mathrm{Dir}\left(\vec{\boldsymbol{\beta}}_k|\vec{\boldsymbol{n}}_{k,\neg i} + \vec{\boldsymbol{\eta}}\right) d\vec{\boldsymbol{\beta}}_k \\
&= \mathbb{E}_{\neg i}\left(\theta_{dk}\right) \cdot \mathbb{E}_{\neg i}\left(\beta_{kv}\right) \\
&= \hat{\theta}_{dk} \cdot \hat{\beta}_{kv}
\end{aligned}
$$





Gregor Heinrich. Parameter estimation for text analysis.

Sampling from $p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\boldsymbol{w}})$

$$
\begin{aligned}
p\left(z_i = k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}\right) &\propto p\left(z_i = k, w_i = v|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \\
&= \int p\left(z_i = k, w_i = v, \vec{\boldsymbol{\theta}}_d, \vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k, \vec{\boldsymbol{\theta}}_d|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \cdot p\left(w_i = v, \vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k|\vec{\boldsymbol{\theta}}_d\right) p\left(\vec{\boldsymbol{\theta}}_d|\overrightarrow{\mathbf{z}}_{\rightarrow i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) \cdot p\left(w_i = v|\vec{\boldsymbol{\beta}}_k\right) p\left(\vec{\boldsymbol{\beta}}_k|\overrightarrow{\mathbf{z}}_{\neg i}, \overrightarrow{\mathbf{w}}_{\neg i}\right) d\vec{\boldsymbol{\theta}}_d d\vec{\boldsymbol{\beta}}_k \\
&= \int p\left(z_i = k|\vec{\boldsymbol{\theta}}_d\right) \mathrm{Dir}\left(\vec{\boldsymbol{\theta}}_d|\vec{\boldsymbol{t}}_{d,\neg i} + \vec{\boldsymbol{\alpha}}\right) d\vec{\boldsymbol{\theta}}_d \cdot \int p\left(w_i = v|\vec{\boldsymbol{\beta}}_k\right) \mathrm{Dir}\left(\vec{\boldsymbol{\beta}}_k|\vec{\boldsymbol{n}}_{k,\neg i} + \vec{\boldsymbol{\eta}}\right) d\vec{\boldsymbol{\beta}}_k \\
&= \int \theta_{dk} \mathrm{Dir}\left(\vec{\boldsymbol{\theta}}_d|\vec{\boldsymbol{t}}_{d,\neg i} + \vec{\boldsymbol{\alpha}}\right) d\vec{\boldsymbol{\theta}}_d \cdot \int \beta_{kv} \mathrm{Dir}\left(\vec{\boldsymbol{\beta}}_k|\vec{\boldsymbol{n}}_{k,\neg i} + \vec{\boldsymbol{\eta}}\right) d\vec{\boldsymbol{\beta}}_k \\
&= \mathbb{E}_{\neg i}\left(\theta_{dk}\right) \cdot \mathbb{E}_{\neg i}\left(\beta_{kv}\right) \\
&= \hat{\theta}_{dk} \cdot \hat{\beta}_{kv}
\end{aligned}
$$

Bayes Estimator

$$\hat{\theta}_{dk,\neg i} = \frac{t_{dk,\neg i} + \alpha}{N_{d,\neg i} + \alpha K} \qquad \hat{\beta}_{kv,\neg i} = \frac{n_{kv,\neg i} + \eta}{N_{k,\neg i} + \eta V}$$

Gregor Heinrich. Parameter estimation for text analysis.

Bayes Estimator for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$

- Document-topic distribution: $p(\boldsymbol{\theta}_d = \boldsymbol{\theta}|\boldsymbol{\alpha}, \boldsymbol{z}) = \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha} + \boldsymbol{t}_d)$

$$\widehat{\boldsymbol{\theta}}_d = \int \boldsymbol{\theta} \cdot \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha} + \boldsymbol{t}_d)\, \mathrm{d}\boldsymbol{\theta} \qquad \hat{\theta}_{dk} = \frac{t_{dk} + \alpha_k}{N_d + \sum_{k'=1}^{K} \alpha_{k'}}$$

- Term-topic distribution: $p(\boldsymbol{\beta}_k = \boldsymbol{\beta}|\boldsymbol{\eta}, \boldsymbol{w}, \boldsymbol{z}) = \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n}_k)$

$$\widehat{\boldsymbol{\beta}}_k = \int \boldsymbol{\beta} \cdot \mathrm{Dir}(\boldsymbol{\beta}|\boldsymbol{\eta} + \boldsymbol{n}_k)\, \mathrm{d}\boldsymbol{\beta} \qquad \hat{\beta}_{kv} = \frac{n_{kv} + \eta_v}{N_k + \sum_{v'=1}^{V} \eta_v}$$

- Griffiths *et al.* 2004 further set $\alpha_k = \alpha$, $\eta_v = \eta$ as all constants.

$$\hat{\theta}_{dk} = \frac{t_{dk} + \alpha}{N_d + \alpha K} \qquad \hat{\beta}_{kv} = \frac{n_{kv} + \eta}{N_k + \eta V}$$

LDA Inference with Gibbs Sampling

- Initialize all latent topics $\{z_{11}, z_{12}, \ldots, z_{DN_D}\}$.

- **Iterate:**
  - For each position ij, sample z_{ij} from posterior:

$$p(z_{ij} = k | \boldsymbol{z}_{\neg ij}, \boldsymbol{w}) \propto \frac{t_{dk,\neg ij} + \alpha}{N_{d,\neg ij} + \alpha K} \cdot \frac{n_{kv,\neg ij} + \eta}{N_{k,\neg ij} + \eta V}$$

  - Compute Bayes Estimator of parameters:

$$\hat{\theta}_{dk} = \frac{t_{dk} + \alpha}{N_d + \alpha K} \qquad \hat{\beta}_{kv} = \frac{n_{kv} + \eta}{N_k + \eta V}$$

Gregor Heinrich. Parameter estimation for text analysis.

Thank You

Questions?

Mingsheng Long
mingsheng@tsinghua.edu.cn
http://ise.thss.tsinghua.edu.cn/~mlong
答疑：东主楼11区413室