Latent Dirichlet Allocation

Huang Chen

2015/09/16

Data Mining Lab, Big Data Research Center,
School of Computer Science and Engineering,
UESTC

Outline

Brief Introduction

About Math

Devil's Game

LDA

Part One

Brief Introduction

Brief Introduction

How to measure the similarity of documents?

VS

- TFIDF + Cosine
- ……
- Topic Model

Topic Model

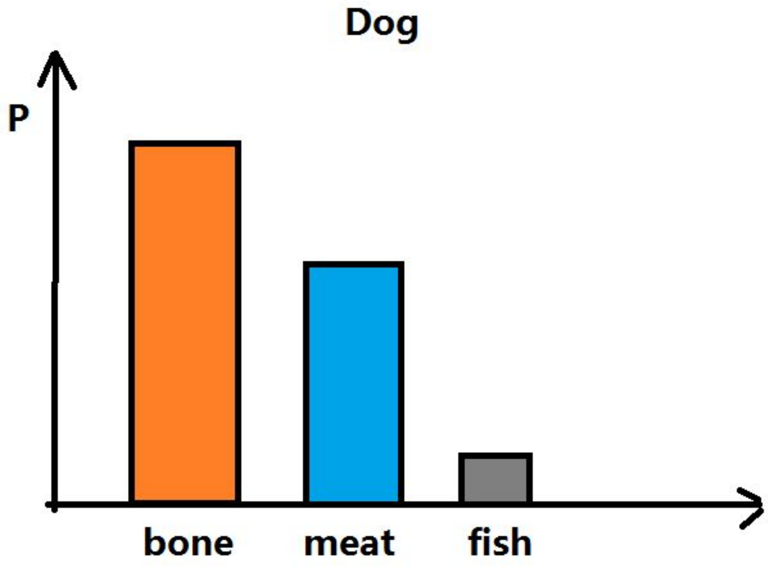
- A type of statistical model for discovering the abstract "**topics**" that occur in a collection of documents. (Wiki)

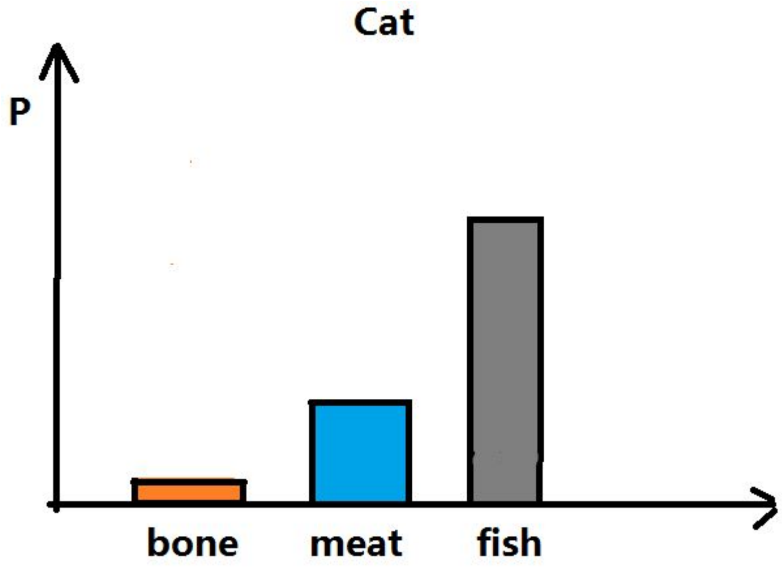
  - Particular words to appear in the document more or less frequently according to its topics
  - More "bones", less "fish" in the doc about dogs

What's topic?

Some words that strongly related to some topic, and them more likely to appear.

Probability distribution over words

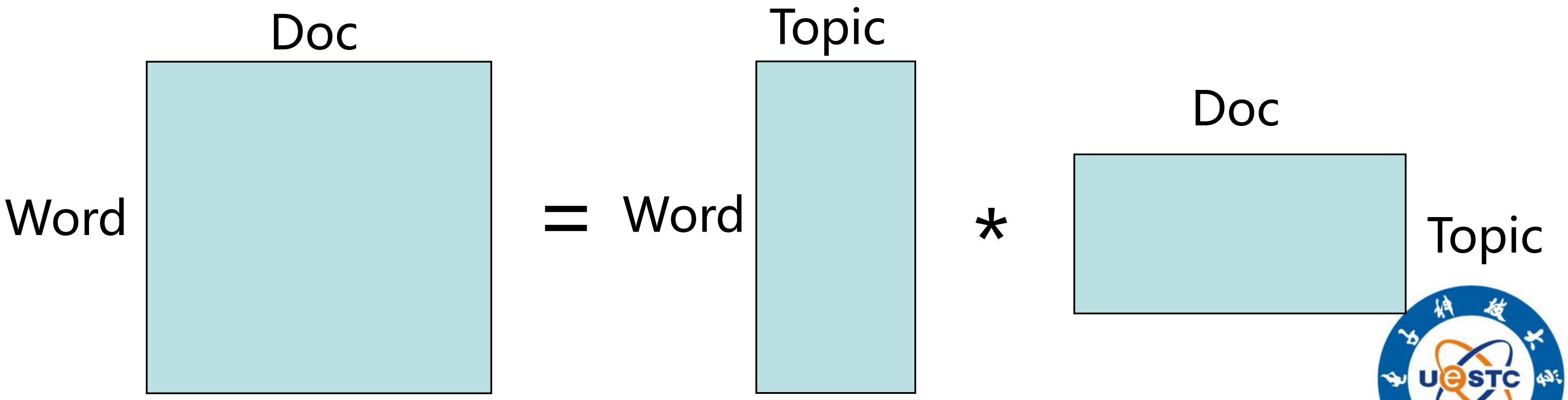




Topic Model

- **Key :**
- A word in a doc should be chosed with a certain rule:
- Choose one topic with a certain probability
- Choose a word from this topic with a certain probability

$$P(\mathrm{Word} \mid Doc) = \sum_{Topic} \mathrm{P}(\mathrm{Word} \mid Topic) * \mathrm{P}(Topic \mid Doc)$$



Part Two

Math

Gamma -> Beta -> Dirichlet

Gamma Function

- Let's say $\mathbf{x^n}$, As you know
- First derivative: $\mathbf{nx^{n-1}}$
- Second derivative: $\mathbf{n(n-1)x^{n-2}}$
- ……
- K-th derivative (n≥k):

$$\frac{n!}{(n-k)!}x^{n-k}$$

So What is the 1/2 order derivative of x ?

What is the 1/2 order derivative of x ?

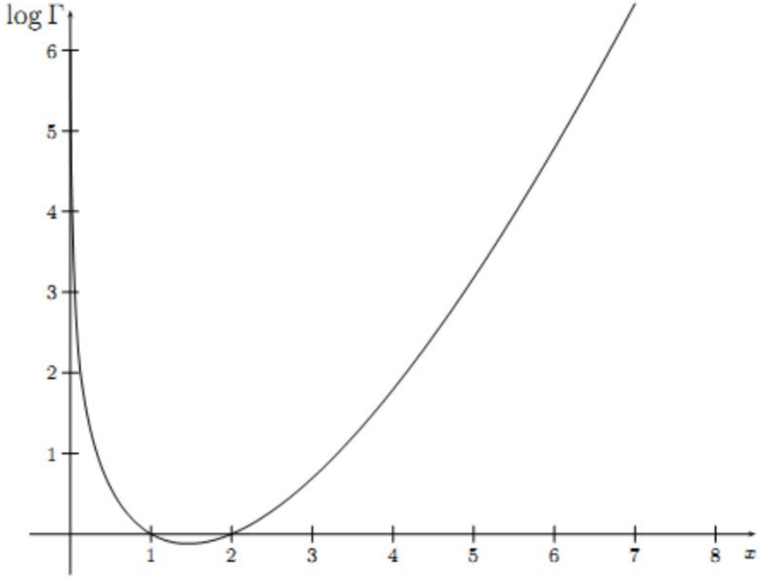
- **Gamma function**

$$\mathrm{T}(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$$

$$\mathrm{T}(n) = (n-1)!$$

- **The k-th derivative:**

$$\frac{\mathrm{T}(n+1)}{T(n-k+1)} x^{n-k}$$



- So when n = 1, k = 1/2:
- The 1/2 order derivative of x = $2\sqrt{\frac{\mathrm{x}}{\pi}}$

Devil's Game (1)

Suppose one day you are caught by a tricky devil, and he wants to play a game with you.

Devil has a button, once he push it, it will output a random number between 0 to 1. He pushed 10 times, and got 10 random numbers. (i.i.d.)

Random number

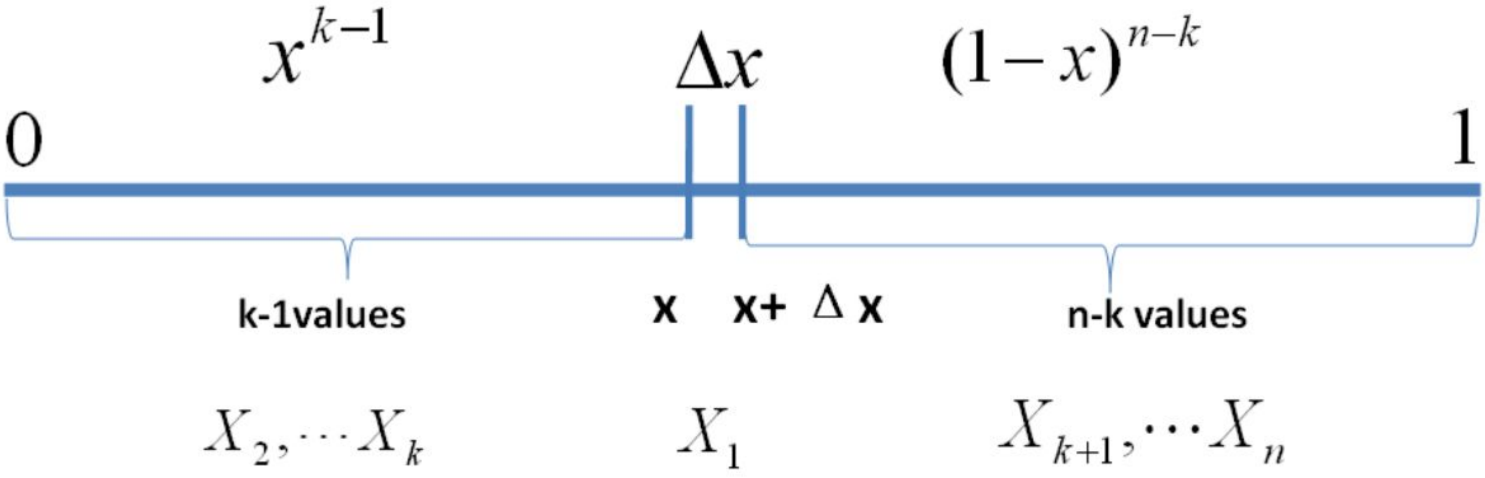
The question is "**What is the seventh big number of these ten ?**" Give your answer and error no more than ξ

The distribution of X(k)

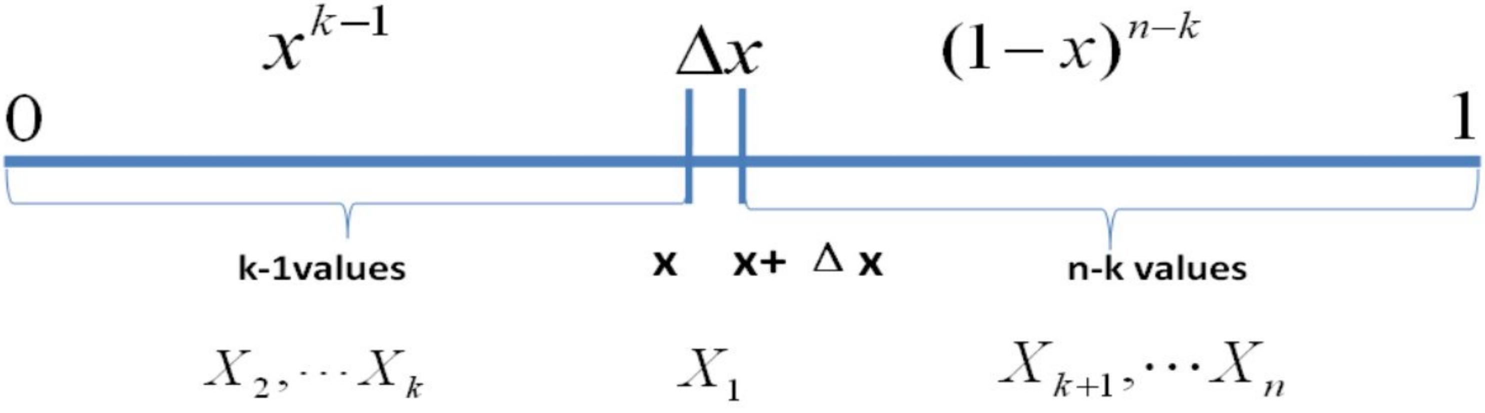
Devil's Game (1)

$$P(x \le X(k) \le x + \Delta x) = ?$$

- **Let's say X1 is located in this area:**



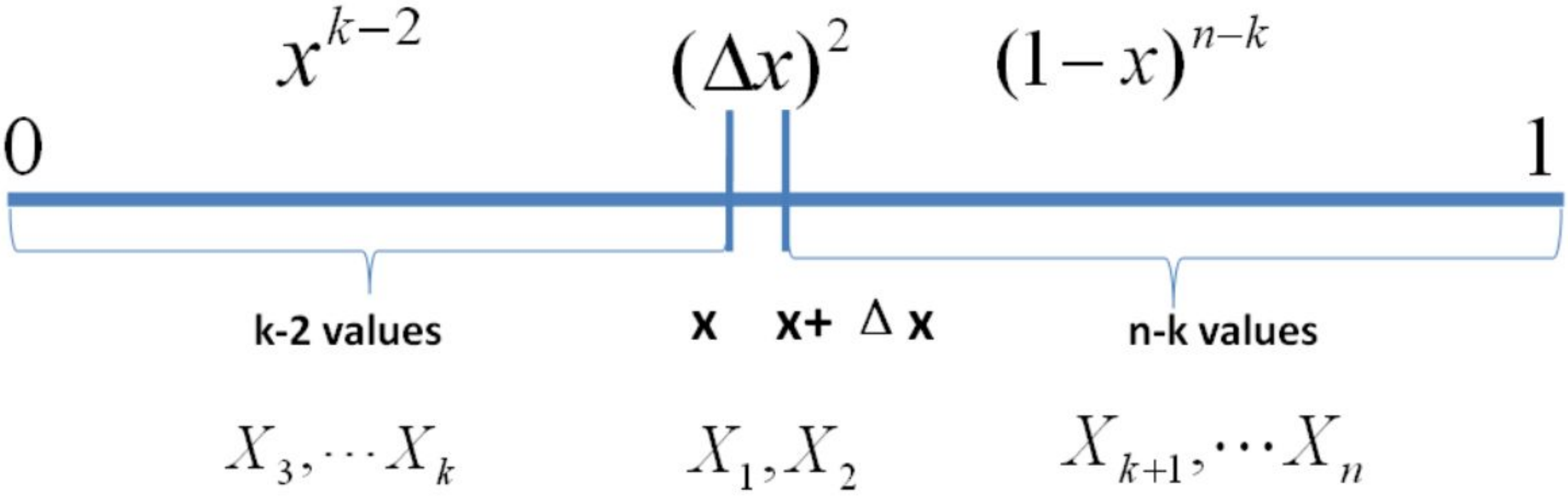
Devil's Game (1)



$$P(E) = \prod_{i=1}^{n} P(X_i) = x^{k-1}(1-x-\Delta x)^{n-k}\Delta x = x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x)$$

Equivalent event : $n\binom{n-1}{k-1}$

Devil's Game (1)



$$\mathrm{P(E')} = \prod_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{P(X_i)} = x^{k-2}(1-x-\Delta x)^{n-k}(\Delta x)^2 = o(\Delta x)$$

Devil's Game (1)

Thus...

$$
\begin{aligned}
&\mathrm{P}(x \leq X(k) \leq x + \Delta x) \\
&= n\binom{n-1}{k-1} P(E) + o(\Delta x) \\
&= n\binom{n-1}{k-1} \mathrm{x}^{k-1}(1-x)^{n-k}\,\Delta x + o(\Delta x)
\end{aligned}
$$

Devil's Game (1)

Thus pdf...

$$f(x)=\lim_{\Delta x\to 0}\frac{P(x\le X(k)\le x+\Delta x)}{\Delta x}$$

$$=n\binom{n-1}{k-1}x^{k-1}(1-x)^{n-k}$$

$$=\frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k}\qquad x\in[0,1]$$

$$=\frac{\mathrm{T}(n+1)}{T(k)T(n-k+1)}x^{k-1}(1-x)^{n-k}$$

$$=\frac{\mathrm{T}(\alpha+\beta)}{T(\alpha)T(\beta)}x^{\alpha-1}(1-x)^{\beta-1}\qquad \alpha=k,\beta=n-k+1$$

Beta

Devil's Game (1)

Okay! Give your answer !

Let n = 10, k = 7

$$f(x) = \frac{10!}{(6)!(3)!} x^6 (1-x)^3 \qquad x \in [0,1]$$

Devil's Game (2)

But you're wrong......bad luck

Devil shows his mercy.

He allows you to push the button 5 times and tells you which is bigger, compared with the seventh big number , among the 5 numbers you got

Devil's Game (2)

Target :

$$\mathrm{P}(X(k) \mid Y_1, Y_2, \cdots, Y_m) \qquad Y_1, Y_2, \cdots, Y_m \overset{\mathrm{ii}d}{\sim} uniform(0,1)$$

What we know :

- Prior knowledge

$$\mathrm{f}(X(k)) = Beta(X \mid k, n-k+1)$$

- Say m_1 numbers smaller than X(k), m_2 bigger

$$\mathrm{Y} \sim \mathrm{B}(m, X(k)) \quad \Longleftarrow$$

Binomial

- Posterior knowledge

$$f(X(k) \mid m_1, m_2) = Beta(X \mid k + m_1, n - k + 1 + m_2)$$

Devil's Game (2)

Thus ...

$$
\begin{aligned}
&Beta(p \mid k, n-k+1) + Binom\mathrm{Coun}t(m_1, m_2) \\
&= Beta(p \mid k+m_1, n-k+1+m_2) \\
&p = P(X(k))
\end{aligned}
$$

Beta-Binomial Conjugate

*Conjugate

If the posterior distributions $p(\theta|x)$ are in the **same family** as the prior probability distribution $p(\theta)$,

The prior and posterior are then called **conjugate distributions**

The prior is called a **conjugate prior** for the likelihood function.

Devil's Game (2)

If there are 2 numbers smaller than X(7) :

Give your answer !

$$\mathrm{Beta}(x\,|\,9,7)=\frac{15!}{(8)!(6)!}x^{8}(1-x)^{6}\qquad x\in[0,1]$$

Devil's Game (3)

Luckily , you're right

However, Devil wants play one more time

Push 20 times, What are the 7-th and 13-th big number?

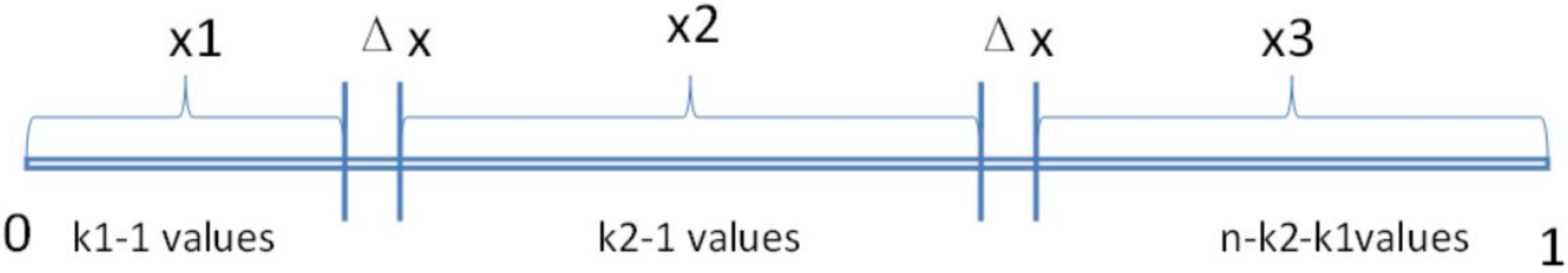
Devil's Game (3)

Target :

Joint distribution of (X(k$_1$), X(K$_1$+K$_2$))

Solution :

$$\mathrm{P}(E) = \frac{n!}{(k_1-1)!(k_2-1)!(n-k_1-k_2)!} x_1^{k_1-1} x_2^{k_2-1} x_3^{n-k_1-k_2} (\Delta x)^2$$



Devil's Game (3)

Thus pdf..

$$\mathrm{f}(x_1,x_2,x_3)=\frac{n!}{(k_1-1)!(k_2-1)!(n-k_1-k_2)!}x_1^{k_1-1}x_2^{k_2-1}x_3^{n-k_1-k_2}$$

$$=\frac{T(n+1)}{T(k_1)T(k_2)T(n-k_1-k_2+1)}x_1^{k_1-1}x_2^{k_2-1}x_3^{n-k_1-k_2}$$

$$=\frac{\mathrm{T}(\alpha_1+\alpha_2+\alpha_3)}{\mathrm{T}(\alpha_1)\mathrm{T}(\alpha_2)\mathrm{T}(\alpha_3)}x_1^{\alpha_1-1}x_2^{\alpha_2-1}x_3^{\alpha_3-1} \qquad \Longleftarrow \quad \textbf{Dirichlet}$$

$$\alpha_1=k_1,\alpha_2=k_2,\alpha_3=n-k_1-k_2$$

Devil's Game (4)

Again, He allows you to push **m** times, and tells you which one is bigger among them

What we know :

- Prior knowledge

$$\mathrm{Di}r(\vec{X} \mid \vec{k})$$

- Say m_1 , m_2 , m_3

$$\mathrm{Multi}(\vec{m} \mid \vec{x}) \qquad \vec{m} = (m_1, m_2, m_3)$$

- Posterior knowledge

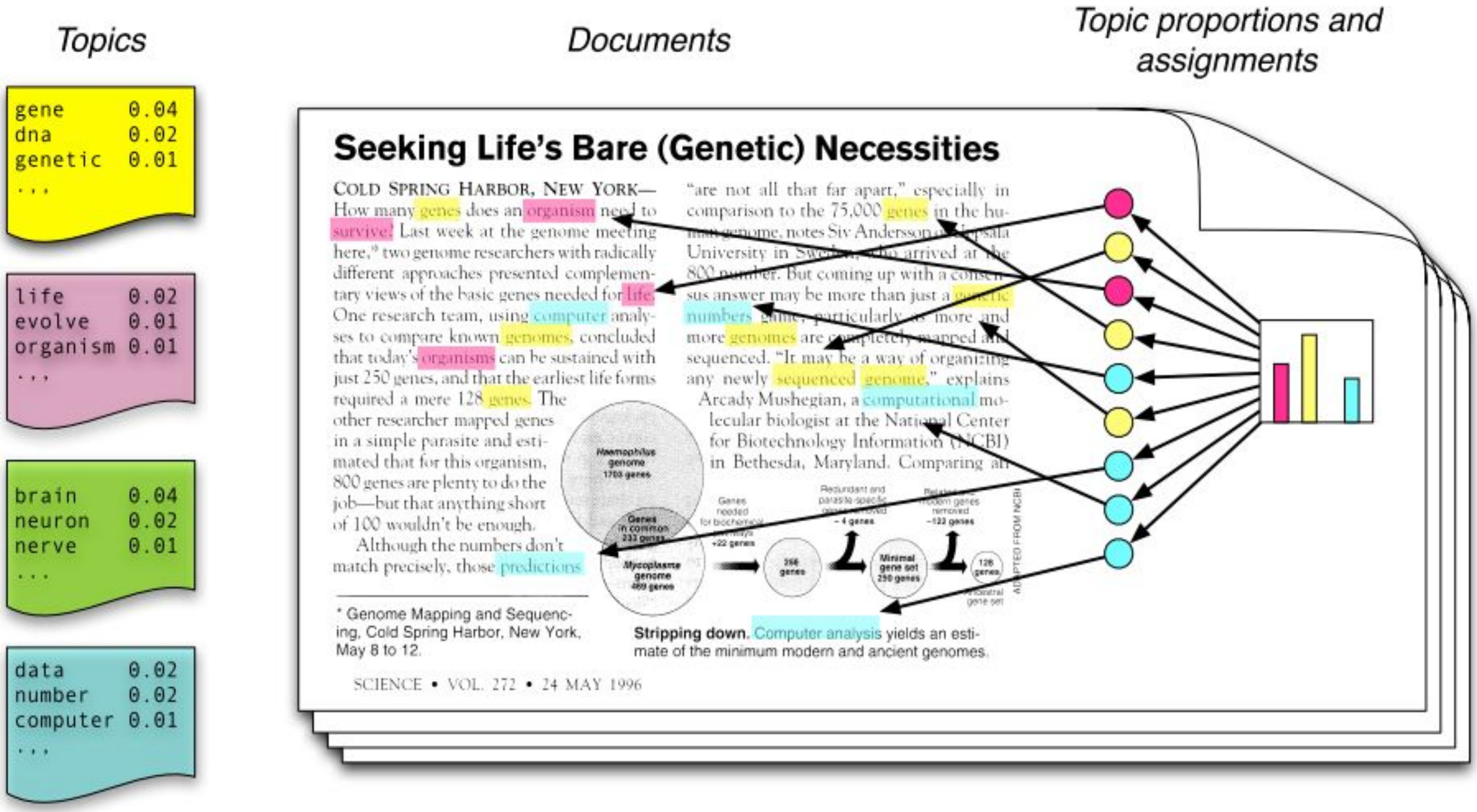
$$\mathrm{Di}r(\vec{X} \mid \vec{k} + \vec{\mathrm{m}})$$

Dirichlet-Multinomial Conjugate

Part Three

Latent Dirichlet Allocation

LDA Topic Model



gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,[*] two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

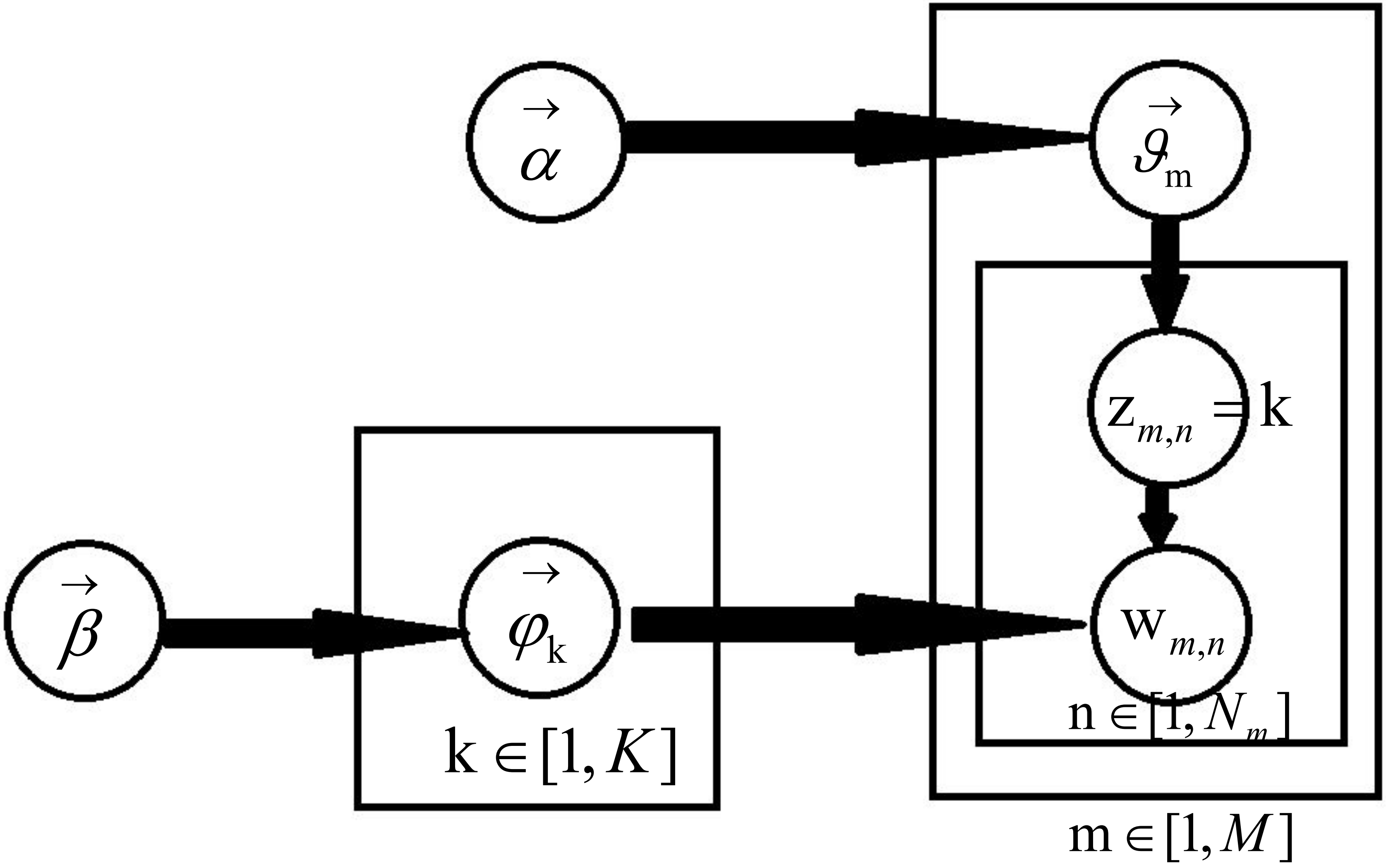
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

[*] Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

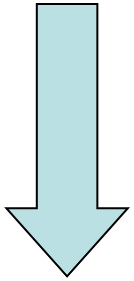
Create Document

Topic distribution of doc m



LDA Topic Model

- **Seen**
- Words in the documents



$P(z|w)$

- **Latent**
- Topic
- Topic distribution of document
- Word distribution of topic

LDA Topic Model

Keys:

- 2 Processes

$$\begin{tikzcd}
\vec{\alpha} \arrow[r] & \vec{\vartheta}_{\mathrm{m}} \arrow[r] & \mathrm{z}_{m,n}
\end{tikzcd}$$

$$\begin{tikzcd}
\vec{\beta} \arrow[r] & \vec{\varphi}_{\mathrm{k}} \arrow[r] & \mathrm{w}_{m,n} \mid \mathrm{k} = z_{m,n}
\end{tikzcd}$$

- (M+K) Dirichlet-Multinomial Conjugate

LDA Topic Model

$$\begin{tikzcd}[column sep=huge] \vec{\alpha} \arrow[r, "\text{Dirichlet}"] & \vec{\vartheta}_{\mathrm{m}} \arrow[r, "\text{Multinomial}"] & z_{m,n} \end{tikzcd}$$

$$\mathrm{P}(\vec{\mathrm{z}}_{\mathrm{m}} \mid \vec{\alpha}) = \int \mathrm{P}(\vec{\mathrm{z}}_{\mathrm{m}} \mid \vec{p})\mathrm{P}(\vec{p} \mid \vec{\alpha})\mathrm{d}\,\vec{p}$$

$$= \int \prod_{\mathrm{k}=1}^{\mathrm{V}} p_{\mathrm{k}}^{\,n_k}\,\mathrm{Dir}(\vec{p} \mid \vec{\alpha})\mathrm{d}\,\vec{p}$$

$$= \int \prod_{\mathrm{k}=1}^{\mathrm{V}} p_{\mathrm{k}}^{\,n_k}\,\frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{V} p_{\mathrm{k}}^{\,\alpha_k-1}\mathrm{d}\,\vec{p}$$

$$= \frac{1}{\Delta(\vec{\alpha})}\int \prod_{\mathrm{k}=1}^{\mathrm{V}} p_{\mathrm{k}}^{\,n_k+\alpha_k-1}\mathrm{d}\,\vec{p}$$

$$= \frac{\Delta(\vec{n}_{\mathrm{m}}+\vec{\alpha})}{\Delta(\vec{\alpha})} \qquad \vec{n}_{\mathrm{m}} = (\mathrm{n}_m^{\,1},\cdots \mathrm{n}_m^{\,K})$$

- n_m^k is the # of word that topic k create in the doc m

LDA Topic Model

$$\vec{\beta} \xrightarrow{\text{Dirichlet}} \vec{\varphi}_{\mathrm{k}} \xrightarrow{\text{Multinomial}} \mathrm{w}_{m,n} \mid \mathrm{k} = z_{m,n}$$

$$\mathrm{P}(\vec{\mathrm{w}}(\mathrm{k}) \mid \vec{\beta}) = \frac{\Delta(\vec{n}_{\mathrm{k}} + \vec{\beta})}{\Delta(\vec{\beta})} \qquad \vec{n}_{\mathrm{k}} = (\mathrm{n}_k^{\,1}, \cdots \mathrm{n}_k^{\,V})$$

- n_k is the word distribution of topic k

LDA Topic Model

- Joint distribution

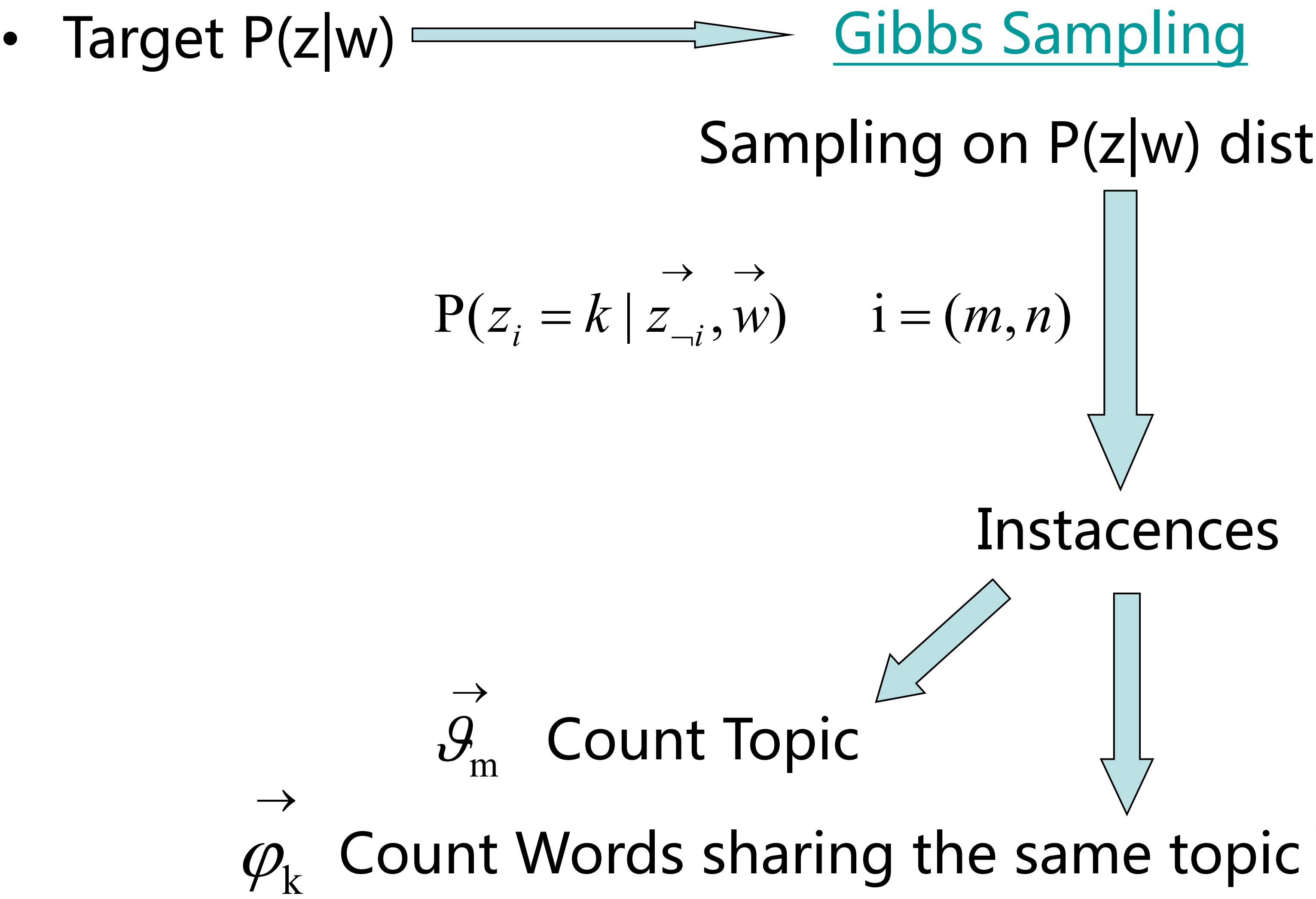
$$\mathrm{P}(\vec{w}, \vec{z} \mid \vec{\alpha}, \vec{\beta})$$

$$= \mathrm{P}(\vec{w} \mid \vec{z}, \vec{\beta})\mathrm{P}(\vec{z} \mid \vec{\alpha})$$

$$= \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k + \vec{\beta})}{\Delta(\vec{\beta})} \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

- Then what's next?

LDA Topic Model - Mind Mapping



- Target P(z|w) → Gibbs Sampling

Sampling on P(z|w) distribution

$$P(z_i = k \mid \vec{z}_{\neg i}, \vec{w}) \qquad \mathrm{i} = (m, n)$$

Instacences

$\vec{\vartheta}_{\mathrm{m}}$ Count Topic

$\vec{\varphi}_{\mathrm{k}}$ Count Words sharing the same topic

LDA + Gibbs Training

- Initialization
- Foreach word in docs, randomly give them a topic

- Gibbs Sampling
- Update the topic of every word

- Repeat step 2 until Gibbs converge

- Output
- Topic-word Matrix

LDA Inference

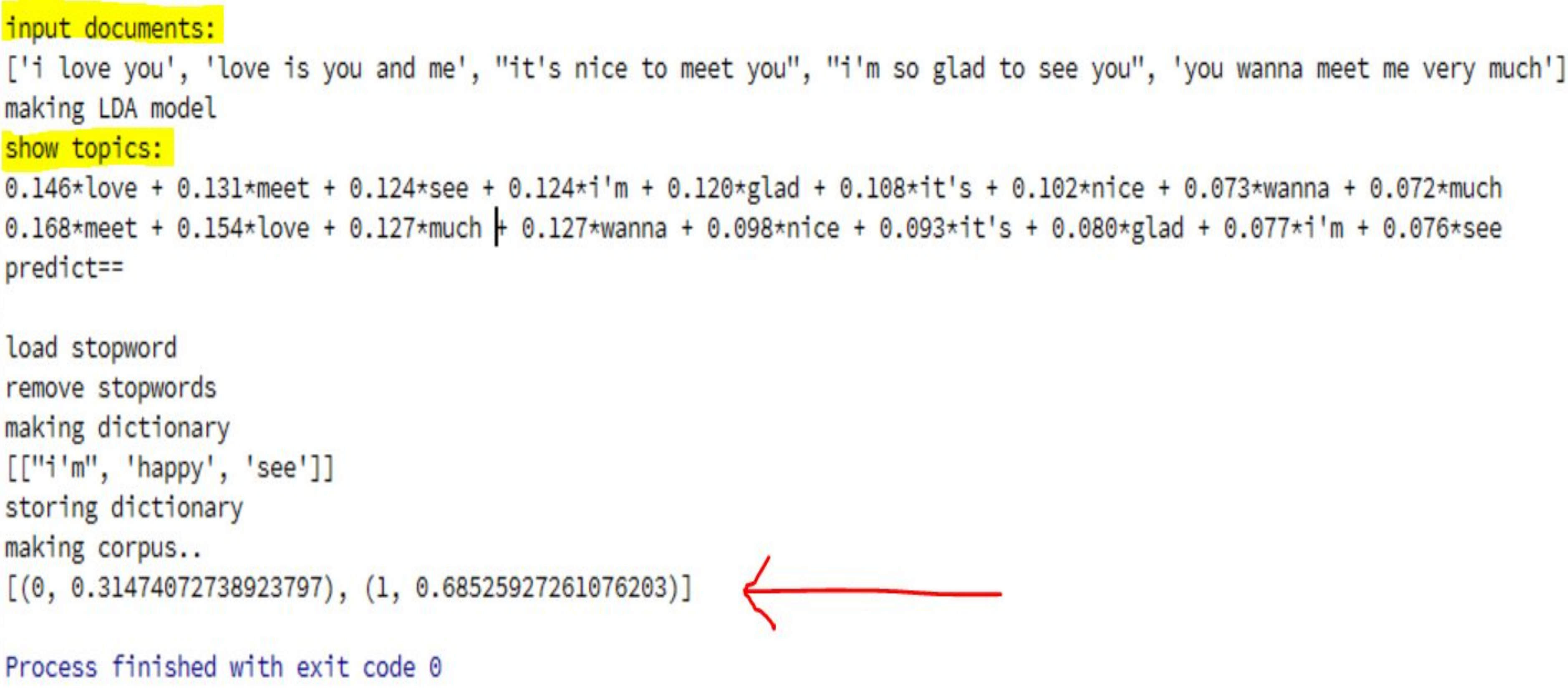
- For a new document D'
- Foreach words in D', randomly pick a topic

- Gibbs Sampling, together with training output, update the topic of every word

- Repeat step 2 until converge

- Count topic in D', then we have its $\vec{\vartheta}_{\text{new}}$

NLTK LDA Test



```
input documents:
['i love you', 'love is you and me', "it's nice to meet you", "i'm so glad to see you", 'you wanna meet me very much']
making LDA model
show topics:
0.146*love + 0.131*meet + 0.124*see + 0.124*i'm + 0.120*glad + 0.108*it's + 0.102*nice + 0.073*wanna + 0.072*much
0.168*meet + 0.154*love + 0.127*much + 0.127*wanna + 0.098*nice + 0.093*it's + 0.080*glad + 0.077*i'm + 0.076*see
predict==

load stopword
remove stopwords
making dictionary
[["i'm", 'happy', 'see']]
storing dictionary
making corpus..
[(0, 0.31474072738923797), (1, 0.68525927261076203)]

Process finished with exit code 0
```

Q&A

Markov Chain Monte Carlo (MCMC)

关键问题：

构造转移矩阵P，使得它的平稳分布就是待采样的分布p(x)

细致平稳条件：

如果非周期的马氏链的转移矩阵P和分布π(x)，满足

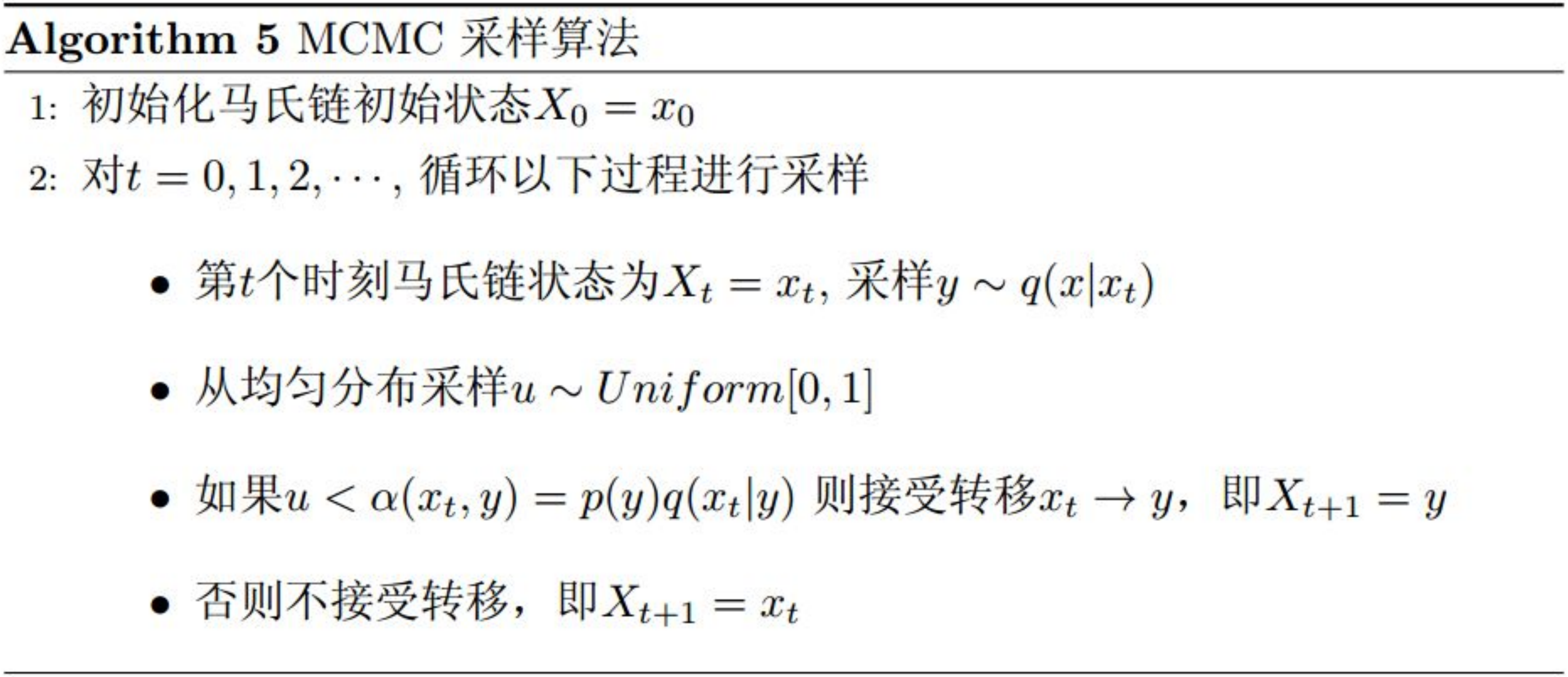
π(i)Pij = π(j)Pji for all i,j

则π(x)就是它的平稳分布

接受率：

$$p(i)\underbrace{q(i,j)\alpha(i,j)}_{Q'(i,j)} = p(j)\underbrace{q(j,i)\alpha(j,i)}_{Q'(j,i)}$$

MCMC



Algorithm 5 MCMC 采样算法

```
初始化马氏链初始状态$X_0 = x_0$
对$t = 0, 1, 2, \cdots$, 循环以下过程进行采样
    • 第$t$个时刻马氏链状态为$X_t = x_t$, 采样$y \sim q(x|x_t)$
    • 从均匀分布采样$u \sim Uniform[0, 1]$
    • 如果$u < \alpha(x_t, y) = p(y)q(x_t|y)$ 则接受转移$x_t \rightarrow y$，即$X_{t+1} = y$
    • 否则不接受转移，即$X_{t+1} = x_t$
```

Gibbs Sampling

考虑坐标轴上的两个点A(x1, y1)，B(x1, y2)

$$p(x_1, y_1)p(y_2|x_1) = p(x_1)p(y_1|x_1)p(y_2|x_1)$$

$$p(x_1, y_2)p(y_1|x_1) = p(x_1)p(y_2|x_1)p(y_1|x_1)$$

$$\Downarrow$$

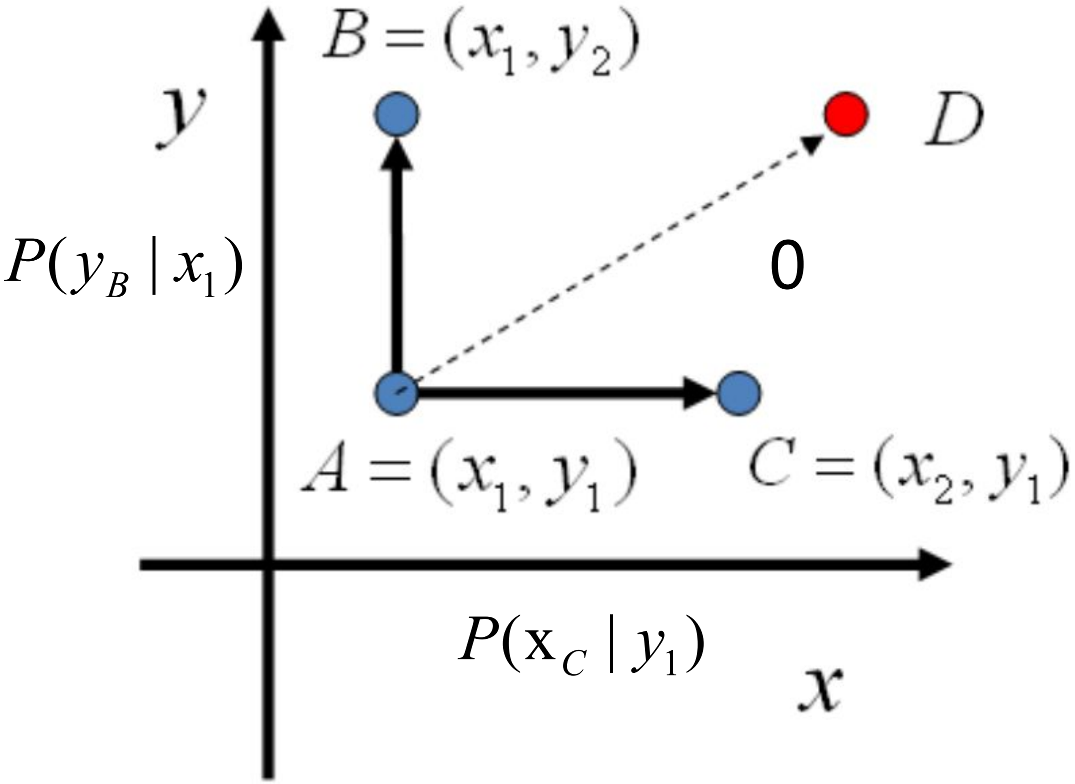
$$p(x_1, y_1)p(y_2|x_1) = p(x_1, y_2)p(y_1|x_1)$$

$$\Downarrow$$

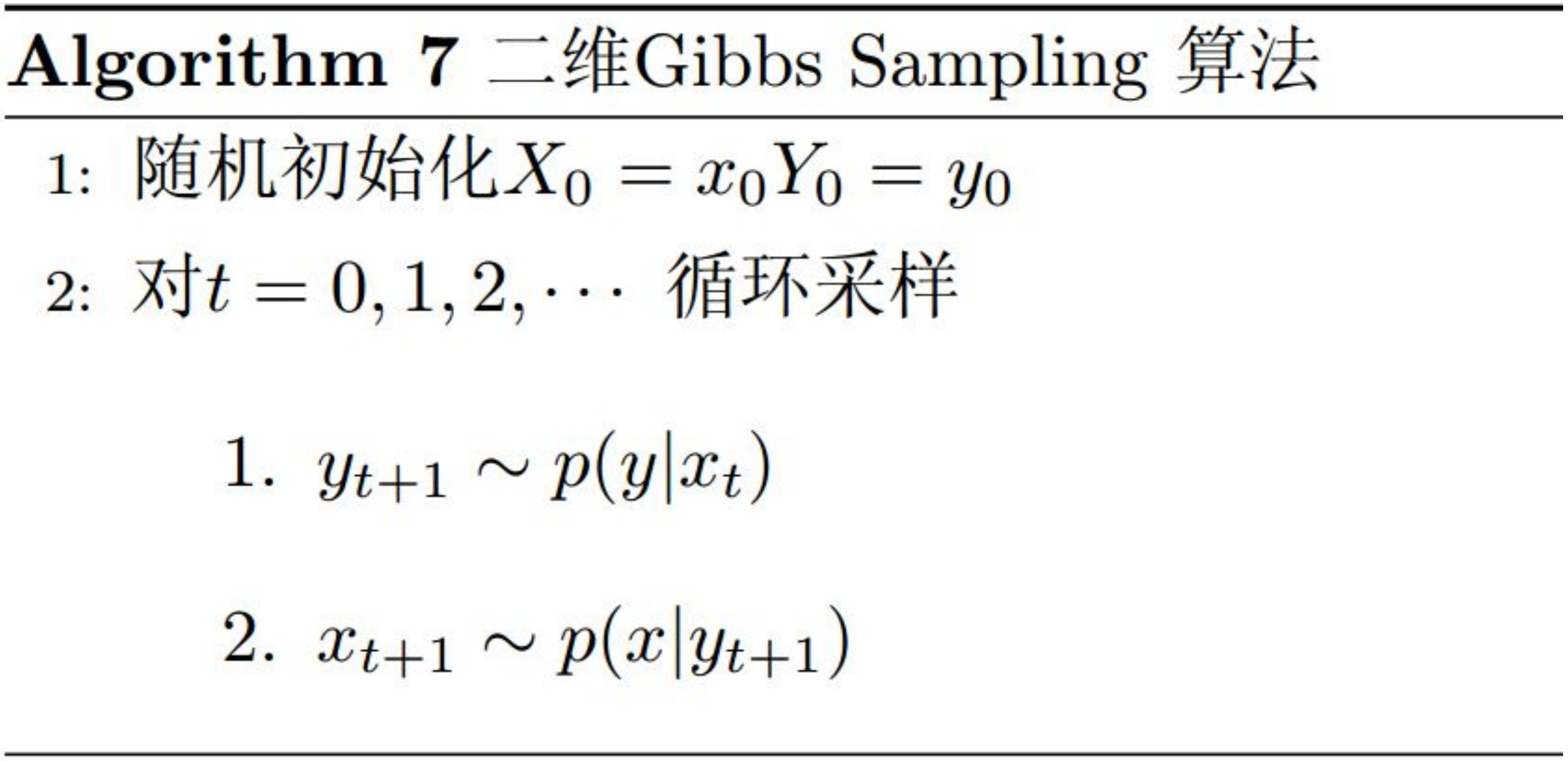
$$p(A)p(y_2|x_1) = p(B)p(y_1|x_1)$$

Gibbs Sampling

在X = X1这条线上，如使用条件分布P(Y|X)作为任何两个点之间的转移概率，那么任何两个点之间的转移概率满足平稳条件



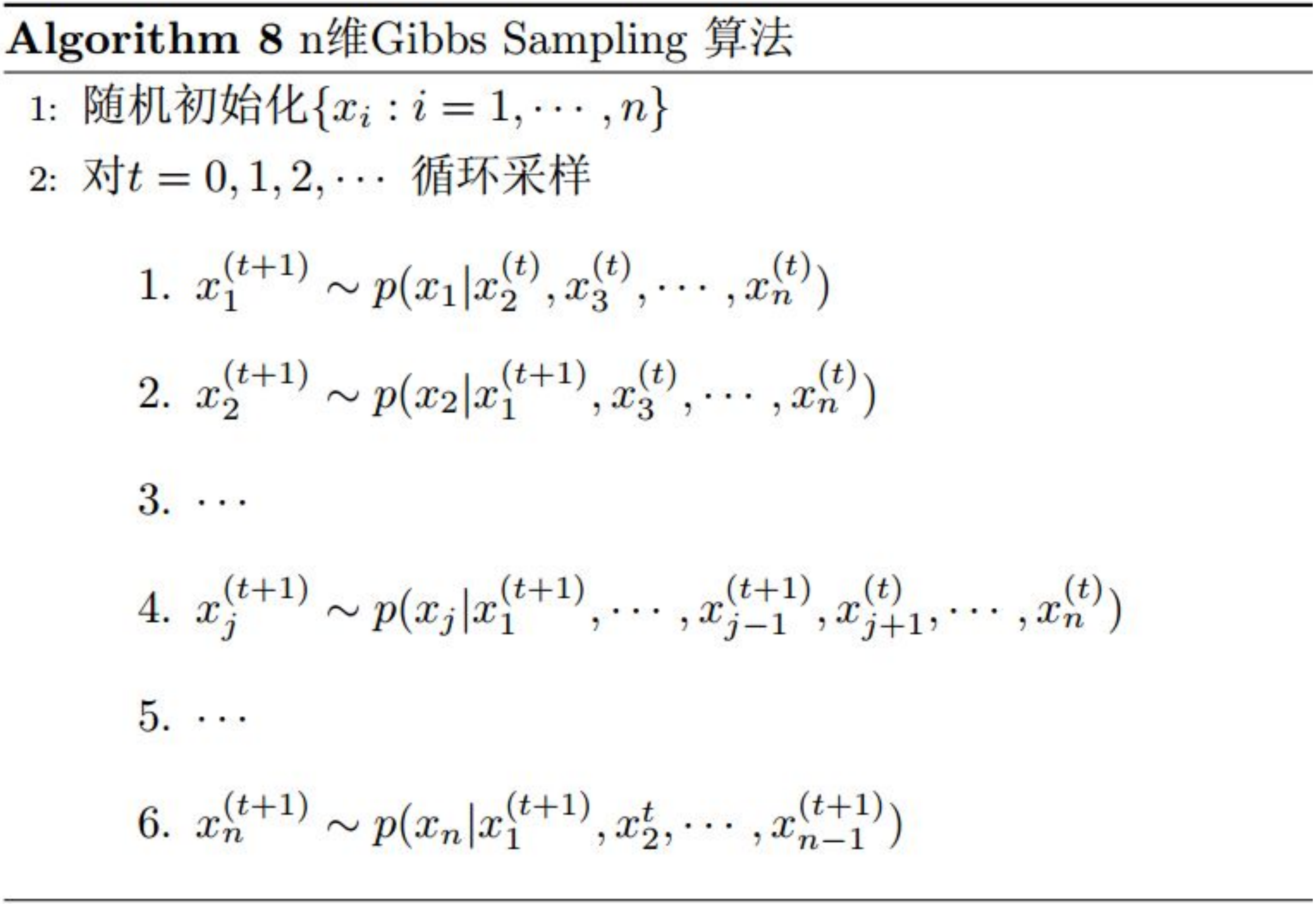
Gibbs Sampling



Algorithm 7 二维Gibbs Sampling 算法

```
1: 随机初始化$X_0 = x_0 Y_0 = y_0$
2: 对$t = 0, 1, 2, \cdots$ 循环采样
       1. $y_{t+1} \sim p(y|x_t)$
       2. $x_{t+1} \sim p(x|y_{t+1})$
```

Gibbs Sampling



Algorithm 8 n维Gibbs Sampling 算法

```
随机初始化$\{x_i : i = 1, \cdots, n\}$
对$t = 0, 1, 2, \cdots$ 循环采样
    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. $x_n^{(t+1)} \sim p(x_n | x_1^{(t+1)}, x_2^{t}, \cdots, x_{n-1}^{(t+1)})$
```

Gibbs Sampling on LDA

$$\begin{aligned}
p(z_i = k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}) \propto\; & p(z_i = k, w_i = t|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) \\
&= \int p(z_i = k, w_i = t, \vec{\theta}_m, \vec{\varphi}_k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) d\vec{\theta}_m d\vec{\varphi}_k \\
&= \int p(z_i = k, \vec{\theta}_m|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) \cdot p(w_i = t, \vec{\varphi}_k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) d\vec{\theta}_m d\vec{\varphi}_k \\
&= \int p(z_i = k|\vec{\theta}_m) p(\vec{\theta}_m|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) \cdot p(w_i = t|\vec{\varphi}_k) p(\vec{\varphi}_k|\vec{\mathbf{z}}_{\neg i}, \vec{\mathbf{w}}_{\neg i}) d\vec{\theta}_m d\vec{\varphi}_k \\
&= \int p(z_i = k|\vec{\theta}_m) Dir(\vec{\theta}_m|\vec{n}_{m,\neg i} + \vec{\alpha}) d\vec{\theta}_m \\
&\quad \cdot \int p(w_i = t|\vec{\varphi}_k) Dir(\vec{\varphi}_k|\vec{n}_{k,\neg i} + \vec{\beta}) d\vec{\varphi}_k \\
&= \int \theta_{mk} Dir(\vec{\theta}_m|\vec{n}_{m,\neg i} + \vec{\alpha}) d\vec{\theta}_m \cdot \int \varphi_{kt} Dir(\vec{\varphi}_k|\vec{n}_{k,\neg i} + \vec{\beta}) d\vec{\varphi}_k \\
&= E(\theta_{mk}) \cdot E(\varphi_{kt}) \\
&= \hat{\theta}_{mk} \cdot \hat{\varphi}_{kt}
\end{aligned}$$

back