Algorithm Foundations of Data Science

Lecture 2: Sampling

MING GAO

DaSE@ECNU
(for course related communications)
mgao@dase.ecnu.edu.cn

Mar. 14, 2018

Outline

1. Monte Carlo Method
2. Markov Chain Monte Carlo
   - MCMC Sampling Algorithm
   - Metropolis-Hastings Algorithm
   - Gibbs Sampling
   - Latent Dirichlet Allocation

Monte Carlo Method

MC methods are a class of computational algorithms that rely on repeated random sampling to obtain numerical results.

1. An early variant of it can be seen in the Buffon's needle experiment;
2. It was central to the simulations required for the Manhattan Project;
3. The founder of MC method were Stanislaw Marcin Ulam, Enrico Fermi, John von Neumann and Nicholas Metropolis.

Monte Carlo Method

MC methods are a class of computational algorithms that rely on repeated random sampling to obtain numerical results.

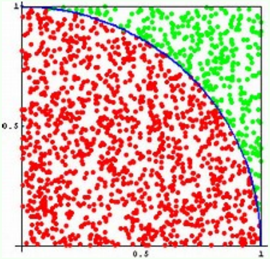
1. An early variant of it can be seen in the Buffon's needle experiment;
2. It was central to the simulations required for the Manhattan Project;
3. The founder of MC method were Stanislaw Marcin Ulam, Enrico Fermi, John von Neumann and Nicholas Metropolis.

Major components of MC methods

1. Define a domain of possible inputs;
2. Generate inputs randomly from a pdf over the domain;
3. Perform a deterministic computation on the inputs;
4. Aggregate the results.

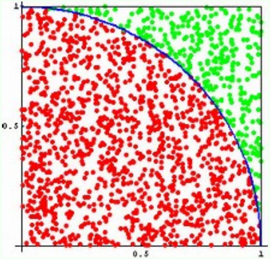
Example I

Algorithm:



Question: How accurate of the probabilistic algorithm?
We cannot answer the question in this moment, once we learn expectation of r.v.s (coming soon).

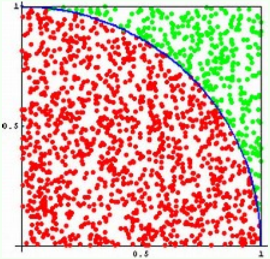
Example I



Algorithm:
Step i: It randomly and uniformly generates a point P_i inside the sample space $\Omega = \{(x, y) | 0 \leq x, y \leq 1\}$.

Question: How accurate of the probabilistic algorithm?
We cannot answer the question in this moment, once we learn expectation of r.v.s (coming soon).

Example I



Algorithm:

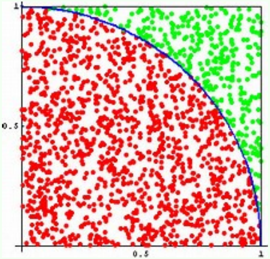
Step i: It randomly and uniformly generates a point P_i inside the sample space $\Omega = \{(x, y) | 0 \leq x, y \leq 1\}$.

Let set $S = \{(x, y) : x^2 + y^2 \leq 1 \wedge x, y \geq 0\}$ be the circle region. And $\forall P_i \in S$, we define $I_S(P_i)$ and $I_{\Omega - S}(P_i)$;

Question: How accurate of the probabilistic algorithm?

We cannot answer the question in this moment, once we learn expectation of r.v.s (coming soon).

Example I



Algorithm:

Step i: It randomly and uniformly generates a point P_i inside the sample space $\Omega = \{(x, y)|0 \leq x, y \leq 1\}$.

Let set $S = \{(x, y) : x^2 + y^2 \leq 1 \wedge x, y \geq 0\}$ be the circle region. And $\forall P_i \in S$, we define $I_S(P_i)$ and $I_{\Omega-S}(P_i)$;

$\frac{\pi}{4} \approx \frac{\sum_{i=1}^{n} I_S(P_i)}{\sum_{i=1}^{n} I_S(P_i)+\sum_{i=1}^{n} I_{\Omega-S}(P_i)}$.

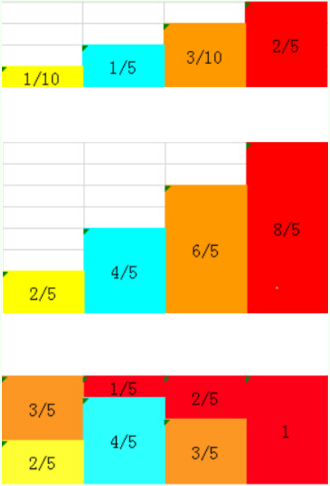
Question: How accurate of the probabilistic algorithm?

We cannot answer the question in this moment, once we learn expectation of r.v.s (coming soon).

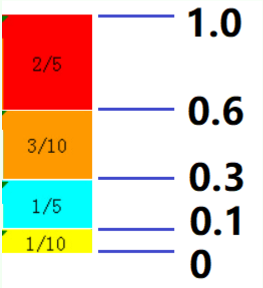
Sample with discrete distribution

How to sample from discrete distribution 0.1, 0.2, 0.3, 0.4?

Aliasing sample:



CDF sample:



$O(\log n)$ for CDF sample, and $O(1)$ for aliasing sample.

Example II: approximating probabilities

In many applications, the probability $P(Y)$ of an observed event Y must be computed as the sum over very many latent variables X of the joint probability $P(Y, X)$. That is,

$$P(Y = y) = \sum_{x \in X} P(Y = y, X = x) = \sum_{x \in X} P(Y = y | X = x) P(X = x).$$

Example II: approximating probabilities

In many applications, the probability $P(Y)$ of an observed event Y must be computed as the sum over very many latent variables X of the joint probability $P(Y, X)$. That is,

$$P(Y = y) = \sum_{x \in X} P(Y = y, X = x) = \sum_{x \in X} P(Y = y | X = x) P(X = x).$$

The term following the last equals sign is the sum over all x of a function of x, weighted by the marginal probabilities $P(X = x)$. Clearly this is an expectation, and therefore may be approximated by Monte Carlo, giving us

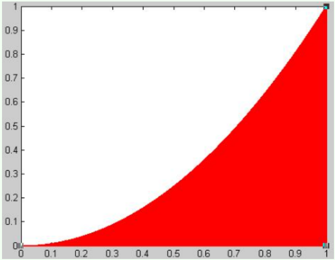
$$P(Y = y) \approx \frac{1}{n} \sum_{i=1}^{n} P(Y = y | X = x_i).$$

Example III: approximating integral $\int_0^1 x^2 dx$

1. Draw a square, then inscribe a parabola within it;
2. Uniformly scatter objects of uniform size over the square;
3. Count # objects inside the parabola and the total number of objects;
4. The ratio (0.3328) of the two counts is an estimate of $\int_0^1 x^2 dx$.

Example III: approximating integral $\int_0^1 x^2 dx$

1. Draw a square, then inscribe a parabola within it;
2. Uniformly scatter objects of uniform size over the square;
3. Count $\#$ objects inside the parabola and the total number of objects;
4. The ratio (0.3328) of the two counts is an estimate of $\int_0^1 x^2 dx$.



For an integral $\int_a^b f(x)dx$, it is hard to find a rectangle to bound the value of $f(x)$, especially for a high-dimensional function. Alternatively, we compute $\int_a^b \frac{f(x)}{p(x)} p(x)dx$.

Example IV: approximating expectation $f(x)$

Computing approximate integrals of the form $\int f(x)p(x)dx$, i.e., computing expectation of $f(x)$ using density $p(x)$.

Example IV: approximating expectation $f(x)$

Computing approximate integrals of the form $\int f(x)p(x)dx$, i.e., computing expectation of $f(x)$ using density $p(x)$.

1. Let $\{x_i\}$ is an i.i.d. random sample drawn from $p(x)$;
2. The strong law of large numbers says:

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) \longrightarrow \int f(x)p(x)dx \text{ (a.s).} \tag{1}$$

3. The rate of convergence is proportional to $\sqrt{N}$;

Example IV: approximating expectation $f(x)$

Computing approximate integrals of the form $\int f(x)p(x)dx$, i.e., computing expectation of $f(x)$ using density $p(x)$.

1. Let $\{x_i\}$ is an i.i.d. random sample drawn from $p(x)$;
2. The strong law of large numbers says:

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i) \longrightarrow \int f(x)p(x)dx \text{ (a.s).} \tag{1}$$

3. The rate of convergence is proportional to $\sqrt{N}$;
4. Major issues:
   - The proportionality constant increases exponentially with the dimension of the integral.
   - Another problem is that sampling from complex distributions is not as easy as uniform.

Rejection sampling: approximating $\int f(x)p(x)dx$

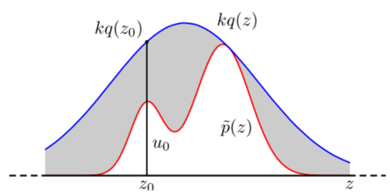
$\frac{1}{N}\sum_{i=1}^{N} f(x_i)$ is difficult to compute since it is hard to draw from $p(x)$.

```
i ← 0;
while i ≠ N do
    x^(i) ~ q(x);
    u ~ U(0,1);
    if u < p(x^(i)) / (k q(x^(i))) then;
        accept x^(i);
        i ← i + 1;
    else
        reject x^(i);
    end if
end while
```

Rejection sampling: approximating $\int f(x)p(x)dx$

$\frac{1}{N}\sum_{i=1}^{N} f(x_i)$ is difficult to compute since it is hard to draw from $p(x)$.

1: $i \leftarrow 0$;
2: while $i \neq N$ do
3: $\quad x^{(i)} \sim q(x)$;
4: $\quad u \sim U(0,1)$;
5: $\quad$ if $u < \frac{p(x^{(i)})}{kq(x^{(i)})}$ then;
6: $\qquad$ accept $x^{(i)}$;
7: $\qquad i \leftarrow i+1$;
8: $\quad$ else
9: $\qquad$ reject $x^{(i)}$;
10: $\quad$ end if
11: end while



where density $q(x)$ (e.g., Gaussian) can sample directly.
What is the average acceptance ratio?
However, it is hard to find the reasonable $q(x)$ and the value of k.

Importance sampling: approximating $I(f) = \int f(x)p(x)dx$

If we have a density $q(x)$ (proposal distribution) which is easy to sample from, we can sample $x^{(i)} \sim q(x)$. We define the importance weight as

$$w(x^{(i)}) = \frac{p(x^{(i)})}{q(x^{(i)})}.$$

Importance sampling: approximating $I(f) = \int f(x)p(x)dx$

If we have a density $q(x)$ (proposal distribution) which is easy to sample from, we can sample $x^{(i)} \sim q(x)$. We define the importance weight as

$$w(x^{(i)}) = \frac{p(x^{(i)})}{q(x^{(i)})}.$$

Consider the weighted Monte Carlo sum:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})w(x^{(i)}) &= \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})\frac{p(x^{(i)})}{q(x^{(i)})} \\ \longrightarrow \int f(x)\frac{p(x)}{q(x)}q(x)dx(\text{a.s}) &= \int f(x)p(x)dx.\end{aligned}$$

Approximating probabilities Cont.d

Going back to Example II with the discrete sum over latent variables X it is clear that the optimal importance sampling function would be the conditional distribution of X given Y, i.e.,

$$P(Y = y) = \sum_{x \in X} \frac{P(Y = y, X = x)}{P(X|Y = y)} P(X|Y = y).$$

Approximating probabilities Cont.d

Going back to Example II with the discrete sum over latent variables X it is clear that the optimal importance sampling function would be the conditional distribution of X given Y, i.e.,

$$P(Y = y) = \sum_{x \in X} \frac{P(Y = y, X = x)}{P(X|Y = y)} P(X|Y = y).$$

Note that the right side is a conditional expectation of a function of X.

Approximating probabilities Cont.d

Going back to Example II with the discrete sum over latent variables X it is clear that the optimal importance sampling function would be the conditional distribution of X given Y, i.e.,

$$P(Y = y) = \sum_{x \in X} \frac{P(Y = y, X = x)}{P(X|Y = y)} P(X|Y = y).$$

Note that the right side is a conditional expectation of a function of X. As before $P(X|Y)$ is not computable.

Approximating probabilities Cont.d

Going back to Example II with the discrete sum over latent variables X it is clear that the optimal importance sampling function would be the conditional distribution of X given Y, i.e.,

$$P(Y=y)=\sum_{x\in X}\frac{P(Y=y,X=x)}{P(X|Y=y)}P(X|Y=y).$$

Note that the right side is a conditional expectation of a function of X. As before $P(X|Y)$ is not computable.
So one must turn to finding some other distribution, i.e., $P^*(X)$, that is close to $P(X|Y)$ but which is more easily sampled from and computed.

Analysis of importance sampling

How to pick $q(x)$

We can sample from any distribution $q(x)$. In practice, we would like to choose $q(x)$ as close as possible to $|f(x)|w(x)$ to reduce the variance of our estimator.

Analysis of importance sampling

How to pick $q(x)$

We can sample from any distribution $q(x)$. In practice, we would like to choose $q(x)$ as close as possible to $|f(x)|w(x)$ to reduce the variance of our estimator.

- We have $Var_{q(x)}f(x)w(x) = \mathbb{E}_{q(x)}f(x)^2w(x)^2 - I(f)^2$.
- Furthermore, we have

$$\begin{aligned}\mathbb{E}_{q(x)}f(x)^2w(x)^2 &\geq (\mathbb{E}_{q(x)}|f(x)|w(x))^2\\ &= (\int |f(x)|p(x)dx)^2.\end{aligned}$$

- The term $I(f)^2$ is independent of $q(x)$. So, the best $q^*(x)$ which makes the variance minimum is given by $q^*(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx}$.

The Main idea of MCMC

We cannot sample directly from the target distribution $p(x)$ in the integral $\int f(x)p(x)dx$.

- Create a Markov chain whose transition matrix does not depend on the normalization term.

The Main idea of MCMC

We cannot sample directly from the target distribution $p(x)$ in the integral $\int f(x)p(x)dx$.

- Create a Markov chain whose transition matrix does not depend on the normalization term.
- Make sure the chain has a stationary distribution and it is equal to the target distribution.

The Main idea of MCMC

We cannot sample directly from the target distribution $p(x)$ in the integral $\int f(x)p(x)dx$.

- Create a Markov chain whose transition matrix does not depend on the normalization term.
- Make sure the chain has a stationary distribution and it is equal to the target distribution.
- After sufficient number of iterations, the chain will converge the stationary distribution.

Markov Chain Monte Carlo

Overview

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its stationary distribution.

Markov Chain Monte Carlo

Overview

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its stationary distribution.

- The algorithm is proposed in 1953, which is the top-10 most important algorithms in the 20th century.

Markov Chain Monte Carlo

Overview

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its stationary distribution.

- The algorithm is proposed in 1953, which is the top-10 most important algorithms in the 20th century.
- MCMC works by generating a sequence of sample values in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired distribution, $\pi(i)$.

Markov Chain Monte Carlo

Overview

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its stationary distribution.

- The algorithm is proposed in 1953, which is the top-10 most important algorithms in the 20th century.
- MCMC works by generating a sequence of sample values in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired distribution, $\pi(i)$.
- That is, a Markov Chain has stationary distribution $\pi(i)$ associated with transition probability matrix P.

Markov Chain Monte Carlo

Overview

Markov Chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its stationary distribution.

- The algorithm is proposed in 1953, which is the top-10 most important algorithms in the 20th century.
- MCMC works by generating a sequence of sample values in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired distribution, $\pi(i)$.
- That is, a Markov Chain has stationary distribution $\pi(i)$ associated with transition probability matrix P.
- The Markov Chain converges to the stationary distribution $\pi(i)$ for arbitrary initial status x_0.

Stationary Distribution

Theorem

Let $X_0, X_1, \cdots,$ be an irreducible and aperiodic Markov chain with transition matrix P. Then, $\lim_{n\to\infty} P_{ij}^n$ exists and independent of i, denoted as $\lim_{n\to\infty} P_{ij}^n = \pi(j)$. We also have

Stationary Distribution

Theorem

Let $X_0, X_1, \cdots$, be an irreducible and aperiodic Markov chain with transition matrix P. Then, $\lim_{n\to\infty} P_{ij}^n$ exists and independent of i, denoted as $\lim_{n\to\infty} P_{ij}^n = \pi(j)$. We also have

- $$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \tag{2}$$

Stationary Distribution

Theorem

Let $X_0, X_1, \cdots$, be an irreducible and aperiodic Markov chain with transition matrix P. Then, $\lim_{n\to\infty} P_{ij}^n$ exists and independent of i, denoted as $\lim_{n\to\infty} P_{ij}^n = \pi(j)$. We also have

- $$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \tag{2}$$
- $\pi(j) = \sum_{i=1}^{\infty} \pi(i) P_{ij}$, and $\sum_{i=1}^{\infty} \pi(i) = 1$.

Stationary Distribution

Theorem

Let $X_0, X_1, \cdots$, be an irreducible and aperiodic Markov chain with transition matrix P. Then, $\lim_{n\to\infty} P_{ij}^n$ exists and independent of i, denoted as $\lim_{n\to\infty} P_{ij}^n = \pi(j)$. We also have

- $$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \pi(1) & \pi(2) & \cdots & \pi(j) & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \tag{2}$$
- $\pi(j) = \sum_{i=1}^{\infty} \pi(i) P_{ij}$, and $\sum_{i=1}^{\infty} \pi(i) = 1$.
- π is the unique and non-negative solution for equation $\pi P = \pi$.

Detailed Balance Condition

Theorem

Let $X_0, X_1, \cdots$, be an aperiodic Markov chain with transition matrix P and distribution π. If the following condition holds,

$$\pi(i)P_{ij} = \pi(j)P_{ji}, \text{ for all } i, j \tag{3}$$

then $\pi(x)$ is the stationary distribution of the Markov chain. The above equation is called the detailed balance condition.

Detailed Balance Condition

Theorem

Let $X_0, X_1, \cdots$, be an aperiodic Markov chain with transition matrix P and distribution π. If the following condition holds,

$$\pi(i)P_{ij} = \pi(j)P_{ji}, \text{ for all } i,j \tag{3}$$

then $\pi(x)$ is the stationary distribution of the Markov chain. The above equation is called the detailed balance condition.

- Proof: $\sum_{i=1}^{\infty} \pi(i)P_{ij} = \sum_{i=1}^{\infty} \pi(j)P_{ji} = \pi(j)\sum_{i=1}^{\infty} P_{ji} = \pi(j) \Rightarrow \pi P = \pi$.

Detailed Balance Condition

Theorem

Let $X_0, X_1, \cdots$, be an aperiodic Markov chain with transition matrix P and distribution π. If the following condition holds,

$$\pi(i)P_{ij} = \pi(j)P_{ji}, \text{ for all } i,j \tag{3}$$

then $\pi(x)$ is the stationary distribution of the Markov chain. The above equation is called the detailed balance condition.

- Proof: $\sum_{i=1}^{\infty} \pi(i)P_{ij} = \sum_{i=1}^{\infty} \pi(j)P_{ji} = \pi(j)\sum_{i=1}^{\infty} P_{ji} = \pi(j) \Rightarrow \pi P = \pi$.
- In general, $\pi(i)P_{ij} \neq \pi(j)P_{ji}$. That is, $\pi(i)$ may not be the stationary distribution.

Detailed Balance Condition

Theorem

Let $X_0, X_1, \cdots$, be an aperiodic Markov chain with transition matrix P and distribution π. If the following condition holds,

$$\pi(i)P_{ij} = \pi(j)P_{ji}, \text{ for all } i, j \tag{3}$$

then $\pi(x)$ is the stationary distribution of the Markov chain. The above equation is called the detailed balance condition.

- Proof: $\sum_{i=1}^{\infty} \pi(i)P_{ij} = \sum_{i=1}^{\infty} \pi(j)P_{ji} = \pi(j)\sum_{i=1}^{\infty} P_{ji} = \pi(j) \Rightarrow \pi P = \pi$.
- In general, $\pi(i)P_{ij} \neq \pi(j)P_{ji}$. That is, $\pi(i)$ may not be the stationary distribution.
- The natural question is how to revise the Markov Chain such that π becomes a stationary distribution. For example, we introduce a function $\alpha(i,j)$ s.t. $\pi(i)P_{ij}\alpha(i,j) = \pi(j)P_{ji}\alpha(j,i)$

Outline

1. Monte Carlo Method

2. Markov Chain Monte Carlo
   - MCMC Sampling Algorithm
   - Metropolis-Hastings Algorithm
   - Gibbs Sampling
   - Latent Dirichlet Allocation

Revising the Markov Chain

Choosing a reasonable parameter

How to choose $\alpha(i,j)$ such that $\pi(i)P_{ij}\alpha(i,j) = \pi(j)P_{ji}\alpha(j,i)$.

Revising the Markov Chain

Choosing a reasonable parameter

How to choose $\alpha(i,j)$ such that $\pi(i)P_{ij}\alpha(i,j) = \pi(j)P_{ji}\alpha(j,i)$.

- In terms of symmetry, we simply choose $\alpha(i,j) = \pi(j)P_{ji},\ \alpha(j,i) = \pi(i)P_{ij}$.

Revising the Markov Chain

Choosing a reasonable parameter

How to choose $\alpha(i,j)$ such that $\pi(i)P_{ij}\alpha(i,j) = \pi(j)P_{ji}\alpha(j,i)$.

- In terms of symmetry, we simply choose $\alpha(i,j) = \pi(j)P_{ji},\ \alpha(j,i) = \pi(i)P_{ij}$.
- Therefore, we have $\pi(i)\underbrace{P_{ij}\alpha(i,j)}_{Q_{ij}} = \pi(j)\underbrace{P_{ji}\alpha(j,i)}_{Q_{ji}}$.

Revising the Markov Chain

Choosing a reasonable parameter

How to choose $\alpha(i,j)$ such that $\pi(i)P_{ij}\alpha(i,j) = \pi(j)P_{ji}\alpha(j,i)$.

- In terms of symmetry, we simply choose $\alpha(i,j) = \pi(j)P_{ji},\ \alpha(j,i) = \pi(i)P_{ij}$.
- Therefore, we have $\pi(i)\underbrace{P_{ij}\alpha(i,j)}_{Q_{ij}} = \pi(j)\underbrace{P_{ji}\alpha(j,i)}_{Q_{ji}}$.
- The transition matrix:
$$\begin{cases} Q_{ij} = P_{ij}\alpha(i,j), & \text{if } j \neq i; \\ Q_{ii} = P_{ii} + \sum_{k\neq i} P_{i,k}(1-\alpha(i,k)), & \text{Otherwise.} \end{cases}$$

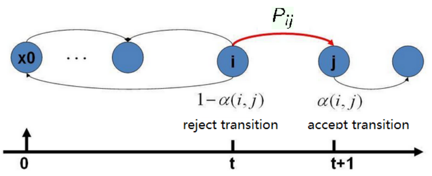
Revising the Markov Chain

Choosing a reasonable parameter

How to choose $\alpha(i,j)$ such that $\pi(i)P_{ij}\alpha(i,j) = \pi(j)P_{ji}\alpha(j,i)$.

- In terms of symmetry, we simply choose $\alpha(i,j) = \pi(j)P_{ji}$, $\alpha(j,i) = \pi(i)P_{ij}$.
- Therefore, we have $\pi(i)\underbrace{P_{ij}\alpha(i,j)}_{Q_{ij}} = \pi(j)\underbrace{P_{ji}\alpha(j,i)}_{Q_{ji}}$.
- The transition matrix:

$$\begin{cases} Q_{ij} = P_{ij}\alpha(i,j), & \text{if } j \neq i; \\ Q_{ii} = P_{ii} + \sum_{k\neq i} P_{i,k}(1-\alpha(i,k)), & \text{Otherwise.} \end{cases}$$



- Accept with probability $\alpha(i,j)$;
- Otherwise stay in the current location.

MCMC Sampling Algorithm

Let $P(x_2|x_1)$ be a proposed distribution.

0: initialize $x^{(0)}$;
1: for $i = 0$ to $N-1$ do
2: sample $u \sim U[0,1]$;
3: sample $x \sim P(x|x^{(i)})$;
4: if $u < \alpha(x, x^{(i)}) = \pi(x)P(x^{(i)}|x)$,
5: then $x^{(i+1)} = x$;
6: else reject x, and $x^{(i+1)} = x^{(i)}$;
7: endif
8: endfor
9: **output** Last N samples;

Observation

Let $\alpha(i,j) = 0.1$, and $\alpha(j,i) = 0.2$ satisfy the detailed balance condition, thus we have

$$\pi(i)P_{ij}0.1 = \pi(j)P_{ji}0.2.$$

The small value of $\alpha(i,j)$ results in a high rejection ratio. We therefore modify the equation as follows:

$$\pi(i)P_{ij}0.5 = \pi(j)P_{ji}1.$$

Outline

1 Monte Carlo Method

2 Markov Chain Monte Carlo
- MCMC Sampling Algorithm
- Metropolis-Hastings Algorithm
- Gibbs Sampling
- Latent Dirichlet Allocation

Metropolis-Hastings algorithm

Let $P(x_2|x_1)$ be a proposed distribution.

0: initialize $x^{(0)}$;
1: for $i = 0$ to max do
2: sample $u \sim U[0,1]$;
3: sample $x \sim P(x|x^{(i)})$;
4: if $u < \min\{1, \frac{\pi(x)P(x^{(i)}|x)}{\pi(x^{(i)})P(x|x^{(i)})}\}$,
5: then $x^{(i+1)} = x$;
6: else reject x, and $x^{(i+1)} = x^{(i)}$;
7: endif
8: endfor
9: **output** Last N samples;

Observation

If we let

$$\alpha(i,j) = \min\left\{1, \frac{\pi(x)P(x^{(i)}|x)}{\pi(x^{(i)})P(x|x^{(i)})}\right\},$$

we can get a high accept ratio, and further improve the algorithm efficiency.

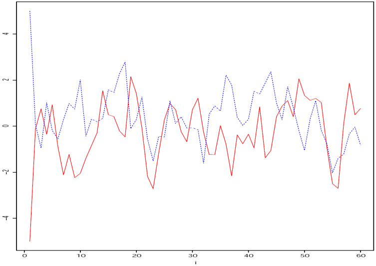
However, for high-dimensional P, Metropolis-Hastings algorithm may be inefficient because of $\alpha < 1$. Is there a way to find a transition matrix with acceptance ratio $\alpha = 1$?

Properties of MCMC

- Trade-off between Mixing rate and Acceptance ratio, where

$$Acceptance\ ratio = \mathbb{E}[\alpha(x_i, x)]$$
$$Mixing\ ratio = rate\ that\ the\ chain\ moves\ around\ the\ dist.$$

- We can have multiple transition matrices P_i (i.e., proposal distribution), and apply them in turn.



Observation

For example:

- **Sample:** $x^{(t+1)}|x^{(t)} \sim \mathcal{N}(0.5x^{(t)}, 1.0)$;
- **Convergence:** $x^{(t)}|x^{(0)} \sim \mathcal{N}(0, 1.33), t \to +\infty$.

Outline

1 Monte Carlo Method

2 Markov Chain Monte Carlo
- MCMC Sampling Algorithm
- Metropolis-Hastings Algorithm
- Gibbs Sampling
- Latent Dirichlet Allocation

Intuition

Example: two-dimensional case

Let $P(x, y)$ be a two-dimensional probability distribution, and two points $A(x_1, y_1)$ and $B(x_1, y_2)$. We have

$$\begin{aligned} P(x_1, y_1)P(y_2|x_1) &= P(x_1)P(y_1|x_1)P(y_2|x_1) \qquad (4) \\ P(x_1, y_2)P(y_1|x_1) &= P(x_1)P(y_2|x_1)P(y_1|x_1) \qquad (5) \end{aligned}$$

That is

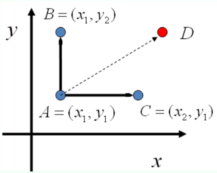
$$P(x_1, y_1)P(y_2|x_1) = P(x_1, y_2)P(y_1|x_1) \qquad (6)$$

i.e.,

$$P(A)P(y_2|x_1) = P(B)P(y_1|x_1) \qquad (7)$$

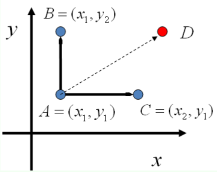
Intuition Cont'd

If $p(y|x_i)$ is considered as the transition probability of two points whose x-axis coordinates are x_i. Therefore, transition between these two points satisfies the *detailed balance condition*, i.e., $P(A)P(y_1|x_1) = P(B)P(y_2|x_1)$ holds.



Intuition Cont'd

If $p(y|x_i)$ is considered as the transition probability of two points whose x-axis coordinates are x_i. Therefore, transition between these two points satisfies the *detailed balance condition*, i.e., $P(A)P(y_1|x_1) = P(B)P(y_2|x_1)$ holds.



Transition matrix

The transition probabilities between two points A and B are given by $T(A \to B)$.

$$T(A \to B) = \begin{cases} p(y_B|x_1), & \text{if } x_A = x_B = x_1; \\ p(x_B|y_1), & \text{if } y_A = y_B = y_1; \\ 0, & \text{otherwise.} \end{cases}$$

It is easy to confirm that the detailed balance condition holds, i.e., $p(A)T(A \to B) = p(B)T(B \to A)$.

Multivariate case

For multivariate case

Let $P(x_i|\mathbf{x}_{-i}) = P(x_i|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$. The transition probabilities are given by $T(\mathbf{x} \to \mathbf{x}^{'}) = P(x_i^{'}|\mathbf{x}_{-i})$. Then, we have:

Multivariate case

For multivariate case

Let $P(x_i|\mathbf{x}_{-i}) = P(x_i|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$. The transition probabilities are given by $T(\mathbf{x} \rightarrow \mathbf{x}^{'}) = P(x_i^{'}|\mathbf{x}_{-i})$. Then, we have:

$$T(\mathbf{x} \rightarrow \mathbf{x}^{'})p(\mathbf{x}) = P(x_i^{'}|\mathbf{x}_{-i})P(x_i|\mathbf{x}_{-i})P(\mathbf{x}_{-i})$$
$$T(\mathbf{x}^{'} \rightarrow \mathbf{x})p(\mathbf{x}^{'}) = P(x_i|\mathbf{x}_{-i}^{'})P(x_i^{'}|\mathbf{x}_{-i}^{'})P(\mathbf{x}_{-i}^{'})$$

Multivariate case

For multivariate case

Let $P(x_i|\mathbf{x}_{-i}) = P(x_i|x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$. The transition probabilities are given by $T(\mathbf{x} \to \mathbf{x}^{'}) = P(x_i^{'}|\mathbf{x}_{-i})$. Then, we have:

$$T(\mathbf{x} \to \mathbf{x}^{'})p(\mathbf{x}) = P(x_i^{'}|\mathbf{x}_{-i})P(x_i|\mathbf{x}_{-i})P(\mathbf{x}_{-i})$$
$$T(\mathbf{x}^{'} \to \mathbf{x})p(\mathbf{x}^{'}) = P(x_i|\mathbf{x}_{-i}^{'})P(x_i^{'}|\mathbf{x}_{-i}^{'})P(\mathbf{x}_{-i}^{'})$$

Note that $\mathbf{x}_{-i}^{'} = \mathbf{x}_{-i}$. That is,

$$T(\mathbf{x} \to \mathbf{x}^{'})p(\mathbf{x}) = T(\mathbf{x}^{'} \to \mathbf{x})p(\mathbf{x}^{'}). \tag{8}$$

Multivariate case

For multivariate case

Let $P(x_i|\mathbf{x}_{-i}) = P(x_i|x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)$. The transition probabilities are given by $T(\mathbf{x}\to\mathbf{x}^{'}) = P(x_i^{'}|\mathbf{x}_{-i})$. Then, we have:

$$T(\mathbf{x}\to\mathbf{x}^{'})p(\mathbf{x}) = P(x_i^{'}|\mathbf{x}_{-i})P(x_i|\mathbf{x}_{-i})P(\mathbf{x}_{-i})$$
$$T(\mathbf{x}^{'}\to\mathbf{x})p(\mathbf{x}^{'}) = P(x_i|\mathbf{x}_{-i}^{'})P(x_i^{'}|\mathbf{x}_{-i}^{'})P(\mathbf{x}_{-i}^{'})$$

Note that $\mathbf{x}_{-i}^{'} = \mathbf{x}_{-i}$. That is,

$$T(\mathbf{x}\to\mathbf{x}^{'})p(\mathbf{x}) = T(\mathbf{x}^{'}\to\mathbf{x})p(\mathbf{x}^{'}). \tag{8}$$

- Therefore, the detailed balance condition also holds.
- Gibbs sampling is feasible if it is easy to sample from the conditional probability distribution.

Gibbs sampling algorithm (proposed distribution $P(x_i|x_{-i})$)

initialize $x_1, \cdots, x_n$;
for $\tau = 0$ to *max* do
  sample $x_1^{\tau+1} \sim P(x_1|x_2^\tau, x_3^\tau, \cdots, x_n^\tau)$;
  $\cdots$;
  sample $x_j^{\tau+1} \sim P(x_j|x_1^{\tau+1}, \cdots, x_{j-1}^{\tau+1}, x_{j+1}^\tau, \cdots, x_n^\tau)$;
  $\cdots$;
  sample $x_n^{\tau+1} \sim P(x_n|x_1^{\tau+1}, x_2^{\tau+1}, \cdots, x_{n-1}^{\tau+1})$;
output Last N samples;

Gibbs sampling algorithm (proposed distribution $P(x_i|x_{-i})$)

initialize $x_1, \cdots, x_n$;
for $\tau = 0$ to max do
    sample $x_1^{\tau+1} \sim P(x_1|x_2^{\tau}, x_3^{\tau}, \cdots, x_n^{\tau})$;
    $\cdots$;
    sample $x_j^{\tau+1} \sim P(x_j|x_1^{\tau+1}, \cdots, x_{j-1}^{\tau+1}, x_{j+1}^{\tau}, \cdots, x_n^{\tau})$;
    $\cdots$;
    sample $x_n^{\tau+1} \sim P(x_n|x_1^{\tau+1}, x_2^{\tau+1}, \cdots, x_{n-1}^{\tau+1})$;
output Last N samples;

Gibbs sampling is a type of random walk through parameter space, and can be considered as a Metroplis-Hastings algorithm with a special proposal distribution.

Gibbs sampling algorithm (proposed distribution $P(x_i|x_{-i})$)

initialize $x_1, \cdots, x_n$;
for $\tau = 0$ to max do
    sample $x_1^{\tau+1} \sim P(x_1|x_2^\tau, x_3^\tau, \cdots, x_n^\tau)$;
    $\cdots$;
    sample $x_j^{\tau+1} \sim P(x_j|x_1^{\tau+1}, \cdots, x_{j-1}^{\tau+1}, x_{j+1}^\tau, \cdots, x_n^\tau)$;
    $\cdots$;
    sample $x_n^{\tau+1} \sim P(x_n|x_1^{\tau+1}, x_2^{\tau+1}, \cdots, x_{n-1}^{\tau+1})$;
output Last N samples;

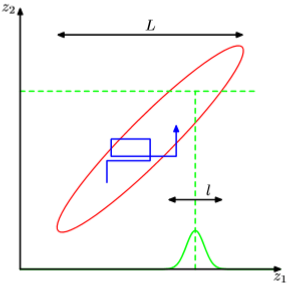
Gibbs sampling is a type of random walk through parameter space, and can be considered as a Metroplis-Hastings algorithm with a special proposal distribution.

At each iteration, we are drawing from conditional posterior probabilities. This means that the proposal move is always accepted. Hence, if we can draw samples from the conditional distributions, Gibbs sampling can be much more efficient than regular Metropolis-Hastings.

Properties of Gibbs Sampling

Properties

- No need to tune the proposal distribution;
- Good trade-off between acceptance and mixing: Acceptance ratio is always 1.
- Need to be able to derive conditional probability distributions.

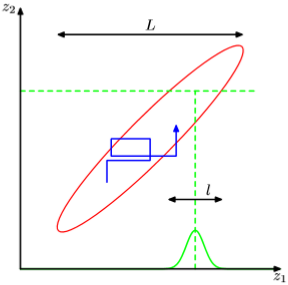


Acceleration of Gibbs sampling: (given $p(a, b, c)$ draw samples from a and c)

Properties of Gibbs Sampling

Properties

- No need to tune the proposal distribution;
- Good trade-off between acceptance and mixing: Acceptance ratio is always 1.
- Need to be able to derive conditional probability distributions.



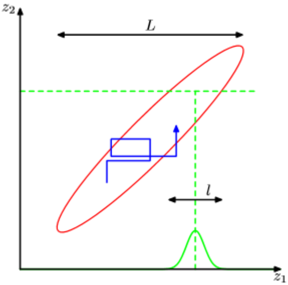
Acceleration of Gibbs sampling: (given $p(a, b, c)$ draw samples from a and c)

- **Blocked Gibbs:**
  (1) Draw (a,b) given c;
  (2) Draw c given (a,b);

Properties of Gibbs Sampling

Properties

- No need to tune the proposal distribution;
- Good trade-off between acceptance and mixing: Acceptance ratio is always 1.
- Need to be able to derive conditional probability distributions.



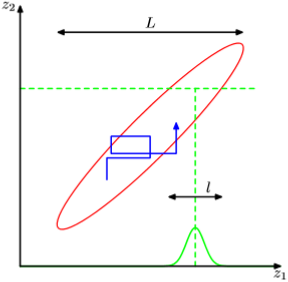
Acceleration of Gibbs sampling: (given $p(a, b, c)$ draw samples from a and c)

- **Blocked Gibbs:**
  (1) Draw (a,b) given c;
  (2) Draw c given (a,b);
- **Collapsed Gibbs:**
  (1) Draw a given c;
  (2) Draw c given a;

Properties of Gibbs Sampling

Properties

- No need to tune the proposal distribution;
- Good trade-off between acceptance and mixing: Acceptance ratio is always 1.
- Need to be able to derive conditional probability distributions.



Acceleration of Gibbs sampling: (given $p(a, b, c)$ draw samples from a and c)

- **Blocked Gibbs:**
  (1) Draw (a,b) given c;
  (2) Draw c given (a,b);
- **Collapsed Gibbs:**
  (1) Draw a given c;
  (2) Draw c given a;
- Marginalize whenever you can.

Outline

1 Monte Carlo Method

2 Markov Chain Monte Carlo
- MCMC Sampling Algorithm
- Metropolis-Hastings Algorithm
- Gibbs Sampling
- Latent Dirichlet Allocation

Topic modeling for text

Latent Dirichlet Allocation (LDA)

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

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

An example article from a corpus.
Each color codes a different topic.

Topic modeling for text

Latent Dirichlet Allocation (LDA)

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

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

An example article from a corpus. Each color codes a different topic.

- There are many models which can be used to represent text data, such as LSA, PLSA, LDA, word2vec, etc.

Topic modeling for text

Latent Dirichlet Allocation (LDA)

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

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

An example article from a corpus. Each color codes a different topic.

- There are many models which can be used to represent text data, such as LSA, PLSA, LDA, word2vec, etc.
- LDA models text in a simple and reasonable manner;

Topic modeling for text

Latent Dirichlet Allocation (LDA)

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

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

An example article from a corpus. Each color codes a different topic.

- There are many models which can be used to represent text data, such as LSA, PLSA, LDA, word2vec, etc.
- LDA models text in a simple and reasonable manner;
- LDA can be applied to many complex applications, such as image, graph, location, etc.

Notations for LDA

symbol	meaning
M	the number of documents
N_m	the number of words in document m
K	the number of topics
$w_{m,n}$	the index of word n in document m
$z_{m,n}$	the topic k of each word $w_{m,n}$
α, β	fixed hyper-parameters
θ	topic distribution for each document
ϕ	topic distribution for each word

Properties of Dirichlet

$$Dir(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_i)}{\Pi_{k=1}^{K}\Gamma(\alpha_k)}\Pi_{k=1}^{K}\theta_k^{\alpha_k - 1} \equiv \frac{1}{\triangle(\alpha)}\Pi_{k=1}^{K}\theta_k^{\alpha_k - 1}$$

$$Mult(m_1, \cdots, m_K|\theta, N) = \binom{N}{m_1 m_2 \cdots m_K}\Pi_{k=1}^{K}\theta_k^{m_k}$$

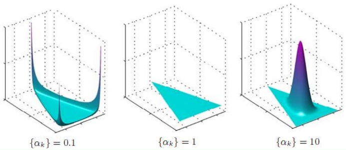
$$Dir(\theta|\mathcal{D}, \alpha) = Dir(\theta|\alpha, m) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_i) + N)}{\Pi_{k=1}^{K}\Gamma(\alpha_k + m_k)}\Pi_{k=1}^{K}\theta_k^{\alpha_k + m_k - 1}$$

Properties of Dirichlet

$$Dir(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_i)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} \equiv \frac{1}{\triangle(\alpha)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$

$$Mult(m_1,\cdots,m_K|\theta,N) = \binom{N}{m_1 m_2 \cdots m_K}\prod_{k=1}^{K}\theta_k^{m_k}$$

$$Dir(\theta|\mathcal{D},\alpha) = Dir(\theta|\alpha,m) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_i)+N)}{\prod_{k=1}^{K}\Gamma(\alpha_k+m_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k+m_k-1}$$

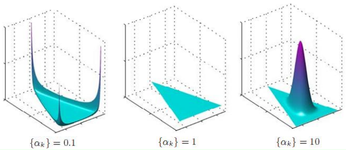


Properties of Dirichlet

$$Dir(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_i)}{\Pi_{k=1}^{K}\Gamma(\alpha_k)} \Pi_{k=1}^{K} \theta_k^{\alpha_k - 1} \equiv \frac{1}{\triangle(\alpha)} \Pi_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

$$Mult(m_1, \cdots, m_K|\theta, N) = \binom{N}{m_1 m_2 \cdots m_K} \Pi_{k=1}^{K} \theta_k^{m_k}$$

$$Dir(\theta|\mathcal{D}, \alpha) = Dir(\theta|\alpha, m) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_i) + N)}{\Pi_{k=1}^{K}\Gamma(\alpha_k + m_k)} \Pi_{k=1}^{K} \theta_k^{\alpha_k + m_k - 1}$$

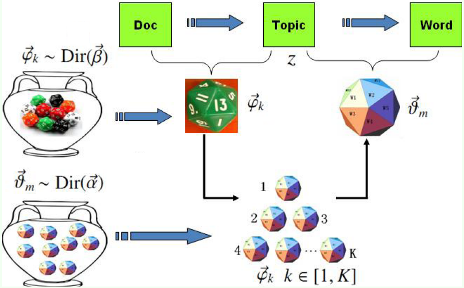


The expectation of Dirichlet is $E(\theta) = (\frac{\alpha_1}{\alpha_0}, \frac{\alpha_2}{\alpha_0}, \cdots, \frac{\alpha_K}{\alpha_0})$, where $\alpha_0 = \sum_{k=1}^{K} \alpha_k$.

LDA: Latent Dirichlet Allocation

LDA assumes the following generative process for each document d in a corpus $\mathbb{D}$:

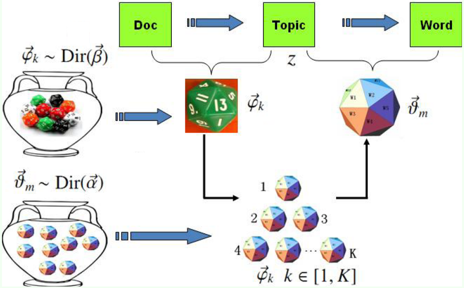
1: for $k = 1$ to K do
2: $\quad \phi^{(k)} \sim Dirichlet(\beta)$;
3: for each document $m \in \mathbb{D}$
4: $\quad \theta_m \sim Dirichlet(\alpha)$;
5: $\quad$ for each word $w_{m,n} \in m$
6: $\quad\quad z_{m,n} \sim Mult(\theta_m)$;
7: $\quad\quad w_{m,n} \sim Mult(\phi^{(z_{m,n})})$;



LDA: Latent Dirichlet Allocation

LDA assumes the following generative process for each document d in a corpus $\mathbb{D}$:

1: for $k = 1$ to K do
2: $\quad \phi^{(k)} \sim Dirichlet(\beta)$;
3: for each document $m \in \mathbb{D}$
4: $\quad \theta_m \sim Dirichlet(\alpha)$;
5: $\quad$ for each word $w_{m,n} \in m$
6: $\qquad z_{m,n} \sim Mult(\theta_m)$;
7: $\qquad w_{m,n} \sim Mult(\phi^{(z_{m,n})})$;



where $\phi^{(k)} \in R^K$ and $\theta_m \in R^{|V|}$.

Joint probability of LDA model

The joint probabilities of observing a word $w_{m,n}$ and the corpus are

$$
\begin{aligned}
&p(w_{m,n}, z_{m,n}, \phi, \theta_m | \alpha, \beta) \\
&= p(w_{m,n} | z_{m,n}, \phi) p(z_{m,n} | \theta_m) p(\phi | \beta) p(\theta_m | \alpha),
\end{aligned}
$$

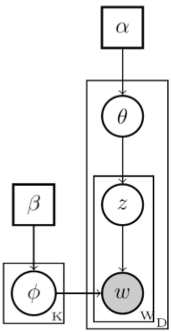
Joint probability of LDA model

The joint probabilities of observing a word $w_{m,n}$ and the corpus are

$$p(w_{m,n}, z_{m,n}, \phi, \theta_m|\alpha, \beta)$$
$$= p(w_{m,n}|z_{m,n}, \phi)p(z_{m,n}|\theta_m)p(\phi|\beta)p(\theta_m|\alpha),$$

In other words,

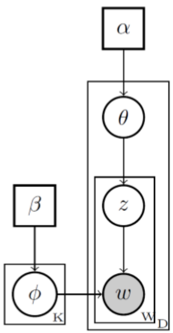
$$p(w, z, \phi, \theta|\alpha, \beta)$$
$$= \Pi_{m=1}^{M}\Pi_{n=1}^{N} p(w_{m,n}, z_{m,n}, \phi, \theta_m|\alpha, \beta)$$
$$= p(\phi|\beta)\Pi_{m=1}^{M} p(\theta_m|\alpha)\Pi_{n=1}^{N} p(w_{m,n}|z_{m,n}, \phi)p(z_{m,n}|\theta_m).$$



Joint probability of LDA model

The joint probabilities of observing a word $w_{m,n}$ and the corpus are

$$p(w_{m,n}, z_{m,n}, \phi, \theta_m | \alpha, \beta)$$
$$= p(w_{m,n}|z_{m,n}, \phi) p(z_{m,n}|\theta_m) p(\phi|\beta) p(\theta_m|\alpha),$$



In other words,

$$p(w, z, \phi, \theta|\alpha, \beta)$$
$$= \Pi_{m=1}^{M} \Pi_{n=1}^{N} p(w_{m,n}, z_{m,n}, \phi, \theta_m|\alpha, \beta)$$
$$= p(\phi|\beta) \Pi_{m=1}^{M} p(\theta_m|\alpha) \Pi_{n=1}^{N} p(w_{m,n}|z_{m,n}, \phi) p(z_{m,n}|\theta_m).$$

LDA Model II

for $k = 1$ to K do
    $\phi^{(k)} \sim Dirichlet(\beta)$;
for each document $m \in \mathbb{D}$
    $\theta_m \sim Dirichlet(\alpha)$;
    for each word $w_{m,n} \in m$
        $z_{m,n} \sim Mult(\theta_m)$;
for each topic $k \in [1, K]$
    for each $z_{m,n} = k$
        $w_{m,n} \sim Mult(\phi^{(k)})$;

LDA Model II

1: for $k = 1$ to K do
2: $\quad \phi^{(k)} \sim Dirichlet(\beta)$;
3: for each document $m \in \mathbb{D}$
4: $\quad \theta_m \sim Dirichlet(\alpha)$;
5: $\quad$ for each word $w_{m,n} \in m$
6: $\quad\quad z_{m,n} \sim Mult(\theta_m)$;
7: for each topic $k \in [1, K]$
8: $\quad$ for each $z_{m,n} = k$
9: $\quad\quad w_{m,n} \sim Mult(\phi^{(k)})$;

We put the words with the same topic together. We have

$$\mathbf{z} = (z_1, z_2, \cdots, z_K),$$
$$\mathbf{w} = (w_1, w_2, \cdots, w_K),$$

where w_k is the set of words generated by the k-th topic, and z_k is a vector whose terms are the IDs of the word topics (k).

LDA Model II

```
for k = 1 to K do
    φ^(k) ~ Dirichlet(β);
for each document m ∈ 𝔻
    θ_m ~ Dirichlet(α);
    for each word w_{m,n} ∈ m
        z_{m,n} ~ Mult(θ_m);
for each topic k ∈ [1, K]
    for each z_{m,n} = k
        w_{m,n} ~ Mult(φ^(k));
```

We put the words with the same topic together. We have

$$\mathbf{z} = (z_1, z_2, \cdots, z_K),$$
$$\mathbf{w} = (w_1, w_2, \cdots, w_K),$$

where w_k is the set of words generated by the k-th topic, and z_k is a vector whose terms are the IDs of the word topics (k).

Now, we have two conjugate structures of Dirichlet-Multinomial:

$$\underbrace{\alpha \longrightarrow}_{Dirichlet} \theta_m \underbrace{\longrightarrow z_m}_{Multinomial}, \text{ and } \underbrace{\beta \longrightarrow}_{Dirichlet} \phi_k \underbrace{\longrightarrow w_k}_{Multinomial} \tag{9}$$

Dice Toss Toy Example

Suppose we have a dice of K sides. We toss the dice and the probability of landing on side k is $p(t = k|f) = f_i$. We throw the dice N times and obtain a set of results $s = \{s_1, s_2, \cdots, s_N\}$. The joint probability is

Dice Toss Toy Example

Suppose we have a dice of K sides. We toss the dice and the probability of landing on side k is $p(t = k|f) = f_i$. We throw the dice N times and obtain a set of results $s = \{s_1, s_2, \cdots, s_N\}$. The joint probability is

$$p(s|f) = \prod_{n=1}^{N} p(s_n|f) = f_1^{n_1} f_2^{n_2} \cdots f_K^{n_K} = \prod_{i=1}^{K} f_i^{n_i} \tag{10}$$

Dice Toss Toy Example

Suppose we have a dice of K sides. We toss the dice and the probability of landing on side k is $p(t = k|f) = f_i$. We throw the dice N times and obtain a set of results $s = \{s_1, s_2, \cdots, s_N\}$. The joint probability is

$$p(s|f) = \prod_{n=1}^{N} p(s_n|f) = f_1^{n_1} f_2^{n_2} \cdots f_K^{n_K} = \prod_{i=1}^{K} f_i^{n_i} \tag{10}$$

Suppose that f is a Dirichlet distribution with α as hyper-parameter. Then we express the probability of f as

$$Dir(f|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} f_k^{\alpha_k - 1} \tag{11}$$

Example Cont'd

If we want to estimate the parameter f based on the observation of s, then we can express f in the following manner

Example Cont'd

If we want to estimate the parameter f based on the observation of s, then we can express f in the following manner

$$p(f|s,\alpha) = \frac{p(s|f,\alpha)p(f|\alpha)}{\int_0^1 p(s|f,\alpha)p(f|\alpha)df}$$

$$= \frac{\prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k - 1}}{\int_0^1 \prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k - 1} df}$$

$$= \frac{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{n_k + \alpha_k - 1}}{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int_0^1 \prod_{k=1}^K f_k^{n_k + \alpha_k - 1} df}$$

$$= \frac{\Gamma(\sum_{k=1}^K (n_k + \alpha_k))}{\prod_{k=1}^K \Gamma(n_k + \alpha_k)} \prod_{k=1}^K f_k^{n_k + \alpha_k - 1}$$

Example Cont'd

If we want to estimate the parameter f based on the observation of s, then we can express f in the following manner

$$
\begin{aligned}
p(f|s,\alpha) &= \frac{p(s|f,\alpha)p(f|\alpha)}{\int_0^1 p(s|f,\alpha)p(f|\alpha)df} \\
&= \frac{\prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k-1}}{\int_0^1 \prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k-1} df} \\
&= \frac{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{n_k+\alpha_k-1}}{\frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int_0^1 \prod_{k=1}^K f_k^{n_k+\alpha_k-1} df} \\
&= \frac{\Gamma(\sum_{k=1}^K (n_k+\alpha_k))}{\prod_{k=1}^K \Gamma(n_k+\alpha_k)} \prod_{k=1}^K f_k^{n_k+\alpha_k-1}
\end{aligned}
$$

Notice that after estimating f based on s observations, f is still a Dirichlet distribution with parameter $\alpha + \mathbf{n}$, where $\mathbf{n} = (n_1, n_2, \cdots, n_k)$. This property is known as conjugate priors. Based on this property, estimating the parameters f_i after observing N trials is a simple counting procedure.

Estimating f_i

Suppose we want to obtain f_i from $f = (f_1, f_2, \cdots, f_{i-1}, f_i, f_{i+1}, \cdots, f_K)$.

$$
\begin{aligned}
E(f_i|s,\alpha) &= \int_0^1 f_i p(f|s,\alpha)df \\
&= \int_0^1 f_i \frac{\Gamma(\sum_{k=1}^K (n_k+\alpha_k))}{\prod_{k=1}^K \Gamma(n_k+\alpha_k)} \prod_{k=1}^K f_k^{n_k+\alpha_k-1} df \\
&= \frac{\Gamma(\sum_{k=1}^K (n_k+\alpha_k))}{\prod_{k=1}^K \Gamma(n_k+\alpha_k)} \int_0^1 f_i \prod_{k=1}^K f_k^{n_k+\alpha_k-1} df \\
&= \frac{\Gamma(\sum_{k=1}^K (n_k+\alpha_k))}{\prod_{k=1}^K \Gamma(n_k+\alpha_k)} \frac{\Gamma(n_i+\alpha_i+1)\prod_{k=1,k\neq i}^K \Gamma(n_i+\alpha_i)}{\Gamma(n_i+\alpha_i+1+\sum_{k=1,k\neq i}^K (n_k+\alpha_k))} \\
&= \frac{n_i+\alpha_i}{\sum_{k=1}^K (n_k+\alpha_k)}
\end{aligned}
$$

Likelihood of Observing s_i

Suppose we want to obtain the likelihood of observing s_i, i.e., $p(s_i|\alpha)$.

$$
\begin{aligned}
p(s_i|\alpha) &= \int_0^1 p(s_i, f|\alpha)df == \int_0^1 p(s_i|f)p(f|\alpha))df \\
&= \int_0^1 \prod_{i=1}^K f_i^{n_i} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K f_k^{\alpha_k - 1} df \\
&= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \int_0^1 \prod_{k=1}^K f_k^{n_k + \alpha_k - 1} df \\
&= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \frac{\prod_{k=1}^K \Gamma(n_k + \alpha_k)}{\Gamma(\sum_{k=1}^K (n_k + \alpha_k))} = \frac{\triangle(\mathbf{n} + \alpha)}{\triangle(\alpha)}.
\end{aligned}
$$

where $\triangle(\alpha) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}$.

Parameter Inference

We integrate θ and ϕ to obtain the following:
$p(z, w|\alpha, \beta) = p(w|z, \beta)p(z|\alpha)$

Parameter Inference

We integrate θ and ϕ to obtain the following:
$p(z, w|\alpha, \beta) = p(w|z, \beta)p(z|\alpha)$

$\underbrace{\beta \longrightarrow}_{\text{Dirichlet}} \phi_k \underbrace{\longrightarrow w_k}_{\text{Multinomial}}$

$$\begin{aligned} p(w|z,\beta) &= \prod_{k=1}^{K} p(w_k|z_k,\beta) \\ &= \prod_{k=1}^{K} \frac{\triangle(\mathbf{n}_k+\beta)}{\triangle(\beta)}, \end{aligned}$$

where $\mathbf{n}_k = (n_k^{(1)}, n_k^{(2)}, \cdots, n_k^{(V)})$, and $n_k^{(v)}$ is the number of words generated by topic k.

Parameter Inference

We integrate θ and ϕ to obtain the following:
$p(z, w|\alpha, \beta) = p(w|z, \beta)p(z|\alpha)$

$$\underbrace{\beta \longrightarrow}_{Dirichlet} \phi_k \underbrace{\longrightarrow w_k}_{Multinomial}$$

$$\begin{aligned} p(w|z,\beta) &= \prod_{k=1}^{K} p(w_k|z_k,\beta) \\ &= \prod_{k=1}^{K} \frac{\triangle(\mathbf{n}_k+\beta)}{\triangle(\beta)}, \end{aligned}$$

where $\mathbf{n}_k = (n_k^{(1)}, n_k^{(2)}, \cdots, n_k^{(V)})$, and $n_k^{(v)}$ is the number of words generated by topic k.

$$\underbrace{\alpha \longrightarrow}_{Dirichlet} \theta_m \underbrace{\longrightarrow z_m}_{Multinomial}$$

$$\begin{aligned} p(z|\alpha) &= \prod_{m=1}^{M} p(z_m|\alpha) \\ &= \prod_{m=1}^{M} \frac{\triangle(\mathbf{n}_m+\alpha)}{\triangle(\alpha)}, \end{aligned}$$

where $\mathbf{n}_m = (n_m^{(1)}, n_m^{(2)}, \cdots, n_m^{(K)})$, and $n_m^{(k)}$ is # words with topic k in the m-th document.

Parameter Inference Cont'd

$p(z|\alpha)$

$$
\begin{aligned}
p(z|\alpha) &= \int p(z,\theta|\alpha)d\theta = \int p(z|\theta,\alpha)p(\theta|\alpha)d\theta \\
&= \int p(z|\theta)p(\theta|\alpha)d\theta = \int \prod_{m=1}^{M}\Big(\prod_{k=1}^{K}\theta_{m,k}^{n_{m,k}}\Big)\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\Big(\prod_{k=1}^{K}\theta_{m,k}^{\alpha_k-1}\Big)d\theta \\
&= \int \prod_{m=1}^{M}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_{m,k}^{n_{m,k}+\alpha_k-1}d\theta \\
&= \prod_{m=1}^{M}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\int\prod_{k=1}^{K}\theta_{m,k}^{n_{m,k}+\alpha_k-1}d\theta \\
&= \prod_{m=1}^{M}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_{m,k})}{\Gamma(\sum_{k=1}^{K}\alpha_k+n_{m,k})} = \prod_{m=1}^{M}\frac{\triangle(\mathbf{n}_m+\alpha)}{\triangle(\alpha)}
\end{aligned}
$$

Parameter Inference Cont'd

$p(w|z$

$$
\begin{aligned}
p(w|z,\beta) &= \int p(w,\phi|z,\beta)d\phi = \int p(w|z,\beta,\phi)p(\phi|z,\beta)d\phi \\
&= \int p(w|z,\phi)p(\phi|\beta)d\phi \\
&= \int \prod_{k=1}^{K}\Big(\prod_{v=1}^{V}\phi_{k,v}^{n_{k,v}}\Big)\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\Big(\prod_{v=1}^{V}\phi_{k,v}^{\beta_v-1}\Big)d\phi \\
&= \int \prod_{k=1}^{K}\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\prod_{v=1}^{V}\phi_{k,v}^{n_{k,v}+\beta_v-1}d\phi \\
&= \prod_{k=1}^{K}\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)}\frac{\prod_{v=1}^{V}\Gamma(\beta_v+n_{k,v})}{\Gamma(\sum_{v=1}^{V}\beta_v+n_{k,v})} = \prod_{k=1}^{K}\frac{\triangle(\mathbf{n}_k+\beta)}{\triangle(\beta)}
\end{aligned}
$$

Gibbs Sampling

Analysis

For simplicity, the topic of the i-th word in the corpus denotes z_i, where $i = (m, n)$. In terms of Gibbs sampling, we need to compute the conditional probability $p(z_i = k|\mathbf{z}_{-i}, \mathbf{w})$.

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}) = p(z_i = k|\mathbf{z}_{-i}, \mathbf{w}_{-i}, w_i = t)$$
$$= \frac{p(z_i = k, w_i = t|\mathbf{z}_{-i}, \mathbf{w}_{-i})}{p(w_i = t|\mathbf{z}_{-i}, \mathbf{w}_{-i})} \propto p(z_i = k, w_i = t|\mathbf{z}_{-i}, \mathbf{w}_{-i}).$$

Notice that $z_i = k, w_i = t$ only involves the $m-$th document and the $k-$th topic, which are related to two Dirichlet-Multinomial (DM) structures, and is independent to $M + K - 2$ DM structures.

$$p(\theta_m|\mathbf{z}_{-i}, \mathbf{w}_{-i}) = Dir(\theta_m|\mathbf{n}_{m,-i} + \alpha)$$
$$p(\phi_k|\mathbf{z}_{-i}, \mathbf{w}_{-i}) = Dir(\phi_k|\mathbf{n}_{k,-i} + \beta)$$

Deriving the Transition Probability

Transition probability

$$p(z_i = k, w_i = t|\mathbf{z}_{-i}, \mathbf{w}_{-i}) = \int p(z_i = k, w_i = t, \theta_m, \phi_k|\mathbf{z}_{-i}, \mathbf{w}_{-i}) d\theta_m d\phi_k$$

$$\int p(z_i = k, \theta_m|\mathbf{z}_{-i}, \mathbf{w}_{-i}) p(w_i = t, \phi_k|\mathbf{z}_{-i}, \mathbf{w}_{-i}) d\theta_m d\phi_k$$

$$= \int p(z_i = k|\theta_m) Dir(\theta_m|\mathbf{n}_{m,-i} + \alpha) d\theta_m$$

$$\cdot \int p(w_i = t|\phi_k) Dir(\phi_k|\mathbf{n}_{k,-i} + \beta) d\phi_k$$

$$= \int \theta_{mk} Dir(\theta_m|\mathbf{n}_{m,-i} + \alpha) d\theta_m \int \phi_{kt} Dir(\phi_k|\mathbf{n}_{k,-i} + \beta) d\phi_k$$

$$= E(\theta_{mk}) E(\phi_{kt}) = \widehat{\theta}_{mk} \widehat{\phi}_{kt},$$

where $\widehat{\theta}_{mk} = \frac{n_{m,-i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}(n_{m,-i}^{(k)} + \alpha_k)}$ and $\widehat{\phi}_{kt} = \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{V}(n_{k,-i}^{(t)} + \beta_t)}$.

Take-home messages

- Monte Carlo method
- Markov Chain Monte Carlo
  - MCMC sampling algorithm
  - Metropolis-Hastings algorithm
  - Gibbs sampling
  - Latent Dirichlet Allocation