# LDA 概述

—— watkins.song

在详细的讲LDA之前， 先需要了解一些基本概念。

## 1. MCMC and Gibbs Sampling

reference:

http://www.52nlp.cn/lda-math-mcmc-%E5%92%8C-gibbs-sampling1

MC(蒙特卡罗方法(Monte Carlo Simulation)) 是统计模拟的方法， 用来进行在连续的概率分布函数中进行采样模拟统计分布。 MC针对的问题是给定一个概率密度函数 p(x) 如何在计算机中生成这个概率密度函数的样本。

当p(x) 不是很复杂的时候， 可以采用**[Box-Muller 变换]** 将p(x)转换到均匀分布然后采用线性同余发生器获取样本， 但是当p(x)非常复杂的时候， 就需要别的方法。

当p(x)比较复杂的时候， 或者p(x)是高维分布的情况， 例如p(x)是以下情况时：

- $p(x)=\tilde{p}(x)/\int\tilde{p}(x)dx$,而 $\tilde{p}(x)$ 我们是可以计算的，但是底下的积分式无法显式计算。
- $p(x,y)$ 是一个二维的分布函数，这个函数本身计算很困难，但是条件分布 $p(x|y),p(y|x)$的计算相对简单;如果 $p(\mathbf{x})$ 是高维的，这种情形就更加明显。

此时就需要使用一些更加复杂的随机模拟的方法来生成样本。MCMC(Markov Chain Monte Carlo) 和 Gibbs Sampling算法就是最常用的一种，这两个方法在现代贝叶斯分析中被广泛使用。

**总结： 因为有些概率密度函数比较难计算， 所以需要采用MCMC或者Gibbs Sampling这些模拟统计的方法进行采样计算概率密度函数， 降低概率密度函数的计算复杂度。**

**MCMC在复杂的概率分布中用于进行样本的生成，当概率分布函数比较复杂的时候，不好计算的时候，就需要通过MCMC进行样本的生成，计算概率分布函数。**

**对于连续的概率分布，不可能在计算机中表示所有的样本，也只能通过MCMC进行有限数量的样本采样。**

## 2. Markov Chain

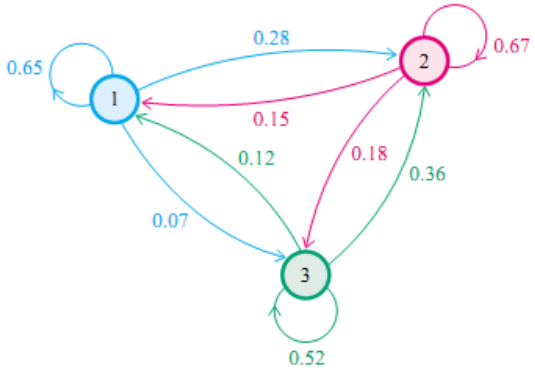
马氏链的数学定义很简单:

$$P(X_{t+1} = x|X_t, X_{t-1}, \cdots) = P(X_{t+1} = x|X_t)$$

也就是状态转移的概率只依赖于前一个状态。

我们先来看马氏链的一个具体的例子。社会学家经常把人按其经济状况分成3类：下层(lower-class)、中层(middle-class)、上层(upper-class)，我们用1,2,3 分别代表这三个阶层。社会学家们发现决定一个人的收入阶层的最重要的因素就是其父母的收入阶层。如果一个人的收入属于下层类别，那么他的孩子属于下层收入的概率是 0.65, 属于中层收入的概率是 0.28, 属于上层收入的概率是 0.07。事实上，从父代到子代，收入阶层的变化的转移概率如下:

| | | 子代 | | |
|---|---|---|---|---|
| | State | 1 | 2 | 3 |
| 父代 | 1 | 0.65 | 0.28 | 0.07 |
| | 2 | 0.15 | 0.67 | 0.18 |
| | 3 | 0.12 | 0.36 | 0.52 |




使用矩阵的表示方式，转移概率矩阵记为

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

假设当前这一代人处在下层、中层、上层的人的比例是概率分布向量 $\pi_0 = [\pi_0(1), \pi_0(2), \pi_0(3)]$，那么他们的子女的分布比例将是 $\pi_1 = \pi_0 P$，他们的孙子代的分布比例将是 $\pi_2 = \pi_1 P = \pi_0 P^2$, ....., 第$n$代子孙的收入分布比例将是 $\pi_n = \pi_{n-1} P = \pi_0 P^n$。

假设初始概率分布为$\pi_0$=[0.21,0.68,0.11]，则我们可以计算前$n$代人的分布状况如下:

| 第$n$代人 | 下层 | 中层 | 上层 |
|---|---|---|---|
| 0 | 0.210 | 0.680 | 0.110 |
| 1 | 0.252 | 0.554 | 0.194 |
| 2 | 0.270 | 0.512 | 0.218 |
| 3 | 0.278 | 0.497 | 0.225 |
| 4 | 0.282 | 0.490 | 0.226 |
| 5 | 0.285 | 0.489 | 0.225 |
| 6 | 0.286 | 0.489 | 0.225 |
| 7 | 0.286 | 0.489 | 0.225 |
| 8 | 0.289 | 0.488 | 0.225 |
| 9 | 0.286 | 0.489 | 0.225 |
| 10 | 0.286 | 0.489 | 0.225 |
| ... | ... | ... | ... |

我们发现，当 $n$ 足够大的时候，这个$P^n$矩阵的每一行都是稳定地收敛到$\pi =$ $[0.286,0.489,0.225]$ 这个概率分布。自然的，这个收敛现象并非是我们这个马氏链独有的，而是绝大多数马氏链的共同行为。

**所有的 MCMC(Markov Chain Monte Carlo) 方法都是以这个定理（Markov Chain定理）作为理论基础的。**

从初始概率分布 $\pi_0$ 出发，我们在马氏链上做状态转移，记$X_i$的概率分布为$\pi_i$, 则有

$$X_0 \sim \pi_0(x)$$

$$X_i \sim \pi_i(x) \qquad \pi_i(x)=\pi_{i-1}(x)P=\pi_0(x)P^n$$

由马氏链收敛的定理, 概率分布$\pi_i(x)$将收敛到平稳分布 $\pi(x)$。假设到第$n$步的时候马氏链收敛，则有

$$
\begin{aligned}
X_0 &\sim \pi_0(x) \\
X_1 &\sim \pi_1(x) \\
&\cdots \\
X_n &\sim \pi_n(x) = \pi(x) \\
X_{n+1} &\sim \pi(x) \\
X_{n+2} &\sim \pi(x) \\
&\cdots
\end{aligned}
$$

所以 $X_n, X_{n+1}, X_{n+2}, \cdots \sim \pi(x)$ 都是同分布的随机变量，当然他们并不独立。如果我们从一个具体的初始状态 $x_0$ 开始,沿着马氏链按照概率转移矩阵做跳转，那么我们得到一个转移序列 $x_0, x_1, x_2, \cdots x_n, x_{n+1} \cdots$, 由于马氏链的收敛行为， $x_n, x_{n+1}, \cdots$ 都将是平稳分布 $\pi(x)$ 的样本。

总结： Markov Chain 是一个状态转移的公式， 并且状态转移最终会实现收敛， 就像PageRank算法的网页相关性矩阵最终收敛差不多。

利用Markov Chain收敛的特性，对于不好计算的概率分布，可以先随机的选定一些样本，然后随机的给样本赋值一个概率分布，然后进行Markov Chain的收敛计算，直到收敛后，就可以得到真实的概率分布。

采用Markov Chain 计算概率的主要原因是由于某些概率分布直接计算不能实现，那么就可以随机的采样一些样本点，然后给每个样本点分配一个随机的概率，并采用Markov Chain方式进行状态转移， 直到概率分布收敛。（最终得到的就是概率分布的近似真实分布）

Markov Chain 最终会达到收敛状态，由于马氏链的收敛行为， $x_n, x_{n+1}, \cdots$ 都将是平稳分布 $\pi(x)$ 的样本。

对于给定的概率分布$p(x)$, 我们希望能有便捷的方式生成它对应的样本。由于马氏链能收敛到平稳分布， 于是一个很的漂亮想法是：如果我们能构造一个转移矩阵为$P$的马氏链，使得该马氏链的平稳分布恰好是$p(x)$, 那么我们从任何一个初始状态$x_0$出发沿着马氏链转移, 得到一个转移序列 $x_0,x_1,x_2,\cdots x_n,x_{n+1}\cdots$,， 如果马氏链在第$n$步已经收敛了，于是我们就得到了$p(x)$的样本$x_n,x_{n+1}\cdots$。

MCMC是基于Markov Chain的采样方法，需要利用Markov Chain的原理，使得状态转移达到收敛的状态，也就是在某个样本采样之后所有的样本都服从同一个概率分布。

由上一节的例子和定理我们看到了，马氏链的收敛性质主要由转移矩阵$P$决定, 所以基于马氏链做采样的**关键问题是如何构造转移矩阵$P$**, 使得平稳分布恰好是我们要的分布$p(x)$。

剧透： 在LDA中为什么要使用Gibbs Sampling? 因为在LDA中根据M篇文档以及每篇文档的单词的topic分布， 计算了一个联合概率公式：

$$p(\vec{\mathbf{w}},\vec{\mathbf{z}}|\vec{\alpha},\vec{\beta}) = p(\vec{\mathbf{w}}|\vec{\mathbf{z}},\vec{\beta})p(\vec{\mathbf{z}}|\vec{\alpha})$$
$$= \prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\vec{\beta})}{\Delta(\vec{\beta})}\prod_{m=1}^{M}\frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}$$

这个联合概率公式是非常难计算的， 所以需要采用Gibbs Sampling算法进行采样计算这个联合概率公式。 在这个公式中， 那么Markov Chain的转换， 是要让谁达到收敛状态呢？ 因为 $\vec{\mathbf{w}}$ 是已经观测到的样本， 所以不要计算， $\vec{\mathbf{z}}$ 是初始计算时随机分配的topic， 所以最终的目的是要达到 $\vec{\mathbf{z}}$ 实现收敛， 这样在LDA的plate model中的每篇文档对应

的topic 分布 $\vec{\theta}_m$ 和每个topic下对应的word的分布 $\vec{\varphi}_k$ 就自然也是收敛的了， 就可以通过 $\vec{\mathbf{z}}$ 计算LDA模型的两种多项分布的参数。

## 3. 继续MCMC and Gibbs Sampling

**Algorithm 5** MCMC 采样算法

1: 初始化马氏链初始状态$X_0 = x_0$

2: 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$X_{t+1} = x_t$

MCMC中存在一个称为接受率的概念， $\alpha(i,j)$， 一般$\alpha(i,j)$表示从状态i 转移到状态 j.

采样$y \sim q(x|x_t)$ 的意思是，有一个概率分布$q(x|x_t)$, 这个概率分布是受第t个时刻马氏链的状态影响的，所以为条件概率分布，然后从这个条件概率分布中选取随机值作为采样y。

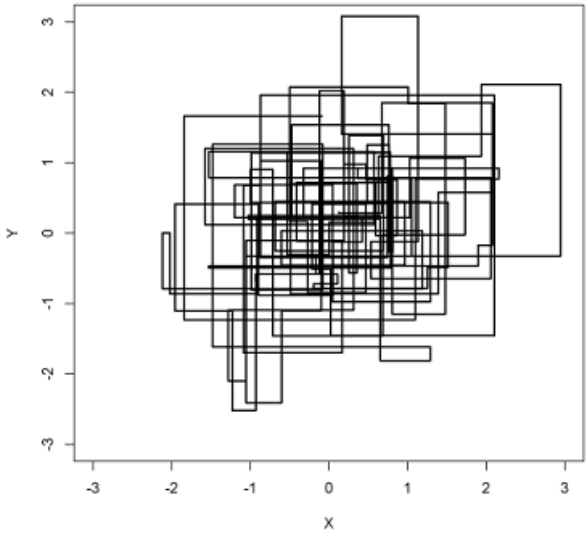
对于高维的情形，由于接受率 $\alpha$的存在(通常 $\alpha$<1), 以上 MCMC及变种算法的效率不够高。所以可以采用Gibbs Sampling提高状态转换效率以达到最快的实现收敛的概率分布状态。

**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. $x_{t+1} \sim p(x|y_{t+1})$




**二维Gibbs Sampling 算法中的马氏链转移**

Gibbs Sampling 算法大都是坐标轴轮换采样的，但是这其实是不强制要求的。最一般的情形可以是，在$t$时刻，可以在$x$轴和$y$轴之间随机的选一个坐标轴，然后按条件概率做转移，马氏链也是一样收敛的。

将2维的Gibbs Sampling算法扩展到N维的采样算法， Gibbs Sampling过程中状态转换只是沿着一个坐标轴， 也就是说只沿着一个随机变量进行状态转换， 所以$n$维空间中对于概率分布 $p(x_1,x_2,\cdots,x_n)$ 可以如下定义转移矩阵：

1. 如果当前状态为$(x_1,x_2,\cdots,x_n)$，马氏链转移的过程中，只能沿着坐标轴做转移。沿着 $x_i$这根坐标轴做转移的时候，转移概率由条件概率 $p(x_i|x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)$ 定义；
2. 其它无法沿着单根坐标轴进行的跳转，转移概率都设置为 0。

**Algorithm 8** n维Gibbs Sampling 算法

1: 随机初始化$\{x_i : i = 1,\cdots,n\}$

2: 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. $x_n^{(t+1)} \sim p(x_n|x_1^{(t+1)},x_2^{t},\cdots,x_{n-1}^{(t+1)})$

以上算法收敛后，得到的就是符合概率分布$p(x_1,x_2,\cdots,x_n)$的样本，当然这些样本并不独立，但是我们此处要求的是采样得到的样本符合给定的概率分布，并不要求独立。

这里不是很理解，采样算法收敛之后应该得到的是概率分布的真实分布，怎么会有样本么？根本就没有任何采样的样本。。。$x_n^t$ 表示的是状态t时刻的第n个维度的概率分布，则Markov Chain收敛了以后，就是获得的概率分布的样本呢。

1. $x_1^{(t+1)} \sim p(x_1|x_2^{(t)},x_3^{(t)},\cdots,x_n^{(t)})$ 这个公式中计算了Gibbs 采样中都某一个维度进行采样的过程，其中 $p(x_1|x_2^{(t)},x_3^{(t)},\cdots,x_n^{(t)})$ 这个条件概率公式根据减去了$x_1$的概率分布计算了一个新的$x_1$的概率分布，然后采用随机数生成器在概率计算得到概率分布中进行采样，获得 $x_1$的新的采样， 也就是在LDA模型中某个单词对应的新的topic值。

当Gibbs采样收敛以后，一次循环计算了所有的维度的采样，所有的维度的采样就对应一个采样样本。

总结 & 剧透：在Gibbs Sampling过程中，每次状态转换只沿着一个变量（或者可以说是一个坐标轴）进行，变量之间的顺序可以采用轮循。在LDA中，Gibbs Sampling采样的过程相当于对于所有的文档的所有的单词，也就是整体语料库里面的每一个单词都当作一个随机变量，每一次采样的时候都选取一个单词作为随机变量，然后根据除了这个单词的所有其他的语料决定当前单词的下一个状态（要分配的topic编号），Gibbs Sampling的过程就是不停的循环对所有的单词从新估计topic分布，直到topic编号分布实现收敛，不再变化（或者变化幅度小于某个阈值），这样就对应于最终的每篇文档的topic分布和每个 topic下 term的多项分布。

总结：有了Markov Chain之后，我们就知道了一个概率分布通过状态转移最终能够实现收敛，但是我们怎么去模拟一个概率分布的状态转移？并不是所有的概率分布都有像经济状况分布那样可以观测到的状态转移，所以我们就需要一个方法进行状态转移，这里MCMC和Gibbs Sampling就是用来通过采样进行状态转移计算的。

？目前Gibbs采样比较明白，还是不太明白MCMC采样到底是个什么过程。

## 4. Gamma函数

Gamma函数：

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

通过分部积分的方法，可以推导出这个函数有如下的递归性质：

$$\Gamma(x+1) = x\Gamma(x)$$

$\Gamma(x)$ 函数可以当成是阶乘在实数集上的延拓，如果x是整数，Gamma函数的定义则为：

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

Gamma函数将阶乘由整数域扩展到了实数域， 同时将倒数的计算由整数域扩展到了实数域。

与Gamma函数相关的一个函数：

$$B(m,n)=\frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

Gamma 分布在概率统计领域也是一个万人迷，众多统计分布和它有密切关系。指数分布和$\chi_2$分布都是特殊的Gamma 分布。另外Gamma 分布作为先验分布是很强大的，在贝叶斯统计分析中被广泛的用作其它分布的先验。如果把统计分布中的共轭关系类比为人类生活中的情侣关系的话，那指数分布、Poission分布、正态分布、对数正态分布都可以是Gamma 分布的情人。

总结：Gamma函数实现了将阶乘和倒数的计算扩展到实数域， 并且Gamma分布也是很多分配的先验分布。 之所以在这里简单的提一些Gamma分布， 是因为在Beta分布和Dirichlet分布中用到了Gamma函数。

# 5. Beta Distribution & Dirichlet Distribution

## 5.1 Beta Distribution

Beta函数的定义为：

$$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

这个就是一般意义上的 Beta 分布！可以证明，在$\alpha,\beta$取非负实数的时候，这个概率密度函数也都是良定义的。 上面公式中， x 可以理解为二项分布中的事件发生的概率p。

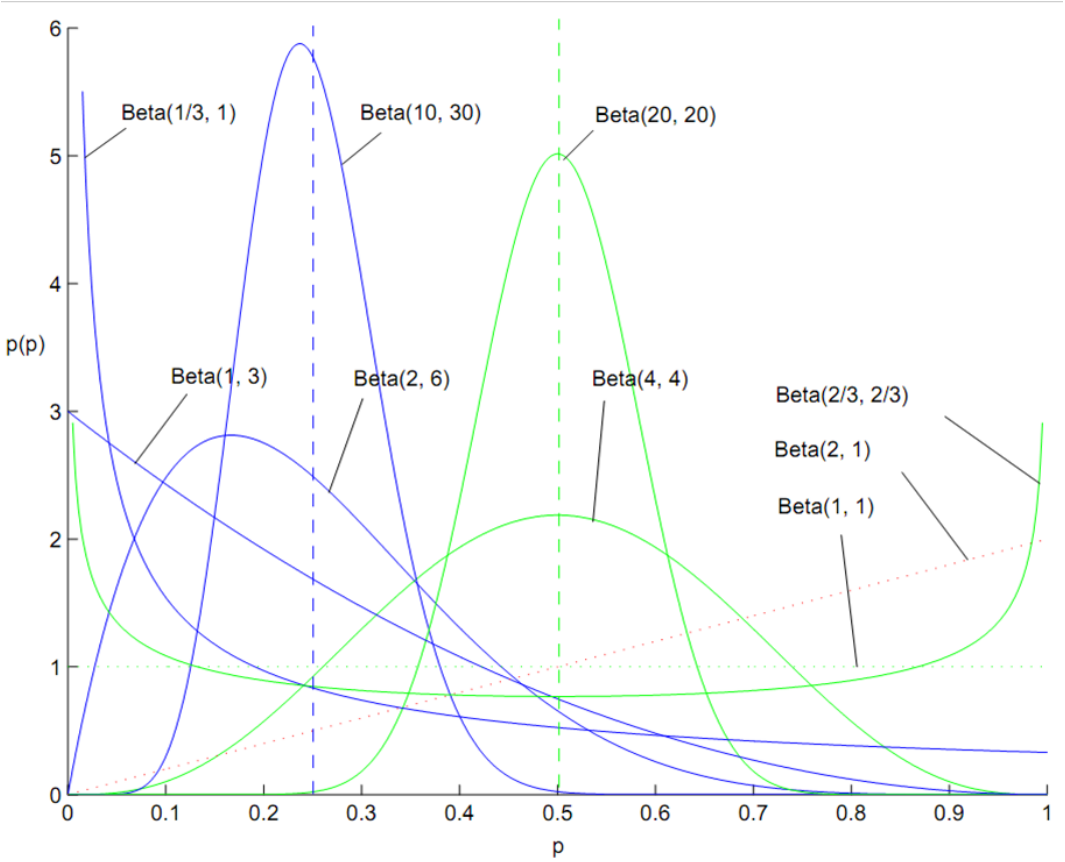
Beta分布也可以将f(x) 换成p(x)， 即表示x的概率分布函数（这里的x表示的就是二项分布的参数p）。 α, β 的物理意义是伪计数， pseudo-count, 因为beta分布是二项分布的参

数的分布（即二项分布中的参数p的分布）， 所以α, β 作为伪计数的情况下， beta分布可以理解为二项分布的先验分布。 因为二项分布存在两种情况： 发生和不发生， 所以对应的先验分布（beta分布）有两个参数α和β， α,β分别作为二项分布的先验分布的伪计数，即我们的先验知识告诉我们这两种事件发生的次数。

Beta分布称为二项分布的共厄先验分布， 因为 ：

$$Beta(p|\alpha, \beta) + Count(m_1, m_2) = Beta(p|\alpha+m_1,\beta+m_2)$$

上面公式可以看出， 二项分布的参数p的后验概率分布也服从Beta分布。 $Count(m1, m2)$ 为先验知识。




**百变星君Beta分布**

Beta分布可以理解为二项分布的参数p的先验分布， 在二项分布加上观测数据以后可以得到二项分布的参数p的后验分布也是beta分布。

Beta分布的一些性质计算：

$$E(x) = \mu = \frac{\alpha}{\alpha + \beta}$$

$$Variance(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\beta}{(\alpha+\beta)(\alpha+\beta+1)} \cdot \mu$$

$$Skewness(x) = \frac{2(\beta-\alpha)\sqrt{1+\alpha+\beta}}{\sqrt{\alpha+\beta}(2+\alpha+\beta)} \quad \text{偏度}$$

$$Kurtosis(x) = \frac{6[\alpha^3 + \alpha^2(1-2\beta) + \beta^2(1+\beta) - 2\alpha\beta(2+\beta)]}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)} \quad \text{峰度}$$

$$mode = \gamma = \frac{\alpha-1}{\alpha+\beta-2} \quad \alpha > 1 \text{and } \beta > 1 \quad \text{众数}$$

## 5.2 Dirichlet Distribution

Dirichlet分布是由Beta分布衍生来的， Dirichlet分布是多项分布的参数p的先验分布。 在一个多项分布中， 每一项的概率$p_i$如何确定？ 在Bayesian学派认为， 多项分布的参数p应该是服从一个概率分布的， 可以认为服从Dirichlet分布， 也就是说Dirichlet分布是多项分布的参数的分布， 被认为是“分布上的分布”。

Dirichlet分布的定义：

$$Dir(\vec{p}|\vec{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k-1}$$

上面公式中， $\vec{\alpha}$ 为Dirichlet分布的参数， 即多项分布的参数 $\vec{p}$ 的分布的参数。 $\vec{\alpha}$ 的物理意义和Beta分布中的伪计数α和β一样， 表示在先验知识中每种事件发生的伪计数。

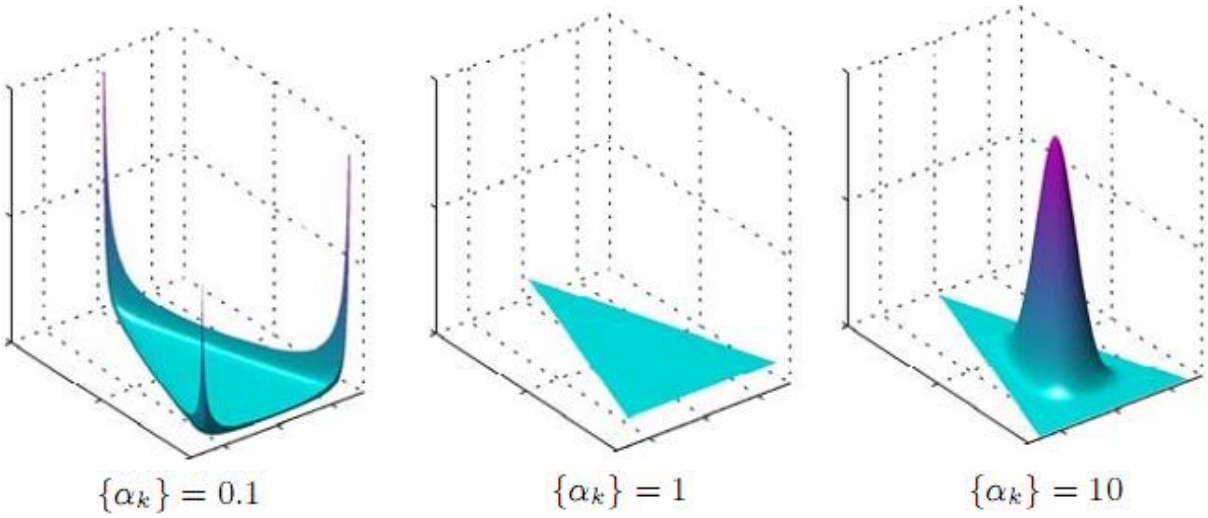
另外， Dirichlet分布还可以表示为 ：

$$Dir(\vec{p}|\vec{\alpha}) = \frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{V} p_k^{\alpha_k - 1}, \quad \vec{\alpha} = (\alpha_1, \cdots, \alpha_V)$$

$\Delta(\vec{\alpha})$ 就是归一化因子 $Dir(\vec{\alpha})$，

$$\Delta(\vec{\alpha}) = \int \prod_{k=1}^{V} p_k^{\alpha_k - 1} d\vec{p}.$$

一般在应用中采用对称的Dirichlet分布， 他和Beta 分布一样也是一个百变星君，密度函数可以展现出多种形态。




**不同 $\alpha$ 下的Dirichlet 分布**

在多项分布中， 多项分布的参数 $\vec{p}$ 的后验概率分布也服从Dirichlet分布， 并且Dirichlet分布和多项分布互为共厄分布。

$$Dir(\vec{p}|\vec{\alpha}) + MultCount(\vec{m}) = Dir(p|\vec{\alpha} + \vec{m})$$

如果 $p$~$Beta(t|\alpha, \beta)$, 则

$$\begin{aligned} E(p) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \\ &= \frac{\alpha}{\alpha+\beta} \end{aligned}$$

如果 $\vec{p} \sim Dir(\vec{t}|\vec{\alpha})$,

$$E(\vec{p}) = \left( \frac{\alpha_1}{\sum_{i=1}^{K} \alpha_i}, \frac{\alpha_2}{\sum_{i=1}^{K} \alpha_i}, \cdots, \frac{\alpha_K}{\sum_{i=1}^{K} \alpha_i} \right)$$

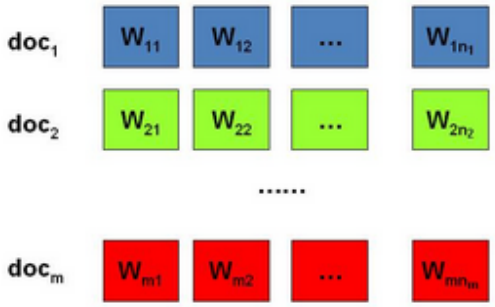
Dirichlet分布中存在两个比较重要的参数， “the scale or the concentration”

$\alpha_0 = \sum_{k=1}^{K} \alpha_k$，还有 “base measure” $\alpha_i' = \frac{\alpha_i}{\alpha 0}$ 。

在LDA模型中， 每篇文档的topic分布服从一个多项分布， 同时这个topic分布（多项分布）的参数又服从一个Dirichlet分布。 每一篇文档的都对应一个不同的topic分布， 即多项分布。 同时， LDA中的每个topic下存在一个term的单词多项分布， 即每个topic下的单词都服从多项分布， 并且每个topic对应的多项分布的参数都服从一个Dirichlet分布，这也是为什么LDA的名字由来， 因为存在两个隐含的Dirichlet分布。

## 6. 文本建模 Unigram Model

在文本中，每一篇文档都可以看作若干词的组合，$d=(w_1,w_2,\cdots,w_n)$




**包含$M$ 篇文档的语料库**

LDA为生成模型，同时Unigram Model也为生成模型。Unigram Model是把所有的文档的词都作为一个语料库，并且所有的词的分布都服从同一个多项分布，即每个词出现的概率。也就是所有的这些词组成的字典中，每个词的频率。假设词典总共有V个词，那么多项分布的情况种类就有V个，多项分布的参数 $\vec{p}$ 的维度为V。

**Game 1** Unigram Model

1: 上帝只有一个骰子，这个骰子有$V$ 个面，每个面对应一个词，各个面的概率不一；

2: 每抛一次骰子，抛出的面就对应的产生一个词；如果一篇文档中有$n$ 个词，上帝就是独立的抛$n$次骰子产生这$n$ 个词；

有每个词的出现的概率，就可以计算语料库中每个文档的生成概率：

$d=\vec{w}=(w_1,w_2,\cdots,w_n)$ 的生成概率为：

$$p(\vec{w})=p(w_1,w_2,\cdots,w_n)=p(w_1)p(w_2)\cdots p(w_n)$$

语料库中的文档可以认为是相对独立的， 所以每个文档之间是没有相关性的， 所以可以获得整个语料库的生成概率， $\mathcal{W}=(\overrightarrow{w_1},\overrightarrow{w_2},\ldots,\overrightarrow{w_m})$， 语料库的概率为：

$$p(\mathcal{W})=p(\overrightarrow{w_1})p(\overrightarrow{w_2})\cdots p(\overrightarrow{w_m})$$

假设语料中总的词频是$N$, 在所有的 $N$ 个词中,如果我们关注每个词 $v_i$ 的发生次数 $n_i$， 那么 $\overrightarrow{n}=(n_1,n_2,\cdots,n_V)$ 正好是一个多项分布， 该多项分布的概率为：

$$p(\overrightarrow{n})=Mult(\overrightarrow{n}|\overrightarrow{p},N)=\binom{N}{\overrightarrow{n}}\prod_{k=1}^{V}p_k^{n_k}$$

此时， 语料的概率是：

$$p(\mathcal{W})=p(\overrightarrow{w_1})p(\overrightarrow{w_2})\cdots p(\overrightarrow{w_m})=\prod_{k=1}^{V}p_k^{n_k}$$

注意： 要特别区分多项分布的概率计算公式和语料的生成概率计算公式， 语料的生成概率计算公式中没有全排列计算， 因为在语料中不用考虑文档以及单词的出现顺序。

我们很重要的一个任务就是估计模型中的参数 $\overrightarrow{p}$， 也就是问上帝拥有的这个骰子的各个面的概率是多大，按照统计学家中频率派的观点，使用最大似然估计最大化$P(\mathrm{W})$，于是参数$p_i$的估计值就是：

$$\hat{p}_i=\frac{n_i}{N}$$

然而， 在上面的基础上， 贝叶斯学派认为上帝只有一个骰子是不对的， 应该上帝有无数多的骰子， 只是在生成语料库的时候上帝随机的选了一个。 这就是说， 生成语料库的模型参数 $\overrightarrow{p}$ （也就是多项分布的参数）应该是服从一个概率分布的， 上帝选择一个骰子的

意思就是在 $\vec{p}$ 的分布上进行采样获得一个具体的 $\vec{p}$ 的取值(即在 $\vec{p}$ 分布上进行采样)，然后根据将这个取值作为多项分布的参数。 这个时候， 一般认为 $\vec{p}$ 服从Dirichlet分布。

**Game 2** 贝叶斯Unigram Model假设

1: 上帝有一个装有无穷多个骰子的坛子，里面有各式各样的骰子，每个骰子有$V$ 个面；

2: 上帝从坛子里面抽了一个骰子出来，然后用这个骰子不断的抛，然后产生了语料中的所有的词；

Dirichlet分布则可以认为是多项分布的参数 $\vec{p}$ 的先验分布。

在贝叶斯学派的游戏规则的假设之下，语料W产生的概率如何计算呢？ 因为W中每个单词出现的概率是受 $\vec{p}$ 影响的， 而 $\vec{p}$ 本分服从一个Dirichlet分布，所以要对所有的可能存在的 $\vec{p}$ 的情况进行全概率计算。 连续变量则进行积分计算：

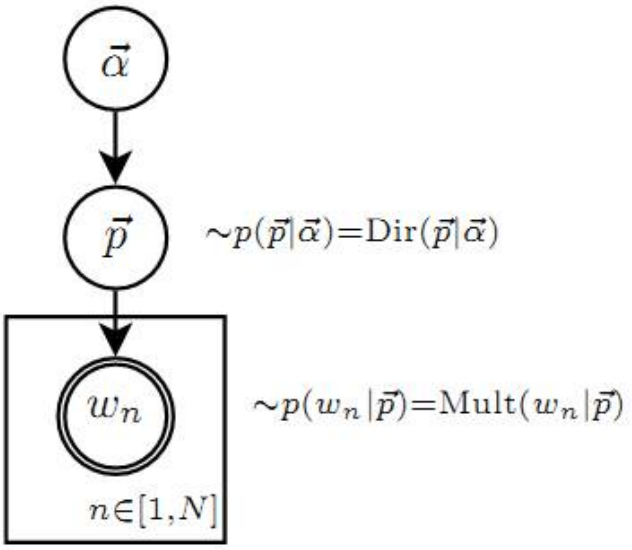
$$p(\mathcal{W}) = \int p(\mathcal{W}|\vec{p})p(\vec{p})d\vec{p}$$

在上面的公式中， $p(\vec{p})$ 应该如何计算？ 对先验分布的一个比较好的选择就是多项分布对应的共轭分布，即 Dirichlet 分布：

$$Dir(\vec{p}|\vec{\alpha}) = \frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{V} p_k^{\alpha_k - 1}, \quad \vec{\alpha} = (\alpha_1, \cdots, \alpha_V)$$

$\Delta(\vec{\alpha})$ 就是归一化因子 $Dir(\vec{\alpha})$，

$$\Delta(\vec{\alpha}) = \int \prod_{k=1}^{V} p_k^{\alpha_k - 1} d\vec{p}.$$




**Unigram Model的概率图模型**

之前提到了以下的概念：

**Dirichlet 先验 + 多项分布的数据 → 后验分布为 Dirichlet 分布**

$$Dir(\overrightarrow{p}|\overrightarrow{\alpha}) + MultCount(\overrightarrow{n}) = Dir(\overrightarrow{p}|\overrightarrow{\alpha} + \overrightarrow{n})$$

所以， 在给定语料库的基础上， 可以得知多项分布的参数 $\overrightarrow{p}$ 的后验概率分布也服从Dirichlet分布， 我们可以计算 $\overrightarrow{p}$ 的后验概率分布：

$$p(\overrightarrow{p}|\mathcal{W}, \overrightarrow{\alpha}) = Dir(\overrightarrow{p}|\overrightarrow{n} + \overrightarrow{\alpha}) = \frac{1}{\Delta(\overrightarrow{n} + \overrightarrow{\alpha})} \prod_{k=1}^{V} p_k^{n_k+\alpha_k-1} d\overrightarrow{p}$$

计算得到参数 $\vec{p}$ 的后验概率分布以后，就可以根据后验概率分布计算参数 $\vec{p}$ 的估计了，可以根据后验概率分布的期望计算参数 $\vec{p}$ 的平均取值，作为 $\vec{p}$ 的估计值。$\vec{p}$ 的后验概率分布为 $Dir(\vec{p}|\vec{n}+\vec{\alpha})$，于是：

$$E(\vec{p}) = \left(\frac{n_1+\alpha_1}{\sum_{i=1}^{V}(n_i+\alpha_i)}, \frac{n_2+\alpha_2}{\sum_{i=1}^{V}(n_i+\alpha_i)}, \cdots, \frac{n_V+\alpha_V}{\sum_{i=1}^{V}(n_i+\alpha_i)}\right)$$

$$\hat{p}_i = \frac{n_i+\alpha_i}{\sum_{i=1}^{V}(n_i+\alpha_i)}$$

$\alpha_i$ 在 Dirichlet 分布中的物理意义是事件的先验的伪计数，这个估计式子的含义是很直观的：每个参数的估计值是其对应事件的先验的伪计数和数据中的计数的和在整体计数中的比例。

进一步，我们可以计算出文本语料的产生概率为：

$$\begin{aligned}
p(\mathcal{W}|\vec{\alpha}) &= \int p(\mathcal{W}|\vec{p})p(\vec{p}|\vec{\alpha})d\vec{p} \\
&= \int \prod_{k=1}^{V} p_k^{n_k} Dir(\vec{p}|\vec{\alpha})d\vec{p} \\
&= \int \prod_{k=1}^{V} p_k^{n_k} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{V} p_k^{\alpha_k-1} d\vec{p} \\
&= \frac{1}{\Delta(\vec{\alpha})} \int \prod_{k=1}^{V} p_k^{n_k+\alpha_k-1} d\vec{p} \\
&= \frac{\Delta(\vec{n}+\vec{\alpha})}{\Delta(\vec{\alpha})}
\end{aligned}$$

总结：在Unigram Model中，所有的文档中的所有的单词都被看作属于统一个语料库，并且都服从同样的分布（这一点和LDA是不同的，在LDA中每一个文档都有自己的topic分布，并且每一个topic都有自己对应的term分布）。贝叶斯学派认为每个单词的出现概率服从多项分布，即要生成一个新的单词有V中选择，针对V中选择有V个不同的概率，即构成 $\vec{p}$ 。多项分布的参数 $\vec{p}$ 又是一个服从Dirichlet分布的随机变量。

## 7. PLSA

Unigram Model过于简单，但是很好的解释了多项分布参数 $\vec{p}$ 的计算过程，同时理解生成模型的过程。

在生活中，写一篇文档总是要先决定这个文档包含哪些主题，以及每个主题的权重（这可能就对应LDA中的每篇文档的topic分布），然后确定了该文档包含哪些主题以后，就需要确定在这个主题下选择该主题下的哪些能够表达该主题的单词（对应于每个主题的term分布），然后重复选择主题和根据主题选择单词的这个过程就可以最终获取一篇指定单词数量的文章。

生成一个文本的生成模型描述：

**Game 3** PLSA Topic Model 假设

1: 上帝有两种类型的骰子，一类是doc-topic 骰子,每个doc-topic 骰子有$K$ 个面，每个面是一个topic 的编号；一类是topic-word 骰子，每个topic-word 骰子有$V$ 个面，每个面对应一个词；

doc-topic

topic-word

2: 上帝一共有$K$ 个topic-word 骰子, 每个骰子有一个编号，编号从1 到$K$；

3: 生成每篇文档之前，上帝都先为这篇文章制造一个特定的doc-topic 骰子，然后重复如下过程生成文档中的词

- 投掷这个doc-topic 骰子,得到一个topic 编号$z$
- 选择$K$ 个topic-word 骰子中编号为$z$的那个，投掷这个骰子，于是得到一个词

上面描述中， 上帝先对一篇文档选择了一个骰子， 即确定了这篇文档的主题分布， 因为是从无数的骰子中选择的， 所以也对应了每篇文档的主题分布的参数也服从一个分布（Dirichlet分布）， 然后根据这个doc-topic骰子进行topic选择， 得到topic以后再选择对应的topic-word分布， 从中选择单词。 在PLSA中每一个topic都对应一个 term 的多项分布， 这些多项分布可以表示为： $\vec{\varphi}_1, \cdots, \vec{\varphi}_K$， 每一个文档都对应于一个topic分布， 可以表示为： $\vec{\theta}_1, \cdots, \vec{\theta}_M$， 如果每个单词有一个编号， 那么一篇文档中每个词的生成概率为： （全概率公式）

$$p(w|d_m) = \sum_{z=1}^{K} p(w|z)p(z|d_m) = \sum_{z=1}^{K} \varphi_{zw}\theta_{mz}$$

然后可以得到一篇文档的生成概率：

$$p(\vec{w}|d_m) = \prod_{i=1}^{n}\sum_{z=1}^{K} p(w_i|z)p(z|d_m) = \prod_{i=1}^{n}\sum_{z=1}^{K} \varphi_{zw_i}\theta_{dz}$$

然后就可以根据EM算法求解模型的参数。

总结：貌似PLSA并没有利用Dirichlet分布，每篇文档对应的topic分布以及每个topic对应的term分布都没有采用Dirichlet分布作为多项分布参数的先验分布，没有充分利用贝叶斯理论，所以在后来提出了基于贝叶斯理论并且使用Dirichlet分布作为先验分布的LDA模型。

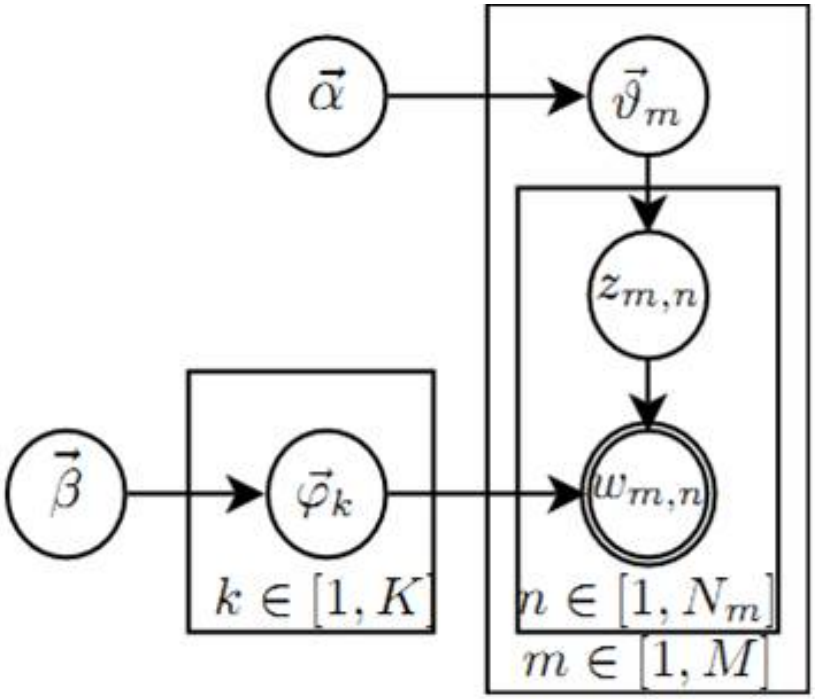
## 8. LDA详解

在了解了上面那么多基本知识以后，终于要聊聊LDA了，我在最早接触LDA的时候发现没有上面的基础知识看LDA真的很费劲。

首先要讲Gibbs Sampling和模型的参数的inference，最早我理解LDA的最大的困惑在于为神马要采用Gibbs Sampling，事实上采用Gibbs Sampling是为了降低参数估计的复杂度，能够实现参数估计。这个接下来会详细说明。

### 8.1 混合模型 & 生成模型

LDA是一个生成模型，生成模型的示例图如下图所示：




对于LDA生成模型的理解：假定我们需要写一篇文章，那么这个文章可能包含多个主题，也就是说我们先要决定我们的文章由哪些主题构成，这些主题所占的比例如何。有了主题分布，也就对应于 $\vec{\theta}_m$（每篇文章的主题分布），有了主题分布，我们就开始写文章，在决定写一个单词的时候，根据上图的模型，先要由主题分布决定这个单词所属的topic，然后有了这个单词的topic，在根据每个topic下面的term的分布选择应该采用哪个单词。这样重复这个过程就可以获得一篇有n个单词的文章。

**LDA 生成模型中，$M$ 篇文档会对应于 $M$ 个独立的 Dirichlet-Multinomial 共轭结构；$K$ 个 topic 会对应于 $K$ 个独立的 Dirichlet-Multinomial 共轭结构。**

LDA中，每个文章都有自己的topic分布，每个topic都对应于自己的term分布，所有的doc-topic分布的参数都服从Dirichlet分布，所有的topic-term分布的参数都服从Dirichlet分布。

LDA生成模型的算法过程描述如下：

```
// topic plate
for all topics k ∈ [1, K] do
    sample mixture components φ⃗_k ~ Dir(β⃗)
// document plate:
for all documents m ∈ [1, M] do
    sample mixture proportion ϑ⃗_m ~ Dir(α⃗)
    sample document length N_m ~ Poiss(ξ)
    // word plate:
    for all words n ∈ [1, N_m] in document m do
        sample topic index z_{m,n} ~ Mult(ϑ⃗_m)
        sample term for word w_{m,n} ~ Mult(φ⃗_{z_{m,n}})
```

因为LDA中包含了多个多项分布， 所以LDA称为混合模型， 在LDA中， 一个单词t出现的概率应该符合全概率公式， 即在所有的topic中这个单词出现的概率和：

$$p(w=t) = \sum_k p(w=t|z=k)p(z=k), \quad \sum_k p(z=k) = 1$$

LDA需要计算的不是全局的topic分布(global topic proportion)， 而是要以文档为单位的基础上一个单词的topic分布。 这样， 可以得知LDA inference的目的为： （1） $p(t|z=k) = \vec{\varphi}_k$， 即为每个topic下的term分布， （2） $p(z|d=m) = \vec{\vartheta}_m$， 每个文档对应的topic分布。 最终要估计的是两个参数：

$\underline{\Phi} = \{\vec{\varphi}_k\}_{k=1}^K$ and $\underline{\Theta} = \{\vec{\vartheta}_m\}_{m=1}^M$ 。

## 8.2 Likelihoods （似然估计）

根据LDA模型， 对于一个文档m 的完全数据的似然估计（关于所有的已知和未知参数的联合分布）为：

$$p(\vec{w}_m, \vec{z}_m, \vec{\vartheta}_m, \underline{\Phi}|\vec{\alpha}, \vec{\beta}) = \overbrace{\underbrace{\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\varphi}_{z_{m,n}}) p(z_{m,n}|\vec{\vartheta}_m)}_{\text{word plate}} \cdot p(\vec{\vartheta}_m|\vec{\alpha}) \cdot \underbrace{p(\underline{\Phi}|\vec{\beta})}_{\text{topic plate}}}^{\text{document plate (1 document)}} .$$

~~注释：这里我一直觉得“word plate”的范围错了，我觉得不应该包括连乘符号。而且，连乘后边继续放一个圆点后放一个条件概率，容易让人误会圆点以后的概率也是对于所有的word都进行乘法操作。~~

是我的以前的理解错了，圆点以后的乘法操作，确实是对所有的word进行的。因为，对已一篇文档，我们可以先根据两个超参数alpha, beta获得topic-word的概率分布以及doc-topic的概率分布。根据beta超参数，可以获得每个topic对应的word的多项分布，根据alpha参数，可以获得当前第m个文档的topic的多项分布，这里为 $\vec{\theta}_m$，因为每个文档对应的topic分布符合多项分布，所有多项分布的参数 $\vec{\theta}_m$ 符合Dirichlet分布，alpha即为多项分布的参数 $\vec{\theta}_m$ 的分布的参数。因为多项分布的参数服从Dirichlet分布，所以可以根据当前的参数计算 $\vec{\theta}_m$ 在Dirichlet分布中的概率。

在文档m中，一个单词t出现的概率为 （通过计算$Z_{m,n}$的边缘分布，并且忽略多项分布的参数分布）

$$p(w_{m,n}=t|\vec{\vartheta}_m, \underline{\Phi}) = \sum_{k=1}^{K} p(w_{m,n}=t|\vec{\varphi}_k) p(z_{m,n}=k|\vec{\vartheta}_m) ,$$

由每个单词出现的概率，我们可以计算整个语料库的生成概率：

$$p(\mathcal{W}|\underline{\Theta},\underline{\Phi}) = \prod_{m=1}^{M} p(\vec{w}_m|\vec{\vartheta}_m,\underline{\Phi}) = \prod_{m=1}^{M}\prod_{n=1}^{N_m} p(w_{m,n}|\vec{\vartheta}_m,\underline{\Phi}) .$$

上面公式给出了两个参数的似然估计， 虽然LDA虽然是比较简单的模型， 但是直接推断这个模型的参数是非常困难的， 可以采用近似推断算法。 所以需要采用Gibbs Sampling进行采样估计。

### 8.3 Inference via Gibbs Sampling

在LDA模型中， 隐藏的模型变量为 $Z_{m,n}$， （$Z_{m,n}$ 表示文档m中第n个单词的 topic值），之前一直在说要用Gibbs Sampling， 这里$Z_{m,n}$ 就相当于Markov Chain中的状态变量， 当Z 收敛以后， 整个语料库的每个文档的topic分布以及每个topic对应的term分布也就收敛了， 也就是 $\vec{\theta}_m$ 和 $\vec{\varphi}_k$ 都实现收敛， 最终可以计算LDA的两个多项分布的参数。

Inference的目的是计算如下公式的概率（其中$Z_i = Z_{m,n}$， 这里是为了方便表示）： （因为在给定观测样本 $\vec{w}$ 的基础上， 计算得到Z的概率分布即可以获得整个LDA模型的所有参数， 即 $\vec{\theta}_m$ 和 $\vec{\varphi}_k$ ， 计算 $p(\vec{z}|\vec{w})$ 概率分布的目的就是看所有的语料库的单词对应的topic分布是否已经收敛了， 如果已经收敛则可以利用Z 计算得到模型参数。 因为上面提到$Z_{m,n}$ 对应于Markov Chain的状态变量， 所以对于不同的状态的Z会有不同的 $p(\vec{z}|\vec{w})$ 概率分布， 但是当Z收敛以后， $p(\vec{z}|\vec{w})$ 应该不会再变化。 这里说了好多为什么要推断 $p(\vec{z}|\vec{w})$ ， 应该很清楚了。）

或者说，当 $p(\vec{z}|\vec{w})$ 比较难计算的时候，我们就可以用Gibbs采样的算法，通过使得Markov Chain达到收敛使得在某个状态以后的所有采样Z 都服从同一个概率分布。

$$p(\vec{z}|\vec{w}) = \frac{p(\vec{z},\vec{w})}{p(\vec{w})} = \frac{\prod_{i=1}^{W} p(z_i, w_i)}{\prod_{i=1}^{W} \sum_{k=1}^{K} p(z_i{=}k, w_i)}$$

但是上面公式中的分母部分很难计算求得，这就需要采用Gibbs Sampling来发挥作用了（发挥作用的意思不是可以用Gibbs Sampling计算分母，而是采用Gibbs Sampling获得收敛的Z）。需要利用Gibbs Sampling 的全条件概率分布 $p(z_i|\vec{z}_{\neg i}, \vec{w})$ 去模拟 $p(\vec{z}|\vec{w})$。关于Gibbs Sampling的公式，存在以下的分析。

Gibbs Sampling是一个特殊的MCMC，并且在采样过程中轮流对不同的维度$x_i$ 进行采样，在对维度$x_i$ 采样的时候，利用所有的其他的不包含$x_i$ 的维度的概率分布计算$x_i$ 的新的采样值，不包含$x_i$ 的维度可以表示为：$\vec{x}_{\neg i}$。

Gibbs Sampling的过程描述如下：

1. choose dimension $i$ (random or by permutation[17])
2. sample $x_i$ from $p(x_i|\vec{x}_{\neg i})$.

为了获取一个Gibbs Sampling采样器，必须得到一个univariate conditionals (or full conditionals) $p(x_i|\vec{x}_{\neg i})$ ：

$$p(x_i|\vec{x}_{\neg i}) = \frac{p(\vec{x})}{p(\vec{x}_{\neg i})} = \frac{p(\vec{x})}{\int p(\vec{x})\,\mathrm{d}x_i} \text{ with } \vec{x} = \{x_i, \vec{x}_{\neg i}\}$$

对于包含隐含变量Z，并且已知后验概率分布$p(\vec{z}|\vec{x})$ 的模型，Gibbs 采样器可以表示为：

$$p(z_i|\vec{z}_{\neg i}, \vec{x}) = \frac{p(\vec{z}, \vec{x})}{p(\vec{z}_{\neg i}, \vec{x})} = \frac{p(\vec{z}, \vec{x})}{\int_Z p(\vec{z}, \vec{x})\, \mathrm{d}z_i},$$

在这个Gibbs采样器中有一个联合概率分布 $p(\vec{z}, \vec{x})$， 我们需要计算这个联合概率分布才能获得具体的Gibbs采样器。

### 8.4 联合概率分布

联合概率分布可以表示为：

$$p(\vec{w}, \vec{z}|\vec{\alpha}, \vec{\beta}) = p(\vec{w}|\vec{z}, \vec{\beta})p(\vec{z}|\vec{\alpha}),$$

我们把这个公式分成两个部分分别计算， 首先是 $p(\vec{w}|\vec{z})$（给定语料库中每个单词的topic值， 根据topic值获得语料库的生成概率）， 得到：

$$p(\vec{w}|\vec{z}, \underline{\Phi}) = \prod_{i=1}^{W} p(w_i|z_i) = \prod_{i=1}^{W} \varphi_{z_i, w_i}.$$

如果表示为另外一种形式， 则得到：

$$p(\vec{w}|\vec{z}, \underline{\Phi}) = \prod_{k=1}^{K} \prod_{\{i:z_i=k\}} p(w_i=t|z_i=k) = \prod_{k=1}^{K} \prod_{t=1}^{V} \varphi_{k,t}^{n_k^{(t)}},$$

$\underline{\Phi}$ 是关于 $\vec{\beta}$ 的分布， 所以 $p(\vec{w}|\vec{z}, \vec{\beta})$ 的计算可以通过对 $\underline{\Phi}$ 进行积分得到：

$$
\begin{aligned}
p(\vec{w}|\vec{z},\vec{\beta}) &= \int p(\vec{w}|\vec{z},\underline{\Phi})\ p(\underline{\Phi}|\vec{\beta})\ \mathrm{d}\underline{\Phi} \\
&= \int \prod_{z=1}^{K} \frac{1}{\Delta(\vec{\beta})} \prod_{t=1}^{V} \varphi_{z,t}^{n_z^{(t)}+\beta_t-1}\ \mathrm{d}\vec{\varphi}_z \\
&= \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})}, \quad \vec{n}_z = \{n_z^{(t)}\}_{t=1}^{V}.
\end{aligned}
$$

在上面的公式中，$p(\underline{\Phi}|\vec{\beta})$ 服从Dirichlet分布，所以可以获得第二步的公式推导。

同样的，$p(\vec{z}|\vec{\alpha})$ 也可以通过以下公式得到：

$$
p(\vec{z}|\underline{\Theta}) = \prod_{i=1}^{W} p(z_i|d_i) = \prod_{m=1}^{M}\prod_{k=1}^{K} p(z_i{=}k|d_i{=}m) = \prod_{m=1}^{M}\prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}},
$$

然后可以通过对 $\underline{\Theta}$ 进行积分计算 $p(\vec{z}|\vec{\alpha})$ ：

$$
\begin{aligned}
p(\vec{z}|\vec{\alpha}) &= \int p(\vec{z}|\underline{\Theta})\ p(\underline{\Theta}|\vec{\alpha})\ \mathrm{d}\underline{\Theta} \\
&= \int \prod_{m=1}^{M} \frac{1}{\Delta(\vec{\alpha})} \prod_{k=1}^{K} \vartheta_{m,k}^{n_m^{(k)}+\alpha_k-1}\ \mathrm{d}\vec{\vartheta}_m \\
&= \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}, \quad \vec{n}_m = \{n_m^{(k)}\}_{k=1}^{K}.
\end{aligned}
$$

最后得到联合概率分布的结果：

$$
p(\vec{z},\vec{w}|\vec{\alpha},\vec{\beta}) = \prod_{z=1}^{K} \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})}.
$$

结合上面提到的Gibbs采样器公式和联合概率公式， 可以最终得到Gibbs Sampling的采样公式为 ：

$$
\begin{aligned}
p(z_i{=}k|\vec{z}_{\neg i},\vec{w}) &= \frac{p(\vec{w},\vec{z})}{p(\vec{w},\vec{z}_{\neg i})} = \frac{p(\vec{w}|\vec{z})}{p(\vec{w}_{\neg i}|\vec{z}_{\neg i})p(w_i)} \cdot \frac{p(\vec{z})}{p(\vec{z}_{\neg i})} \\
&\propto \frac{\Delta(\vec{n}_z+\vec{\beta})}{\Delta(\vec{n}_{z,\neg i}+\vec{\beta})} \cdot \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{n}_{m,\neg i}+\vec{\alpha})} \\
&= \frac{\Gamma(n_k^{(t)}+\beta_t)\,\Gamma(\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t)}{\Gamma(n_{k,\neg i}^{(t)}+\beta_t)\,\Gamma(\sum_{t=1}^{V} n_k^{(t)}+\beta_t)} \cdot \frac{\Gamma(n_m^{(k)}+\alpha_k)\,\Gamma(\sum_{k=1}^{K} n_{m,\neg i}^{(k)}+\alpha_k)}{\Gamma(n_{m,\neg i}^{(k)}+\alpha_k)\,\Gamma(\sum_{k=1}^{K} n_m^{(k)}+\alpha_k)} \\
&= \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t} \cdot \frac{n_{m,\neg i}^{(k)}+\alpha_k}{[\sum_{k=1}^{K} n_m^{(k)}+\alpha_k]-1} \\
&\propto \frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)}+\beta_t}(n_{m,\neg i}^{(k)}+\alpha_k)
\end{aligned}
$$

注释： 当看到得到这个公式以后， 我最初很迷茫有了这个公式到底怎么采样？ 得到的就是一个概率， 采毛样。。。。， 后来看了资料以后， 是需要计算当前的term （单词）所有对应的 k 的概率取值， 这样就获得的所有的 K 个 新的这个单词将要划分的topic的值， 然后生成随机数在这个K 个topic中选择 这个单词 需要重新划分的topic 。 对于所有的单词都进行这个过程操作， 直到 Z 收敛以后。

Gibbs采样就是勇条件分布的采样拉哎替代全概率分布的采样。

这个Gibbs采样公式要利用联合概率分布 $p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta})$ ， 或者更加详细的联合概率分布公式 $p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta}) = p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha})$, ， 所以在计算Gibbs采样中非常重要的步骤是计算这个联合概率分布，所以会有上面的联合概率分布计算推导。

### 8.5 参数估计

根据以前提到的Dirichlet分布的特征可以得知， LDA模型中对应的每个文档的topic分布 $\vec{\theta}_m$ 和每个 topic下term的分布 $\vec{\varphi}_k$ 都服从Dirichlet分布(即多项分布的参数服从Dirichlet分布)， 并且 $\vec{\theta}_m$ ， $\vec{\varphi}_k$ 的后验概率分布也服从Dirichlet分布：

$$p(\vec{\vartheta}_m|\vec{z}_m,\vec{\alpha}) = \frac{1}{Z_{\vartheta_m}} \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\vartheta}_m) \cdot p(\vec{\vartheta}_m|\vec{\alpha}) = \mathrm{Dir}(\vec{\vartheta}_m|\vec{n}_m + \vec{\alpha}),$$

$$p(\vec{\varphi}_k|\vec{z},\vec{w},\vec{\beta}) = \frac{1}{Z_{\varphi_k}} \prod_{\{i:z_i=k\}} p(w_i|\vec{\varphi}_k) \cdot p(\vec{\varphi}_k|\vec{\beta}) = \mathrm{Dir}(\vec{\varphi}_k|\vec{n}_k + \vec{\beta})$$

其中， $\vec{n}_m$ 表示第 m 篇文档中每个topic的单词数量向量（维度为 K）， $\vec{n}_k$ 表示第 k 个topic下 词典 V 中 每个单词在整个语料库中出现的次数 （维度为 V）， 其中， $\vec{n}_m$ ， $\vec{n}_k$ 都可以根据Gibbs Sampling采样达到收敛状态后的 Z 统计得到（计算很简单）。 然后可以得到模型的参数计算：

$$\varphi_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V} n_k^{(t)} + \beta_t},$$

$$\vartheta_{m,k} = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_m^{(k)} + \alpha_k}.$$

这里是根据Dirichlet分布的期望计算公式计算 $\vec{\theta}_m$ 和 $\vec{\varphi}_k$ 的。

终于通过各种分析， Gibbs Sampling计算得到了LDA的模型参数， 并且在计算LDA的Gibbs Sampling 过程中计算了联合概率分布 $p(\vec{w},\vec{z}|\vec{\alpha},\vec{\beta}) = p(\vec{w}|\vec{z},\vec{\beta})p(\vec{z}|\vec{\alpha})$, 其实在最初看论文的过程中最大的问题就是思维比较混乱， 不知到哪里的公式为什么使用

，为什么要计算联合概率分布，为什么要计算 $p(\vec{z}|\vec{w})$ ，等到多看几遍论文以后，便发现所有的论证都是围绕着Gibbs Sampling进行的，都是在为引出或者计算Gibbs Sampling公式服务的。

Gibbs Sampling的过程描述如下：

**Algorithm** LdaGibbs($\{\vec{w}\}, \alpha, \beta, K$)
**Input**: word vectors $\{\vec{w}\}$, hyperparameters $\alpha, \beta$, topic number $K$
**Global data**: count statistics $\{n_m^{(k)}\}$, $\{n_k^{(t)}\}$ and their sums $\{n_m\}$, $\{n_k\}$, memory for full conditional array $p(z_i|\cdot)$
**Output**: topic associations $\{\vec{z}\}$, multinomial parameters $\underline{\Phi}$ and $\underline{\Theta}$, hyperparameter estimates $\alpha, \beta$
// initialisation
zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$
**for** all documents $m \in [1, M]$ **do**
    **for** all words $n \in [1, N_m]$ in document $m$ **do**
        sample topic index $z_{m,n}=k \sim \text{Mult}(1/K)$
        increment document–topic count: $n_m^{(k)}$ += 1
        increment document–topic sum: $n_m$ += 1
        increment topic–term count: $n_k^{(t)}$ += 1
        increment topic–term sum: $n_k$ += 1
// Gibbs sampling over burn-in period and sampling period
**while** not finished **do**
    **for** all documents $m \in [1, M]$ **do**
        **for** all words $n \in [1, N_m]$ in document $m$ **do**
            // for the current assignment of $k$ to a term $t$ for word $w_{m,n}$:
            decrement counts and sums: $n_m^{(k)}$ −= 1; $n_m$ −= 1; $n_k^{(t)}$ −= 1; $n_k$ −= 1
            // multinomial sampling acc. to Eq. 78 (decrements from previous step):
            sample topic index $\tilde{k} \sim p(z_i|\vec{z}_{\neg i}, \vec{w})$
            // for the new assignment of $z_{m,n}$ to the term $t$ for word $w_{m,n}$:
            increment counts and sums: $n_m^{(\tilde{k})}$ += 1; $n_m$ += 1; $n_{\tilde{k}}^{(t)}$ += 1; $n_{\tilde{k}}$ += 1
    // check convergence and read out parameters
    **if** converged and $L$ sampling iterations since last read out **then**
        // the different parameters read outs are averaged.
        read out parameter set $\underline{\Phi}$ according to Eq. 81
        read out parameter set $\underline{\Theta}$ according to Eq. 82

### 8.6 new come document

对于未知的新来的文档， 如何根据已经训练好的模型将文档表示为 topic 的向量或者获取文档的topic特征呢？

对于一个 query document， 可以看作 term 向量 $\vec{\tilde{w}}$， 我们可以在已有的LDA模型参数的基础上计算 query document的 topic分布 $\vec{\tilde{z}}$ 的后验概率， 也可以计算文档的topic分布 $\vec{\tilde{\vartheta}}_m$ 。 计算得到文档的 topic则可以根据这个topic分布的向量计算文档之间的相似度。

计算query document 的topic分布 $\vec{\tilde{z}}$ 的过程和LDA模型的参数估计过程有一些区别， （1） 在query document的Gibbs采样过程中使用LDA模型中已经得到的 $\underline{\Phi}$ 和训练模型的时候使用的 $\alpha$ ， 在query document的 topic分布计算过程中需要使用 已经得到参数 $\underline{\Phi}$， 即我们假设所有的topic对应的term分布都符合从语料库中得到的参数。 （2） 参数 $\underline{\tilde{\Theta}}$ 只对测试文档有效。（不知到理解的对不对。。。。）

和Gibbs Sampling一样， 首先给query document 的每一个word随机分配一个 topic， 然后进行Gibbs Sampling， 直到query document的 $\vec{\tilde{z}}$ 实现收敛， 采样公式如下所示：

$$p(\tilde{z}_i=k|\tilde{w}_i=t, \vec{\tilde{z}}_{\neg i}, \vec{\tilde{w}}_{\neg i}; \mathcal{M}) \propto \varphi_{k,t} \ (n^{(k)}_{\tilde{m},\neg i} + \alpha_k) \ .$$

有了上面公式的Gibbs Sampling采样过程以后， 到达 $\vec{\tilde{z}}$ 收敛之后， 就可以计算query document 的topic 分布 $\vec{\tilde{\vartheta}}_m$ 了：

$$\vartheta_{\tilde{m},k} = \frac{n_{\tilde{m}}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{\tilde{m}}^{(k)} + \alpha_k}.$$

这里还有个大疑问：在进行Gibbs Sampling的过程中，怎么确定（计算）Z已经收敛了呢？貌似论文中没有提到。。。（难道根据Z的变化？。。。）

应该就是根据Z每次迭代的变化幅度，例如变化不超过0.001则认为收敛。