Categorical Distributions in Natural Language Processing
Version 0.1

MURAWAKI Yugo

12 May 2016

Categorical distribution

Suppose random variable x takes one of K values. x is generated according to categorical distribution $\mathrm{Cat}(\boldsymbol{\theta})$, where

$$\boldsymbol{\theta} = \quad (0.1, 0.6, 0.3).$$
$$\text{R} \quad \text{Y} \quad \text{G}$$

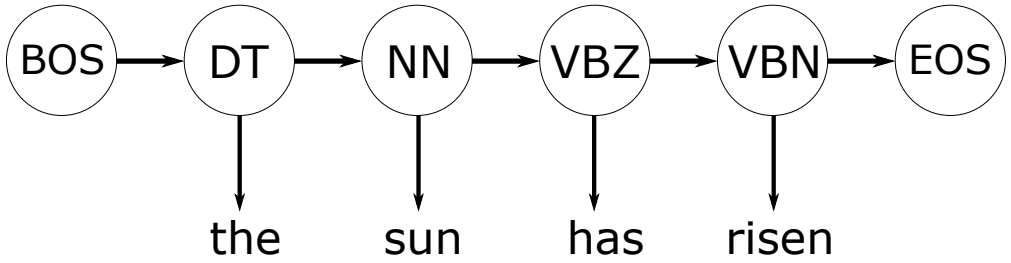
In many task settings, we do not know the true $\boldsymbol{\theta}$ and need to infer it from observed data $\boldsymbol{x} = (x_1, \cdots, x_N)$. Once we infer $\boldsymbol{\theta}$, we often want to predict new variable x'.

NOTE: In Bayesian settings, $\boldsymbol{\theta}$ is usually integrated out and x' is predicted directly from $\boldsymbol{x}$.

Categorical distributions are a building block of natural language models

- N-gram language model (predicting the next word)
- POS tagging based on a Hidden Markov Model (HMM)
- Probabilistic context-free grammar (PCFG)
- Topic model (Latent Dirichlet Allocation (LDA))

Example: HMM-based POS tagging



Let K be the number of POS tags and V be the vocabulary size (ignore BOS and EOS for simplicity).

The transition probabilities can be computed using K categorical distributions $(\boldsymbol{\theta}_{\text{DT}}^{\text{TRANS}}, \boldsymbol{\theta}_{\text{NN}}^{\text{TRANS}}, \cdots)$, with the dimension K.

$$\boldsymbol{\theta}_{\text{DT}}^{\text{TRANS}} = (0.21, 0.27, 0.09, \cdots)$$

NN NNS ADJ

Similarly, the emission probabilities can be computed using K categorical distributions $(\boldsymbol{\theta}_{\text{DT}}^{\text{EMIT}}, \boldsymbol{\theta}_{\text{NN}}^{\text{EMIT}}, \cdots)$, with the dimension V.

$$\boldsymbol{\theta}_{\text{NN}}^{\text{EMIT}} = (0.012, 0.002, 0.005, \cdots)$$

sun rose cat

Outline

- Categorical and multinomial distributions
- Conjugacy and posterior predictive distribution
- LDA (Latent Dirichlet Applocation) as an application
- Gibbs sampling for inference

Categorical distribution: 1 observation

Suppose $\boldsymbol{\theta}$ is known.

Th probability of generating random variable $x \in \{1, \cdots, K\}$ is

$$p(x = k|\boldsymbol{\theta}) = \theta_k.$$

$0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{K} \theta_k = 1$.

For example, if $\boldsymbol{\theta} = (0.1, 0.6, 0.3)$, then $p(x = 2|\boldsymbol{\theta}) = \theta_2 = 0.6$.

Categorical distribution: N observations

The probability of generating a sequence of random variables with length N, $\boldsymbol{x} = (x_1, \cdots, x_N)$, is

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \theta_{x_i} = \prod_{k=1}^{K} \theta_k^{n_k},$$

where n_k is the number of times value k is observed in $\boldsymbol{x}$ ($\sum_{k=1}^{K} n_k = N$).

NOTE: $p(\boldsymbol{x}|\boldsymbol{\theta})$ does not depend on the ordering of $\boldsymbol{x}$ but only on the total number of observed values in the sequence (sufficient statistics).

Multinomial distribution: the probability of counts

Replacing $\boldsymbol{x} = (x_1, \cdots, x_N)$ with the counts $\boldsymbol{n} = (n_1, \cdots, n_K)$, we obtain the multinomial distribution:

$$\begin{aligned}
\mathrm{Multi}(\boldsymbol{n}|\boldsymbol{\theta}) &= \binom{N}{n_1 \cdots n_k} \prod_{k=1}^{K} \theta_k^{n_k} \\
&= \frac{N!}{n_1! \cdots n_k!} \prod_{k=1}^{K} \theta_k^{n_k} \\
&= \frac{\Gamma(N+1)}{\prod_{k=1}^{K} \Gamma(n_k+1)} \prod_{k=1}^{K} \theta_k^{n_k}.
\end{aligned}$$

Note $\Gamma(n) = (n-1)!$. The second term is the same as $p(\boldsymbol{x}|\boldsymbol{\theta})$. The first term is the combinatorial number of $\boldsymbol{x}$ mapped to $\boldsymbol{n}$.

NOTE: In NLP, the categorical distribution is often called a multinomial distribution.

Estimating θ from N observations

If $\boldsymbol{\theta}$ is unknown, we may want to infer it from $\boldsymbol{x}$.

First, the likelihood function is defined as follows:

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta}).$$

In maximum likelihood estimation (MLE), we estimate $\boldsymbol{\theta}^{\mathrm{ML}}$ that satisfies:

$$\boldsymbol{\theta}^{\mathrm{ML}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta}).$$

It has the following analytical solution (use the method of Lagrange multipliers):

$$\theta_k^{\mathrm{ML}} = \frac{n_k}{N}.$$

A problem with MLE

If $n_k = 0$, then $\theta_k^{\mathrm{ML}} = 0$ (zero frequency problem).

Would larger observed data fix the problem? Not necessarily. Natural language symbols usually follow a power low (cf. Zipf's law). There are always low-frequency symbols.

Bayes' theorem

Now we introduce $p(\boldsymbol{\theta})$, a prior distribution over $\boldsymbol{\theta}$. A prior distribution is an (arbitrary) distribution that expresses our beliefs about $\boldsymbol{\theta}$.

Bayes' theorem:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \\ &\propto p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \end{aligned}$$

$p(\boldsymbol{\theta}|\boldsymbol{x})$ is a posterior distribution, $p(\boldsymbol{x}|\boldsymbol{\theta})$ is a likelihood, and $p(\boldsymbol{\theta})$ is a prior distribution.
$p(\boldsymbol{\theta}|\boldsymbol{x})$ can be interpreted as the distribution of $\boldsymbol{\theta}$ *after* observing data.

A prior for categorical distributions

For a categorical distribution, we usually set the Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ as its prior becaues it has nice analytical properties.

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha}) \propto \mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}),$$

where

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(A)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$, $\alpha_k > 0$, and $A = \sum_{k=1}^{K} \alpha_k$. $\boldsymbol{\alpha}$ is a parameter given a priori (hyperparameter). We usually set $\alpha_i = \alpha_j$. α_k can be a real number (e.g. 1.5).

Note that the Dirichlet distribution resembles a categorical distribution:

$$\mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{n_k}.$$

Posterior distribution

Posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha}) \propto \mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ is (proof omitted):

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{\alpha}) &= \frac{\Gamma(N + A)}{\Gamma(n_1 + \alpha_1) \cdots \Gamma(n_K + \alpha_K)} \prod_{k=1}^{K} \theta_k^{n_k + \alpha_k - 1} \\ &= \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{n} + \boldsymbol{\alpha}). \end{aligned}$$

The posterior has the same form as the prior. The parameter $\boldsymbol{\alpha}$ is replaced with $\boldsymbol{n} + \boldsymbol{\alpha}$. This property is called conjugacy.

Maximum a posteriori (MAP) estimatoin

Maximum a posteriori (MAP) estimation employs $\boldsymbol{\theta}^{\mathrm{MAP}}$ that maximizes the posterior probability:

$$\begin{aligned}\boldsymbol{\theta}^{\mathrm{MAP}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{x}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}).\end{aligned}$$

For Dirichlet-categorical parameter $\boldsymbol{\theta}^{\mathrm{MAP}} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{n}+\boldsymbol{\alpha})$, the solution is

$$\theta_k^{\mathrm{MAP}} = \frac{n_k + \alpha_k - 1}{\sum_{k=1}^{K}(n_k + \alpha_k - 1)} = \frac{n_k + \alpha_k - 1}{N + A - K}.$$

Compare it with ML estimate $\theta_k^{\mathrm{ML}} = \frac{n_k}{N}$. The additional term $\alpha_k - 1$ is used to smooth the esimate (additive smoothing).

Marginal likelihood / posterior predictive distribution

Instead of using one estimate of $\boldsymbol{\theta}$, we now consider all possible values of $\boldsymbol{\theta}$. To do so, we integrate out $\boldsymbol{\theta}$ to obtain marginal likelihood:

$$p(\boldsymbol{x}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

The probability of generating new variable x' given $\boldsymbol{x}$ (predictive probability) is:

$$\begin{aligned} p(x'|\boldsymbol{x}) &= \int_{\boldsymbol{\theta}} p(x'|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(x'|\boldsymbol{\theta}) \frac{p(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{x})}\, d\boldsymbol{\theta} \end{aligned}$$

Thanks to conjugacy, we have analytical solutions for Dirichlet-categorical (next slide).

Marginal likelihood for Dirichlet-categorical

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\alpha}) &= \int_{\boldsymbol{\theta}} \mathrm{Cat}(\boldsymbol{x}|\boldsymbol{\theta})\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})\, d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \left(\prod_{k=1}^{K} \theta_k^{n_k} \right) \left(\frac{\Gamma(A)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \right) d\boldsymbol{\theta} \\
&= \frac{\Gamma(A)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \int_{\boldsymbol{\theta}} \prod_{k=1}^{K} \theta_k^{n_k + \alpha_k - 1}\, d\boldsymbol{\theta} \\
&= \frac{\Gamma(A)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(n_k + \alpha_k)}{\Gamma(N + A)} \\
&= \frac{\Gamma(A)}{\Gamma(N + A)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)}
\end{aligned}
$$

Posterior predictive for Dirichlet-categorical

The probability of generating x' after observing $\boldsymbol{x}$ is

$$
\begin{aligned}
p(x' = k' | \boldsymbol{x}, \boldsymbol{\alpha}) &= \frac{p(x' = k', \boldsymbol{x} | \boldsymbol{\alpha})}{p(\boldsymbol{x} | \boldsymbol{\alpha})} \\
&= \frac{\frac{\Gamma(A)}{\Gamma(N+1+A)} \prod_{k=1}^{K} \frac{\Gamma(n_k + I(k=k') + \alpha_k)}{\Gamma(\alpha_k)}}{\frac{\Gamma(A)}{\Gamma(N+A)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)}} \\
&= \frac{\Gamma(N+A)}{\Gamma(N+1+A)} \frac{\Gamma(n_{k'} + 1 + \alpha_{k'})}{\Gamma(n_{k'} + \alpha_{k'})} \\
&= \frac{n_{k'} + \alpha_{k'}}{N + A},
\end{aligned}
$$

where $I(\text{statement})$ is an indicator function. It gives 1 if statement is true; 0 otherwise.

In analogy with n_k (the count of observations), α_k is called a pseudo-count.

Sequential updates of x

Let $\boldsymbol{x}_{1:N} = (x_1, \cdots, x_N)$. $p(\boldsymbol{x}_{1:N}|\boldsymbol{\alpha})$ is the product of the probabilities of generating x one by one.

$$
\begin{aligned}
p(\boldsymbol{x}_{1:N}|\boldsymbol{\alpha}) &= p(x_N|\boldsymbol{x}_{1:N-1}, \boldsymbol{\alpha}) p(\boldsymbol{x}_{1:N-1}|\boldsymbol{\alpha}) \\
&= \frac{n_{x_N}(\boldsymbol{x}_{1:N-1}) + \alpha_{x_N}}{N - 1 + A} p(\boldsymbol{x}_{1:N-1}|\boldsymbol{\alpha}) \\
&= \frac{n_{x_N}(\boldsymbol{x}_{1:N-1}) + \alpha_{x_N}}{N - 1 + A} p(x_{N-1}|\boldsymbol{x}_{1:N-2}, \boldsymbol{\alpha}) p(\boldsymbol{x}_{1:N-2}|\boldsymbol{\alpha}) \\
&= \frac{n_{x_N}(\boldsymbol{x}_{1:N-1}) + \alpha_{x_N}}{N - 1 + A} \frac{n_{x_{N-1}}(\boldsymbol{x}_{1:N-2}) + \alpha_{x_{N-1}}}{N - 2 + A} p(\boldsymbol{x}_{1:N-2}|\boldsymbol{\alpha}) \\
&= \frac{\prod_{k=1}^{K} \prod_{i=1}^{n_k(\boldsymbol{x}_{1:N})} i - 1 + \alpha_k}{\prod_{i=1}^{N} i - 1 + A} \\
&= \frac{\Gamma(A)}{\Gamma(N + A)} \prod_{k=1}^{K} \frac{\Gamma(n_k(\boldsymbol{x}_{1:N}) + \alpha_k)}{\Gamma(\alpha_k)},
\end{aligned}
$$

where $n_k(\boldsymbol{x})$ is the number of times k appears in $\boldsymbol{x}$.

Exchangeability

As stated earlier, $p(\boldsymbol{x}|\boldsymbol{\theta})$ does not depend on the ordering of $\boldsymbol{x}$ but on the total number of times each value is observed in the sequence (sufficient statistics). Let us create $\boldsymbol{x}'$ by swapping an arbitrary pair of variables x_i, x_j in $\boldsymbol{x}$. Since the ordering does not matter, the following is hold:

$$p(\boldsymbol{x}'|\boldsymbol{\alpha}) = p(\boldsymbol{x}|\boldsymbol{\alpha})$$

This property is called exchangeability.

As we will see later, exchangeability makes inference easy.

Example: Sequential updates of x

Suppose random variable x takes one of three values: R, Y, G. The probability of generating the sequence $\boldsymbol{x} = (\mathrm{R}, \mathrm{R}, \mathrm{Y})$ is

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{0 + \alpha_{\mathrm{R}}}{0 + A} \frac{1 + \alpha_{\mathrm{R}}}{1 + A} \frac{0 + \alpha_{\mathrm{Y}}}{2 + A}$$

Verify that reordering does not affect the probability.

Generative story

For Dirichlet-categorical, $\boldsymbol{\theta}$ is first generated from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$ and then each random variable x_n is generated from a categorical distribution with parameter $\boldsymbol{\theta}$. This process can be summarized as follows:

$$\begin{aligned} \boldsymbol{\theta}|\boldsymbol{\alpha} &\sim \mathrm{Dir}(\boldsymbol{\alpha}) \\ x_n|\boldsymbol{\theta} &\sim \mathrm{Cat}(\boldsymbol{\theta}) \end{aligned}$$

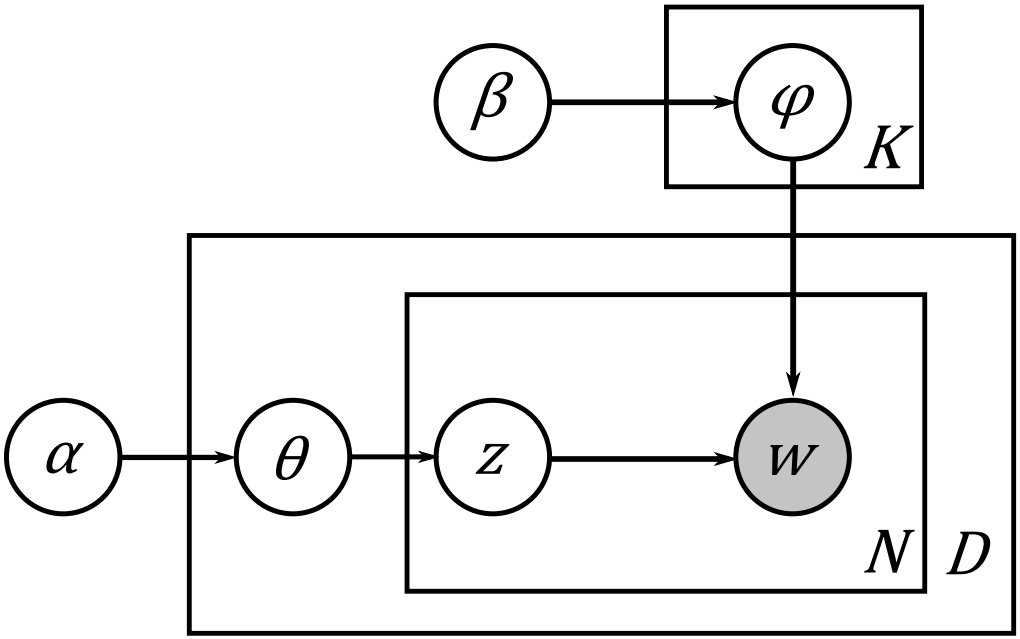
Application: LDA (Latent Dirichlet allocation)

NOTE: We use the modified model by Griffiths et al. (2004), not the original one by Blei et al. (2002).

We are given D documents. Document i contains N_i words, and $w_{i,j}$ is the j-th word in document i. LDA assumes that each document is a mixture of K topics. Document i is associated with a mixing proportion $\boldsymbol{\theta}_i$. Each topic k has a corresponding vocabulary distribution $\mathrm{Cat}(\boldsymbol{\varphi}_k)$. Each word $w_{i,j}$ is tied with a latent variable $z_{i,j} = k'$. $z_{i,j}$ is generated from $\mathrm{Cat}(\boldsymbol{\theta}_i)$, and then $w_{i,j}$ is generated from $\mathrm{Cat}(\boldsymbol{\varphi}_{k'})$.

Application: LDA (Latent Dirichlet allocation)

The generative story can be represented by a directed graph called a graphical model.



Each arrow indicates a dependency between variables. Each plate represents repetition with the number of repetition is shown in the corner. The shaded node represents an observed variable.

Application: LDA (Latent Dirichlet allocation)

The generative story is as follows:

1. For each topic $k \in \{1, \cdots, K\}$, draw vocabulary distribution $\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})$
2. For each document $i \in \{1, \cdots, D\}$, draw topic distribution $\boldsymbol{\theta}_i \sim \mathrm{Dir}(\boldsymbol{\alpha})$
3. For each word $j \in \{1, \cdots, N_i\}$ in document i,
    1. Draw topic assignment $z_{i,j} \sim \mathrm{Cat}(\boldsymbol{\theta}_i)$, and
    2. Draw word $w_{i,j} \sim \mathrm{Cat}(\boldsymbol{\varphi}_{z_{i,j}})$

Let $\boldsymbol{W}$ be the set of all observed words in the collection of documents and $\boldsymbol{Z}$ be the corresponding latent topic assignments. Given $\boldsymbol{W}$, we want to infer $\boldsymbol{Z}$.

Application: LDA (Latent Dirichlet allocation)

We begin with the joint probability:

$$p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \alpha, \boldsymbol{\beta}) = \prod_{k=1}^{K} p(\boldsymbol{\varphi_k} | \boldsymbol{\beta}) \times \prod_{i=1}^{D} \left(p(\boldsymbol{\theta_i} | \alpha) \prod_{j=1}^{N_i} p(z_{i,j} | \boldsymbol{\theta_i}) p(w_{i,j} | \boldsymbol{\varphi}, z_{i,j}) \right).$$

$\boldsymbol{\theta}_i$ and $\boldsymbol{\varphi}_k$ can be integrated out to obtain marginal likelihood.

$$\begin{aligned} p(\boldsymbol{W}, \boldsymbol{Z} | \alpha, \boldsymbol{\beta}) &= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\varphi}} p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \alpha, \boldsymbol{\beta}) \, d\boldsymbol{\theta} d\boldsymbol{\varphi} \\ &= \prod_{i=1}^{D} \left(\frac{\Gamma(\alpha_{(\cdot)})}{\Gamma(n_{i,(\cdot)}^{(\cdot)} + \alpha_{(\cdot)})} \prod_{k=1}^{K} \frac{\Gamma(n_{i,(\cdot)}^{k} + \alpha_k)}{\Gamma(\alpha_k)} \right) \\ &\quad \times \prod_{k=1}^{K} \left(\frac{\Gamma(\beta_{(\cdot)})}{\Gamma(n_{(\cdot),(\cdot)}^{k} + \beta_{(\cdot)})} \prod_{v} \frac{\Gamma(n_{(\cdot),v}^{k} + \beta_v)}{\Gamma(\beta_v)} \right), \end{aligned}$$

where $n_{i,v}^{k}$ is the number of times word type v in document i is tied with topic k. We can see that LDA is simply a composite of $D + K$ Dirichlet-categorical distributions.

Application: LDA (Latent Dirichlet allocation)

Now consider sequential updates of topic assignments and words.
It is easy to verify that exchangeability is hold.

Doc ID	Topic	Word	Prob.
1	Politics	Obama	$\frac{0+\alpha_{\text{Politics}}}{0+\alpha_{(\cdot)}} \times \frac{0+\beta_{\text{Obama}}}{0+\beta_{(\cdot)}}$
	Economy	FRB	$\frac{0+\alpha_{\text{Economy}}}{1+\alpha_{(\cdot)}} \times \frac{0+\beta_{\text{FRB}}}{0+\beta_{(\cdot)}}$
	Politics	election	$\frac{1+\alpha_{\text{Politics}}}{2+\alpha_{(\cdot)}} \times \frac{0+\beta_{\text{election}}}{1+\beta_{(\cdot)}}$
	Politics	Obama	$\frac{2+\alpha_{\text{Politics}}}{3+\alpha_{(\cdot)}} \times \frac{1+\beta_{\text{Obama}}}{2+\beta_{(\cdot)}}$
2	Economy	Yellen	$\frac{0+\alpha_{\text{Economy}}}{0+\alpha_{(\cdot)}} \times \frac{0+\beta_{\text{Yellen}}}{1+\beta_{(\cdot)}}$
	Economy	FRB	$\frac{1+\alpha_{\text{Economy}}}{1+\alpha_{(\cdot)}} \times \frac{1+\beta_{\text{FRB}}}{2+\beta_{(\cdot)}}$

Application: LDA (Latent Dirichlet allocation)

We want to infer $\boldsymbol{Z}$ that satisfies

$$\operatorname*{argmax}_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \operatorname*{argmax}_{\boldsymbol{Z}} p(\boldsymbol{W}, \boldsymbol{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta})$$

The number of possible assignments is $K^{|\boldsymbol{Z}|}$. The search space is too large for exhaustive search.

We resort to approximate inference using random walk.
Specificall,y we often use Gibbs sampling as a theoretically-grounded algorithm.

Sampling

Sampling is the procedure of obtaining a sample from a probability distribution. It is typically implemented by modifying the values of the rand function r $(0 \leq r < 1)$.

Suppose x is a sample from $\boldsymbol{\theta} = (0.1, 0.6, 0.3)$. Using r, we can create a sample x as follows:

$$x = \begin{cases} 1 & 0 \leq r < 0.1, \\ 2 & 0.1 \leq r < 0.7, \text{and} \\ 3 & 0.7 \leq r < 1. \end{cases}$$

Sampling from Dirichlet-categorical posterior predictive

Recall the predictive predictive distribution of Dirichlet-categorical is

$$p(x' = k' | \boldsymbol{x}, \boldsymbol{\alpha}) = \frac{n_{k'}(\boldsymbol{x}) + \alpha_{k'}}{n(\boldsymbol{x}) + \alpha_{(\cdot)}}.$$

Let $\boldsymbol{x} = (\mathrm{R}, \mathrm{R}, \mathrm{Y}, \mathrm{R}, \mathrm{R})$, and $\boldsymbol{\alpha} = (\alpha_{\mathrm{R}}, \alpha_{\mathrm{Y}}, \alpha_{\mathrm{G}}) = (1, 1, 1)$. Then

$$\begin{aligned}
p(x' = \mathrm{R} | \boldsymbol{x}, \boldsymbol{\alpha}) &= \frac{4 + 1}{5 + 3} = 0.625 \\
p(x' = \mathrm{Y} | \boldsymbol{x}, \boldsymbol{\alpha}) &= \frac{1 + 1}{5 + 3} = 0.25 \\
p(x' = \mathrm{G} | \boldsymbol{x}, \boldsymbol{\alpha}) &= \frac{0 + 1}{5 + 3} = 0.125.
\end{aligned}$$

Draw a sample from $\boldsymbol{\theta} = (0.625, 0.25, 0.125)$, and that is the sample we want to obtain.

Gibbs sampling

LDA is a composite of Dirichlet-categorical distributions $p(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta})$. Directly sampling from this complex distribution is difficult.
However, it is easy to draw a sample from $p(z_i|\boldsymbol{W}, \boldsymbol{Z}^{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where $\boldsymbol{Z}^{-i} = \boldsymbol{Z} \setminus \{z_i\}$. i is now a corpus-wide index.
Gibbs sampling is a method of obtaining samples of $\boldsymbol{Z}$ by repeatedly sampling from $p(z_i|\boldsymbol{W}, \boldsymbol{Z}^{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.
The algorithmic overview of Gibbs sampling is as follows:

1. initialize $\boldsymbol{Z}$
2. For each iteration $\tau = 1, \ldots, T$,
   - randomly select i
   - sample z_i' from $p(z_i|\boldsymbol{W}, \boldsymbol{Z}^{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta})$
   - update $\boldsymbol{Z}$ by replacing z_i with z_i'

After a sufficient number of iterations, $\boldsymbol{Z}$ can be seen as samples from $p(\boldsymbol{Z}|\boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

Exchangeability again

Gibbs sampling is applicable if we can easily draw a sample from $p(z_i|\boldsymbol{Z}^{-i})$.

Due to exchangeability, the probability of observing the sequence $(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_N)$ is the same as the probability of observing the sequence $(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N, z_i)$.

This means that $p(z_i|\boldsymbol{Z}^{-i})$ is is the posterior predictive distribution of z_i after observing $(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N)$, the sequence excluding z_i.

Gibbs sampling for LDA 1/2

$$
\begin{aligned}
& p(z_i = k' | \boldsymbol{Z}^{-i}, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
\propto \quad & p(z_i = k', \boldsymbol{Z}^{-i}, \boldsymbol{W} | \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
= \quad & \prod_{j=1}^{D} \left(\frac{\Gamma(\alpha_{(\cdot)})}{\Gamma(n^{(\cdot)}_{j,(\cdot)}(\boldsymbol{Z}^{-i}) + I(w_i \in \boldsymbol{W}_j) + \alpha_{(\cdot)})} \right. \\
& \left. \prod_{k=1}^{K} \frac{\Gamma(n^{k}_{j,(\cdot)}(\boldsymbol{Z}^{-i}) + I(k' = k \,\&\, w_i \in \boldsymbol{W}_j) + \alpha_k)}{\Gamma(\alpha_k)} \right) \\
& \times \prod_{k=1}^{K} \left(\frac{\Gamma(\beta_{(\cdot)})}{\Gamma(n^{k}_{(\cdot),(\cdot)}(\boldsymbol{Z}^{-i}) + I(k' = k) + \beta_{(\cdot)})} \right. \\
& \left. \prod_{v} \frac{\Gamma(n^{k}_{(\cdot),v}(\boldsymbol{Z}^{-i}) + I(k' = k \,\&\, w_i = v) + \beta_v)}{\Gamma(\beta_v)} \right).
\end{aligned}
$$

$\boldsymbol{W}_j$ is the sequence of words in the j-th document. Most terms are irrelevant with z_i and can be dropped.

Gibbs sampling for LDA 2/2

$$
\begin{aligned}
& p(z_i = k' | \boldsymbol{Z}^{-i}, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
\propto & \frac{\Gamma(n^{(\cdot)}_{j',(\cdot)}(\boldsymbol{Z}^{-i}) + \alpha_{(\cdot)})}{\Gamma(n^{(\cdot)}_{j',(\cdot)}(\boldsymbol{Z}^{-i}) + 1 + \alpha_{(\cdot)})} \times \frac{\Gamma(n^{k'}_{j',(\cdot)}(\boldsymbol{Z}^{-i}) + 1 + \alpha_{k'})}{\Gamma(n^{k'}_{j',(\cdot)}(\boldsymbol{Z}^{-i}) + \alpha_{k'})} \\
& \times \frac{\Gamma(n^{k'}_{(\cdot),(\cdot)}(\boldsymbol{Z}^{-i}) + \beta_{(\cdot)})}{\Gamma(n^{k'}_{(\cdot),(\cdot)}(\boldsymbol{Z}^{-i}) + 1 + \beta_{(\cdot)})} \times \frac{\Gamma(n^{k'}_{(\cdot),v'}(\boldsymbol{Z}^{-i}) + 1 + \beta_{v'})}{\Gamma(n^{k'}_{(\cdot),v'}(\boldsymbol{Z}^{-i}) + \beta_{v'})} \\
= & \frac{n^{k'}_{j',(\cdot)}(\boldsymbol{Z}^{-i}) + \alpha_{k'}}{n^{(\cdot)}_{j',(\cdot)}(\boldsymbol{Z}^{-i}) + \alpha_{(\cdot)}} \times \frac{n^{k'}_{(\cdot),v'}(\boldsymbol{Z}^{-i}) + \beta_{v'}}{n^{k'}_{(\cdot),(\cdot)}(\boldsymbol{Z}^{-i}) + \beta_{(\cdot)}},
\end{aligned}
$$

where $v' = w_i$, and j' is the index of the document w_i belongs to. The first term is the posterior predictive probability of adding topic assignment with value k' to document j' (the denominator can be dropped because it is constant wrt z_i). The second term is the posterior predictive probability of drawing word type v' from the vocabulary distribution of topic k'.

Intuitive explanation of Gibbs sampling for LDA

$$p(z_i = k' | \boldsymbol{Z}^{-i}, \boldsymbol{W}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{n^{k'}_{j',(\cdot)}(\boldsymbol{Z}^{-i}) + \alpha_{k'}}{n^{(\cdot)}_{j',(\cdot)}(\boldsymbol{Z}^{-i}) + \alpha_{(\cdot)}} \times \frac{n^{k'}_{(\cdot),v'}(\boldsymbol{Z}^{-i}) + \beta_{v'}}{n^{k'}_{(\cdot),(\cdot)}(\boldsymbol{Z}^{-i}) + \beta_{(\cdot)}}$$

- To make the first term larger, we need to select k' that appears often in document j'.
- To make the second term larger, we need to select k' from which $w_i = v'$ is often drawn.
- z_i reflects a balance between the first and second terms.
- After repeating the procedure, $\boldsymbol{Z}$ will generally satisfy these conditions.