Text Mining for the Social Sciences
Lecture 5: Discrete Probability Models

Stephen Hansen

Introduction

In the last lecture, we viewed topics as orthogonal dimensions that explained most of the variance in the document-term matrix.

In this lecture, we introduce probability models for discrete data.

These can then be used to formally incorporate uncertainty into topic models.

Language Model

We can view a document d as an (ordered) list of words $\mathbf{w}_d = (w_{d,1}, \ldots, w_{d,N_d})$.

A language model is a probability distribution over this list.

We can express the joint probability in terms of the conditional probabilities as follows

$$\Pr[w_{d,1}, \ldots, w_{d,N_d}] =$$
$$\Pr[w_{d,1}] \Pr[w_{d,2} \mid w_{d,1}] \Pr[w_{d,3} \mid w_{d,1}, w_{d,2}] \ldots \Pr[w_{d,N_d} \mid w_{d,1}, \ldots, w_{d,N_d-1}].$$

Modeling the full set of dependencies is difficult, so some simplifications nearly always made.

N-gram Models

Recall the bag of words assumption that word order does not matter. In the context of language models, this is called the *unigram model*.

We assume that

$$\Pr[w_{d,1}, \ldots, w_{d,N_d}] = \Pr[w_{d,1}] \Pr[w_{d,2}] \Pr[w_{d,3}] \ldots \Pr[w_{d,N_d}].$$

Moreover, we assume that $w_{d,n}$ is drawn iid.

In a bigram model, the assumption is

$$\Pr[w_{d,1}, \ldots, w_{d,N_d}] = \Pr[w_{d,1}] \Pr[w_{d,2} \mid w_{d,1}] \ldots \Pr[w_{d,N_d} \mid w_{d,N_d-1}].$$

We can also extend the model to trigrams, 4-grams, etc. In practice higher-order models rarely used.

Motivation for Unigram Model

In this course, we will study the unigram model.

Recall that we are most interested in content analysis.

'I like dogs', 'My wife likes dogs', and 'My grandmother thinks dogs are dirty' are all clearly different sentences, but all are basically about dogs.

N-gram models used for tasks like speech recognition, machine translation, and spelling correction.

Categorical Distribution

Suppose the words in document d are all drawn from some vocabulary indexed by v with V terms.

Let $\Pr[w_{d,n}] = \theta_v$ and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_V)$.

This defines a *categorical distribution* parametrized by $\boldsymbol{\theta}$.

Under the unigram model, we can write

$$\Pr[w_{d,1}, \ldots, w_{d,N_d}] = \prod_v \theta_v^{n_d(v)}$$

where $n_d(v)$ is the number of times term v appears in document d.

Note confusion in literature between categorical and multinomial distributions. Here we express the probability of a specific sequence of words, not the probability of observing a particular aggregate word count.

Maximum Likelihood Inference

To recover the maximum likelihood estimate of $\boldsymbol{\theta}$ given data, maximize Lagrangean

$$\ell(\boldsymbol{\theta}, \lambda) = \underbrace{\sum_v n_d(v) \log(\theta_v)}_{\text{log-likelihood}} + \lambda \underbrace{\left(1 - \sum_v \theta_v\right)}_{\text{Constraint on } \boldsymbol{\theta}}.$$

First order condition is $\frac{n_d(v)}{\theta_v} - \lambda = 0 \Rightarrow \theta_v = \frac{n_d(v)}{\lambda}$.

Constraint gives $\frac{\sum_v n_d(v)}{\lambda} = 1 \Rightarrow \lambda = \sum_v n_d(v) \equiv N_d$.

So MLE estimate is $\widehat{\theta}^{MLE} = \frac{n_d(v)}{N_d}$.

This is the observed frequency of term v in document d.

Black Swan Paradox

To maximize the probability of the observed data, we get parameters that exactly match the observed frequencies.

What would the model predict is the probability of seeing an unobserved term?

This is sometimes called the *black swan paradox*. Europeans assumed that the fact that they had never observed a black swan implied black swans could not exist.

Since the document-term matrix is sparse, the black swan paradox will be particularly relevant for text mining.

Bottom line is we need to incorporate some additional uncertainty in our inference procedure to not drive our beliefs about unobserved events to zero.

Bayesian Inference

One way around the issue is to adopt a Bayesian inference approach.

Recall that Bayes' rule states that

$$\Pr[\boldsymbol{\theta} \mid \mathbf{w}_d] = \frac{\Pr[\mathbf{w}_d \mid \boldsymbol{\theta}] \Pr[\boldsymbol{\theta}]}{\Pr[\mathbf{w}_d]}$$

where

- $\Pr[\boldsymbol{\theta} \mid \mathbf{w}_d]$ is the posterior distribution.
- $\Pr[\mathbf{w}_d \mid \boldsymbol{\theta}]$ is the likelihood function.
- $\Pr[\boldsymbol{\theta}]$ is the prior distribution on the parameter vector.
- $\Pr[\mathbf{w}_d]$ is a normalizing constant sometimes called the evidence.

We introduce initial parameter uncertainty with $\Pr[\boldsymbol{\theta}]$.

Dirichlet Prior

We've seen that $\Pr[\mathbf{w}_d \mid \boldsymbol{\theta}] = \prod_v \theta_v^{n_d(v)}$.

We want to find prior distribution over the simplex that is conjugate to the categorical distribution.

The Dirichlet distribution is parametrized by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_V)$ and has a pdf $\mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$ that satisfies

$$\mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \propto \prod_v \theta_v^{\alpha_v - 1}.$$

$\mathrm{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$ is normalized by $B(\boldsymbol{\alpha}) \equiv \frac{\prod_{v=1}^{V} \Gamma(\alpha_v)}{\Gamma\left(\sum_{v=1}^{V} \alpha_v\right)}$.

Interpreting the Dirichlet

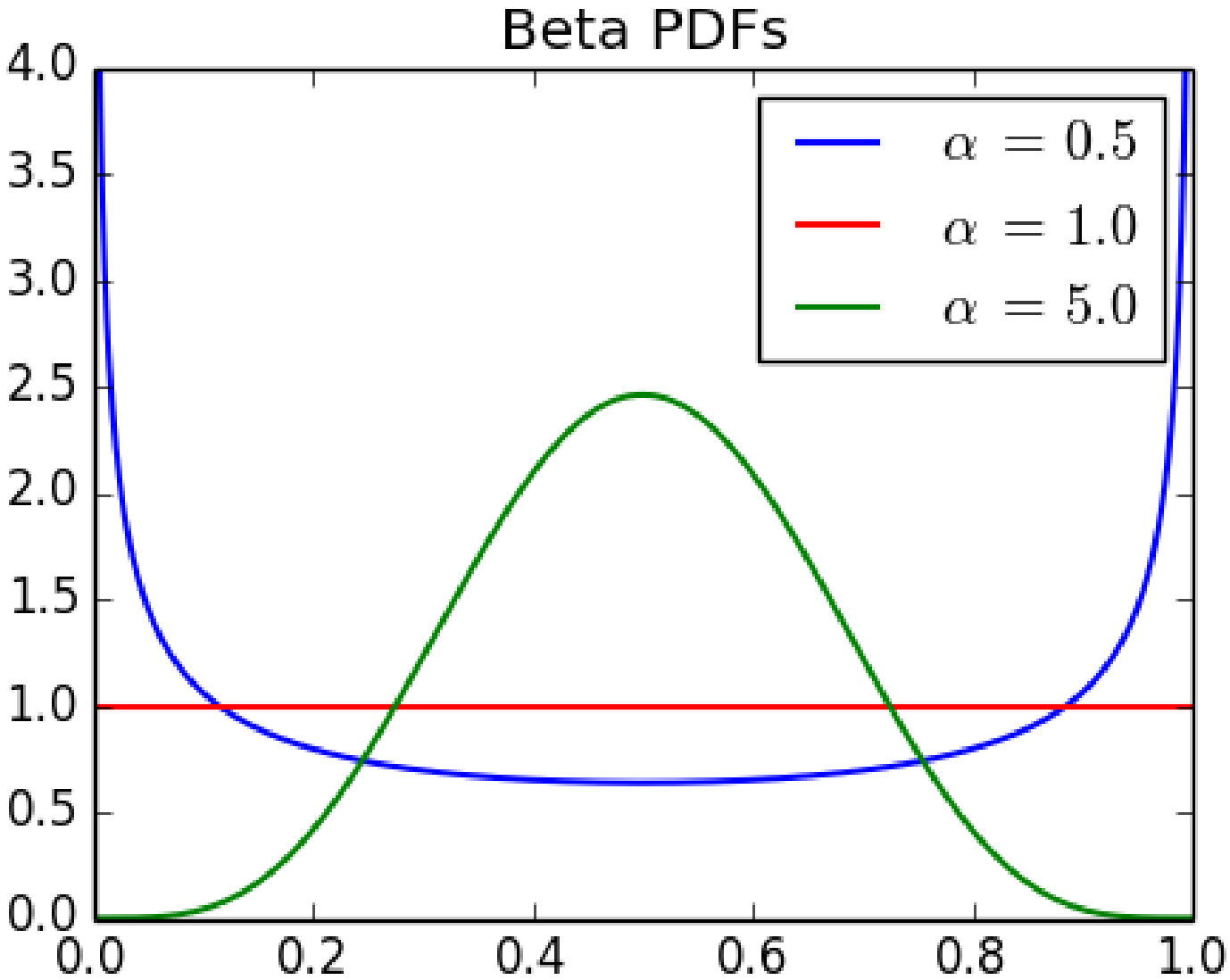
For this course, we will typically consider symmetric Dirichlet priors in which $\alpha_v = \alpha$ for all v. Agnostic about favoring one component over another.

Here the α parameter measures the concentration of distribution on the center of the simplex, where the mass on each term is more evenly spread out:

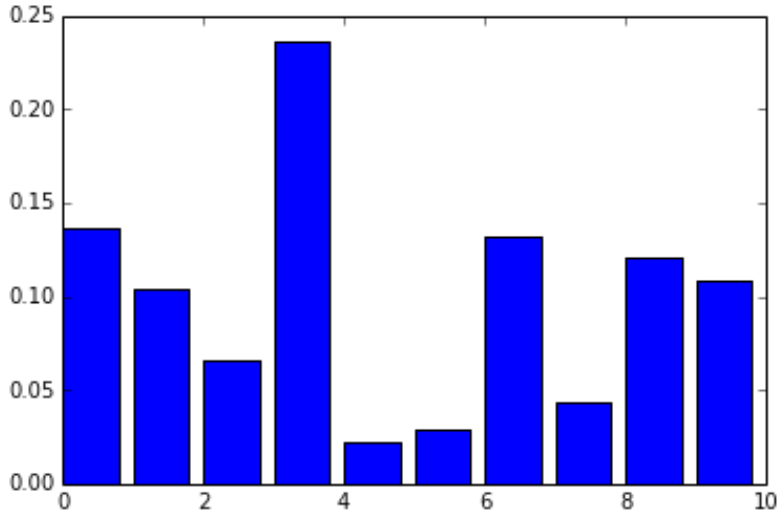
1. $\alpha = 1$ is a uniform distribution.
2. $\alpha > 1$ puts relatively more weight in center of simplex.
3. $\alpha < 1$ puts relatively more weight on corners of simplex.

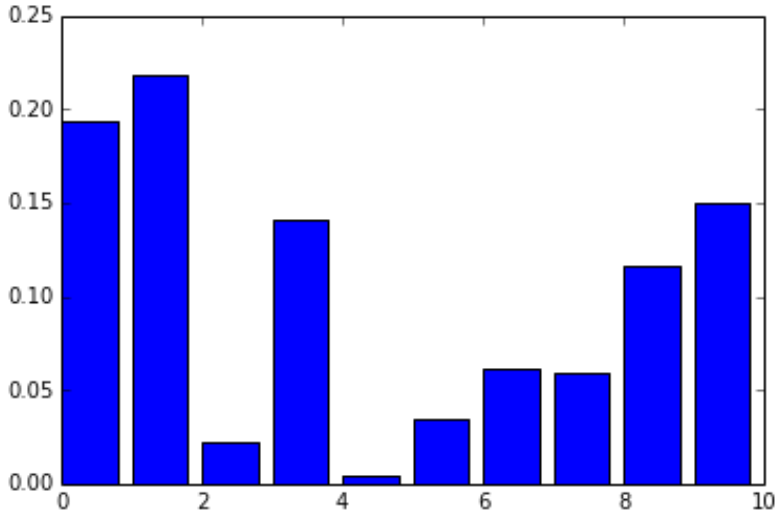
When $V = 2$, the Dirichlet becomes the beta distribution.

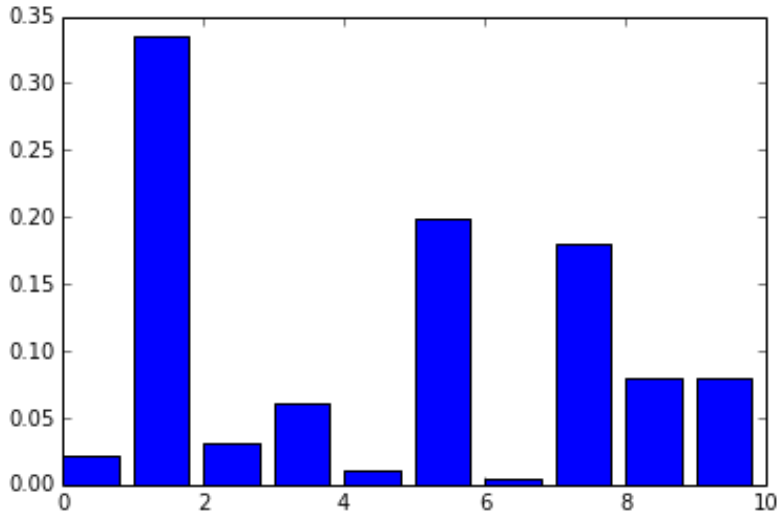
Beta with Different Parameters

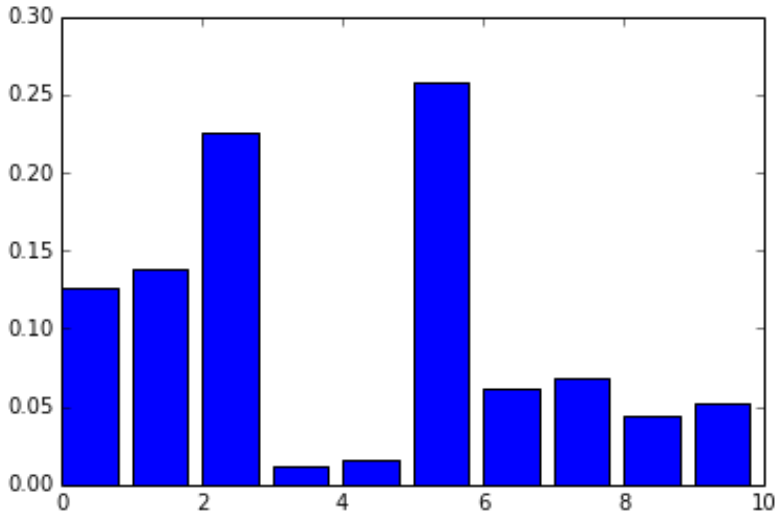


Draws from Dirichlet with $\alpha = 1$

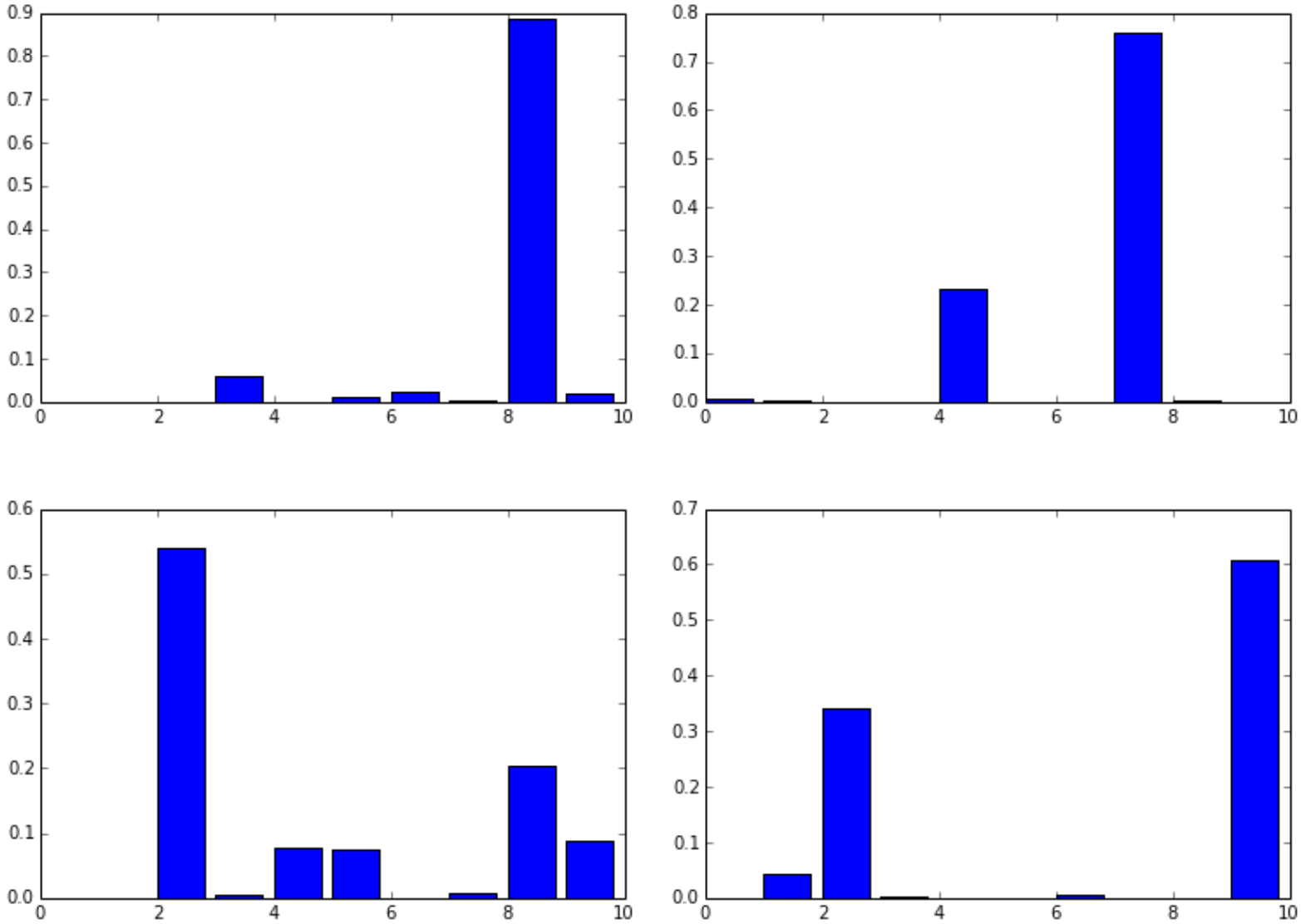




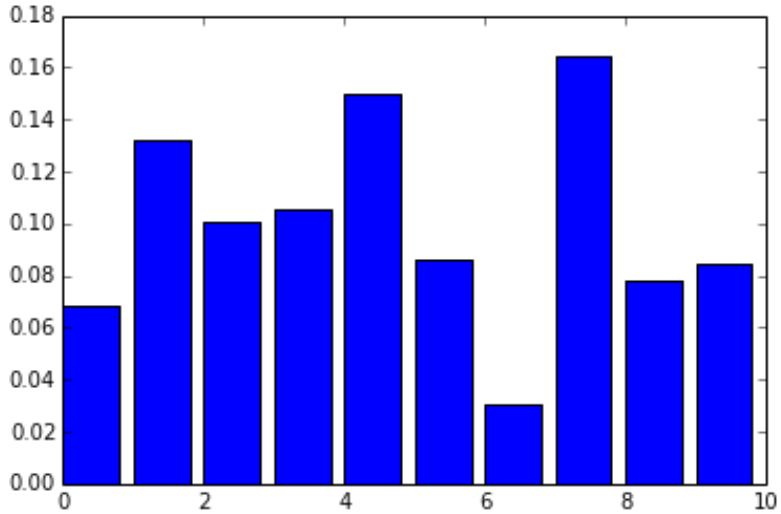


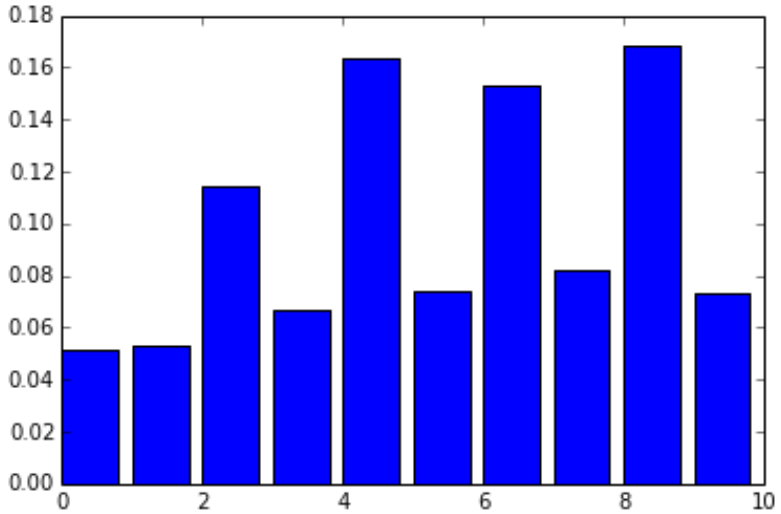


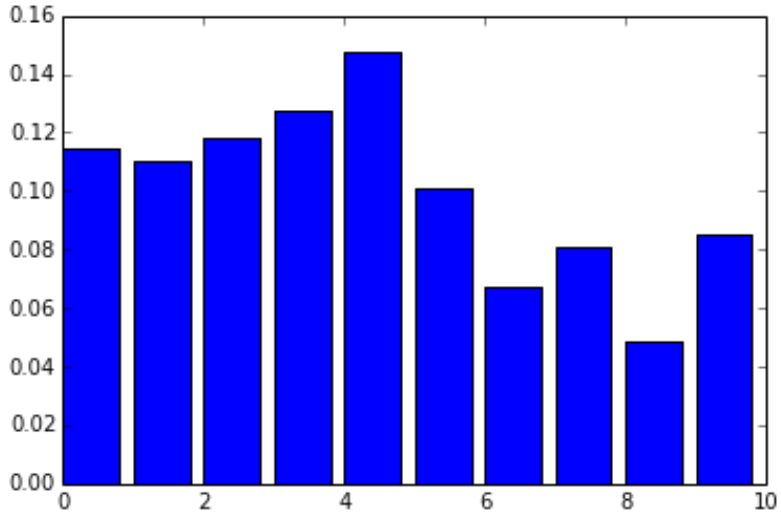
Draws from Dirichlet with $\alpha = 0.1$

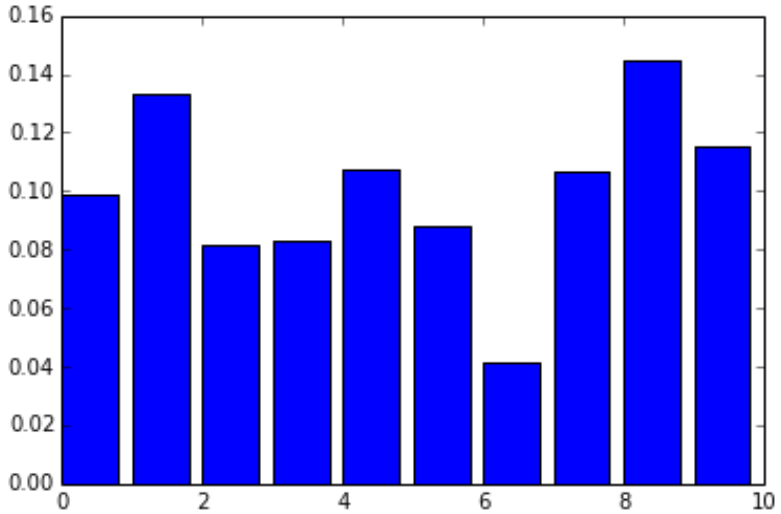


Draws from Dirichlet with $\alpha = 10$









Posterior Distribution

$$\Pr[\boldsymbol{\theta} \mid \mathbf{w}_d] \propto \Pr[\mathbf{w}_d \mid \boldsymbol{\theta}] \Pr[\boldsymbol{\theta}] \propto \prod_{v=1}^{V} \theta_v^{n_d(v)} \prod_{v=1}^{V} \theta_v^{\alpha_v - 1} = \prod_{v=1}^{V} \theta_v^{n_d(v) + \alpha_v - 1}.$$

Posterior is a Dirichlet with parameters $[n_d(1) + \alpha_1, \ldots, n_d(V) + \alpha_V]$.

We have simply added the empirical frequencies to the Dirichlet parameters to form posterior.

The Dirichlet parameters are often called *pseudo-counts*, and can be viewed as observations made before $\mathbf{w}_d$.

The higher are the parameters, the less the data affects the posterior.

Evidence

Later on, computing the evidence $\Pr[\mathbf{w}_d] = \int_{\boldsymbol{\theta}} \Pr[\mathbf{w}_d \mid \boldsymbol{\theta}] \Pr[\boldsymbol{\theta} \mid \alpha] d\boldsymbol{\theta}$ will be important.

By Bayes' Rule

$$\Pr[\mathbf{w}_d] = \frac{\Pr[\mathbf{w}_d \mid \boldsymbol{\theta}] \Pr[\boldsymbol{\theta}]}{\Pr[\boldsymbol{\theta} \mid \mathbf{w}_d]}$$

By the above we know that

$$\Pr[\mathbf{w}_d \mid \boldsymbol{\theta}] \Pr[\boldsymbol{\theta}] = \frac{\prod_v \theta_v^{n_d(v)} \prod_v \theta_v^{\alpha-1} \Gamma(V\alpha)}{\prod_v \Gamma(\alpha_v)}$$

$$\Pr[\boldsymbol{\theta} \mid \mathbf{w}_d] = \frac{\prod_v \theta_v^{n_d(v)+\alpha-1} \Gamma(V\alpha + N_d)}{\prod_v \Gamma(\alpha_v + n_d(v))}$$

so that

$$\Pr[\mathbf{w}_d] = \frac{\Gamma(\alpha V)}{\Gamma(\alpha V + N_d)} \frac{\prod_v \Gamma(\alpha + n_d(v))}{\prod_v \Gamma(\alpha)}.$$

Predictive Distribution

Recall the predictive distribution is a distribution over new data given observed data (rather than the unknown parameter $\boldsymbol{\theta}$).

What's the probability that a $N_d + 1$th word drawn for document d is term v?

$$\Pr[\, w_{d,N_d+1} = v \mid \mathbf{w}_d \,] = \int \Pr[\, w_{d,N_d+1} = v \mid \boldsymbol{\theta} \,] \Pr[\, \boldsymbol{\theta} \mid \mathbf{w}_d \,] d\boldsymbol{\theta} =$$

$$\int \theta_v \Pr[\, \theta_v \mid \mathbf{w}_d \,] d\theta_v = \mathbb{E}[\, \theta_v \mid \mathbf{w}_d \,].$$

One can show that the expectation of a Dirichlet is

$$\mathbb{E}[\, \theta_v \mid \mathbf{w}_d \,] = \frac{\alpha_v + n_d(v)}{\sum_{v=1}^{V} \alpha_v + N_d}.$$

Pseudo-counts act to smooth predictive likelihood and relax black swan paradox.

Topics

We can use categorical distribution to model topics in corpus.

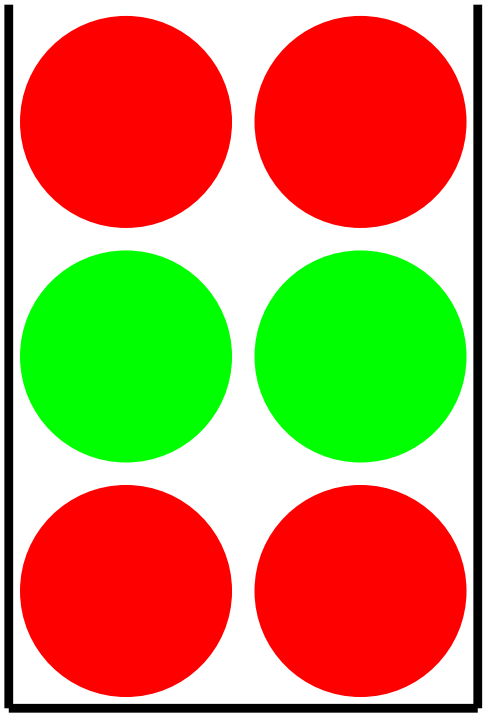
Suppose there are K topics, and let $\beta_k \in \Delta^V$ be the kth topic.

β_k then defines the parameters of a categorical distribution, where $\beta_{k,v}$ is the probability that the vth term appears in the kth topic.

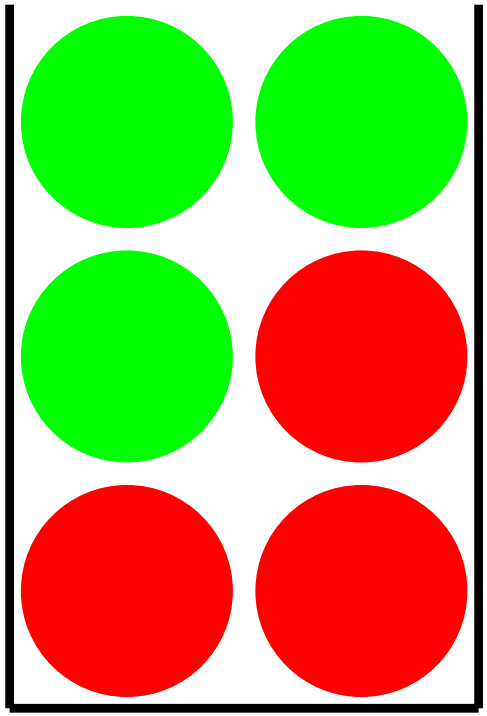
Informally, topics are weighted word lists where the weights are constrained to sum to 1.

Note that any given word potentially has positive probability in all topics, capturing polysemy; any given topic potentially gives all words positive probability, capturing synonomy.

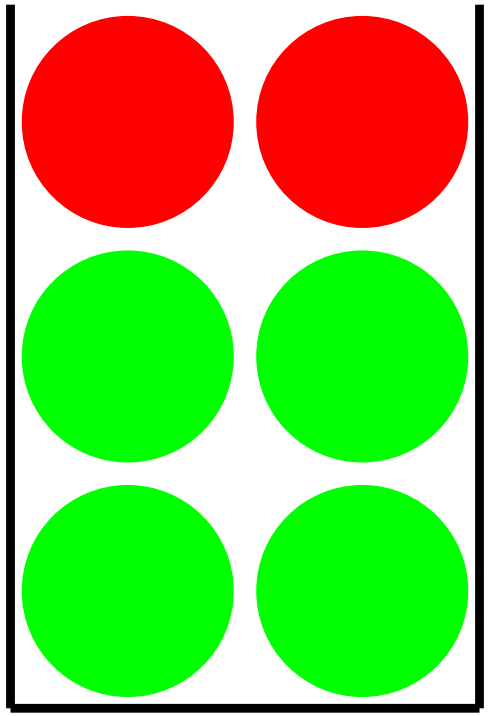
Topics as Urns



Topic 1



Topic 2



Topic 3

Modeling Documents

To complete the description of a topic model, we have to decide how to model documents.

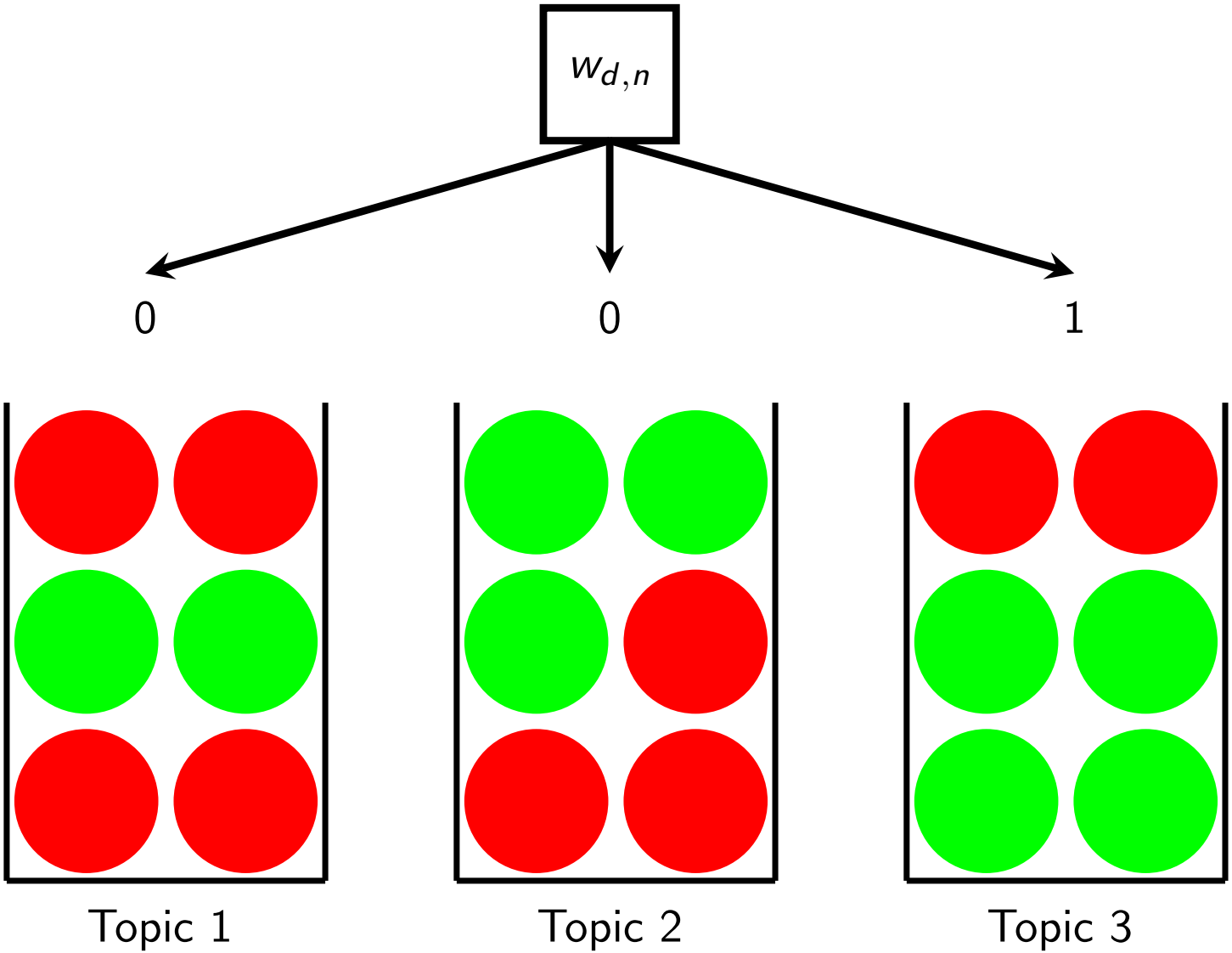
An important idea in topic modeling is that documents are composed of latent topics, and the words we observe are random draws from those topics.

A simple topic model is one in which each document d in the corpus is assigned a single topic z_d drawn from a categorical distribution with parameters $\boldsymbol{\theta}$.

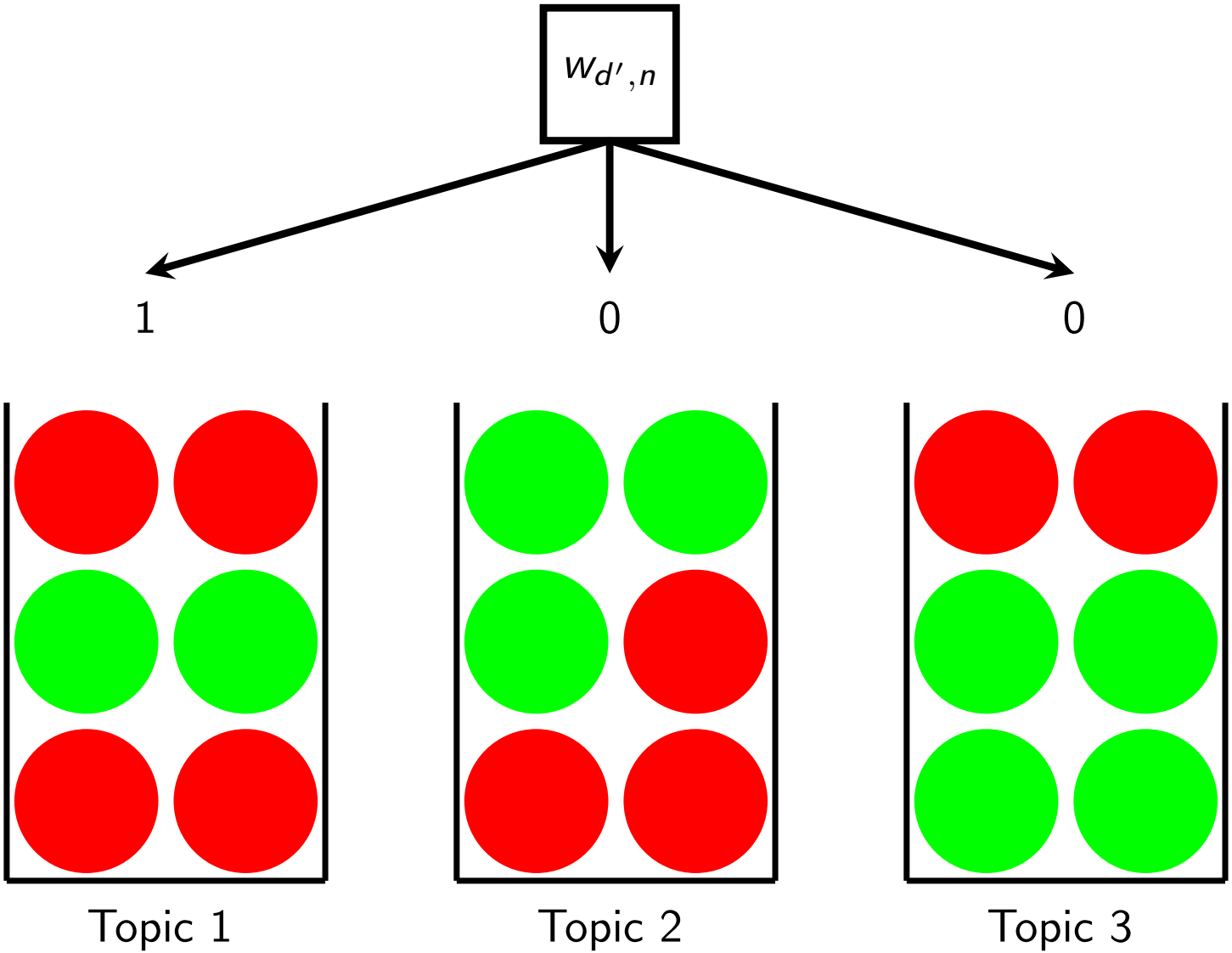
Each element of $\mathbf{w}_d$ is then drawn from β_{z_d}.

This defines a multinomial mixture model.

Mixture Model



Mixture Model



Mixed-Membership Model

A feature of the mixture model is that every word in a document is forced to be drawn from the same distribution.

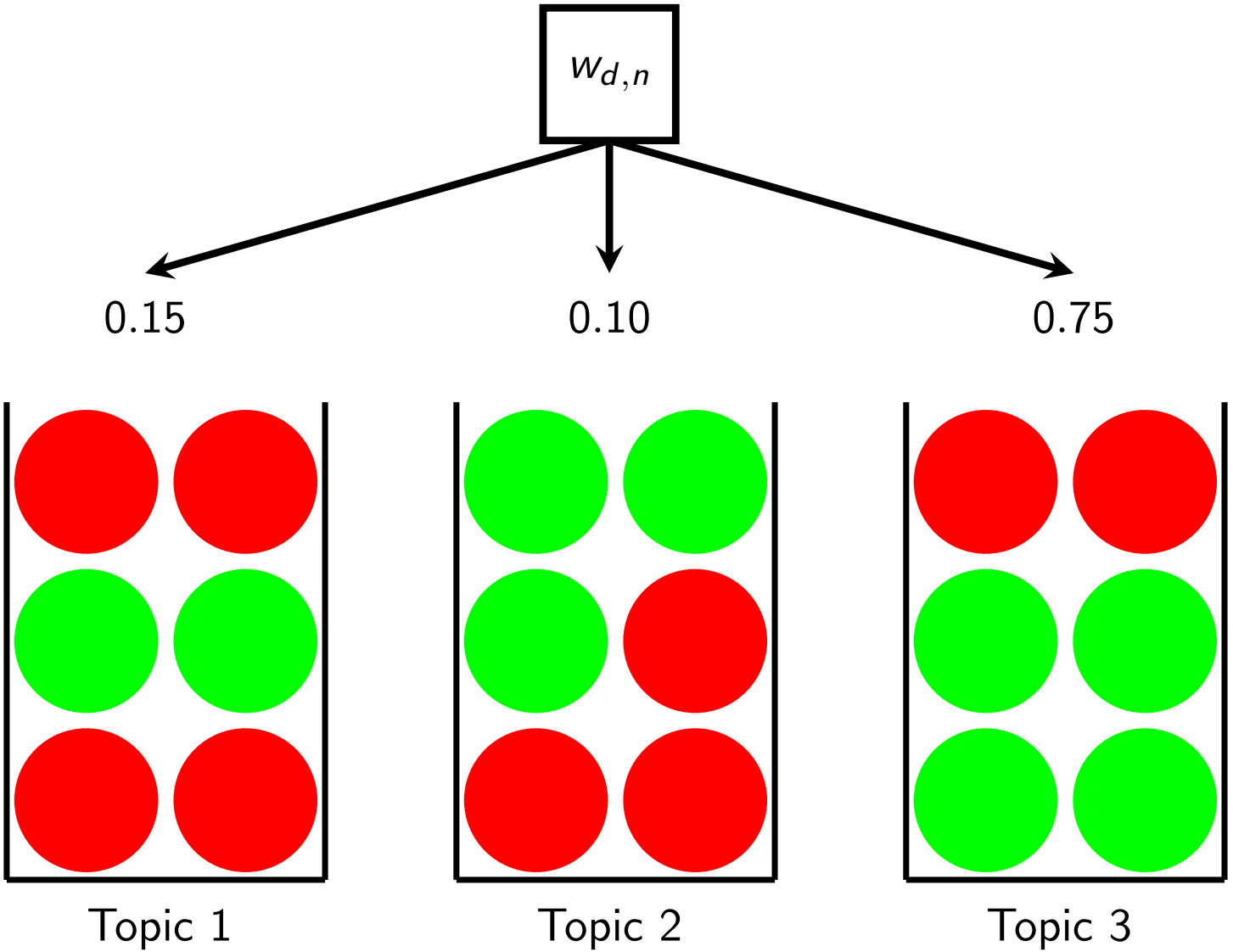
An alternative to this model is that documents can cover multiple topics (but may be inclined to cover some topics more than others).

This is called a *mixed-membership model* because the same document can belong to multiple topics.

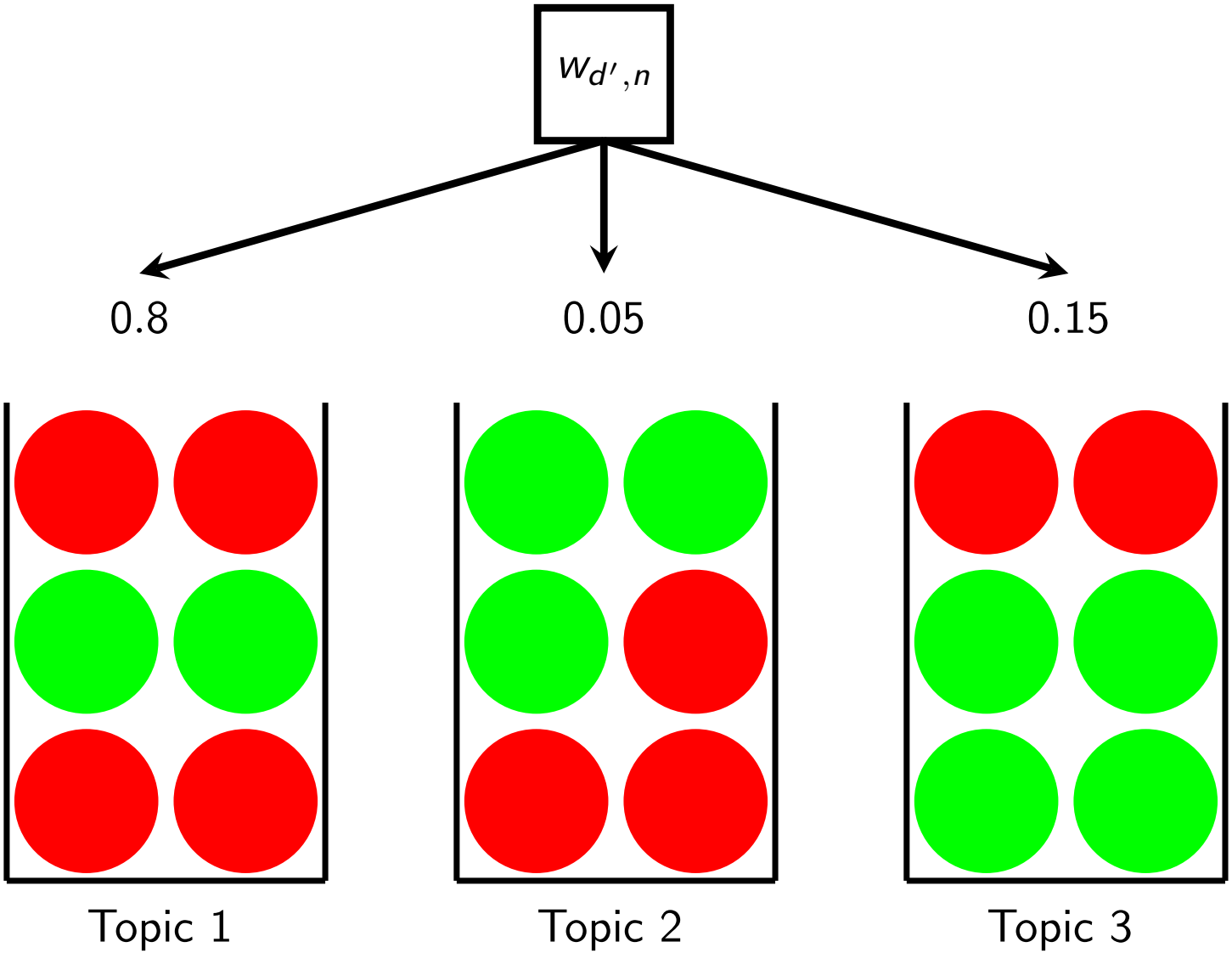
However, each word $w_{d,n}$ in a document belongs to a single topic $z_{d,n}$.

Let $\mathbf{z}_d = (z_{d,1}, \ldots, z_{d,N_d})$.

Mixed-Membership Model



Mixed-Membership Model



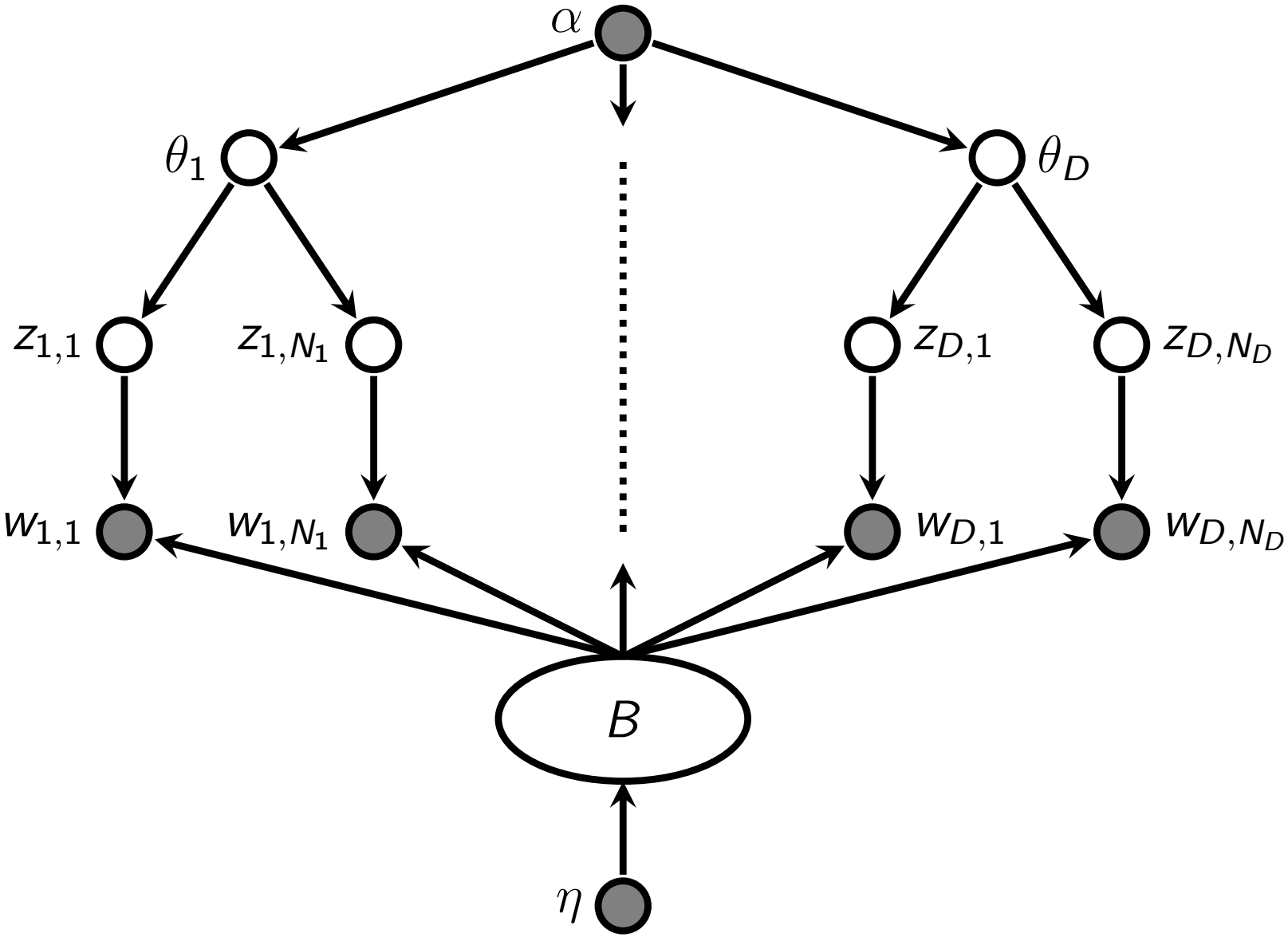
Latent Dirichlet Allocation

Latent Dirichlet Allocation (Blei, Ng, Jordan 2003) is a hugely influential mixed-membership topic model.

The data generating process of LDA is the following:

1. Draw β_k independently for $k = 1, \ldots, K$ from Dirichlet(η). (Note that original model did not have Dirichlet prior).
2. Draw θ_d independently for $d = 1, \ldots, D$ from Dirichlet(α).
3. Each word $w_{d,n}$ in document d is generated from a two-step process:
   - 3.1 Draw topic assignment $z_{d,n}$ from θ_d.
   - 3.2 Draw $w_{d,n}$ from $\beta_{z_{d,n}}$.

LDA as a Directed Graph



Graph Properties

Let $B = (\beta_1, \ldots, \beta_K)$ and $\Theta = (\theta_1, \ldots, \theta_D)$.

Object	Parents	Children
α	$\emptyset$	Θ
θ_d	α	$\mathbf{z}_d$
$z_{d,n}$	θ_d	$w_{d,n}$
$w_{d,n}$	$z_{d,n}, B$	$\emptyset$
β_k	η	$\mathbf{w}_1, \ldots, \mathbf{w}_D$
η	$\emptyset$	B

Joint probability factors as

$$\Pr[\mathbf{w}, \mathbf{z}, \Theta, B \mid \alpha, \eta] = \Pr[B \mid \eta] \prod_d \Pr[\mathbf{w}_d \mid \mathbf{z}_d, B] \Pr[\mathbf{z}_d \mid \theta_d] \Pr[\theta_d \mid \alpha]$$

Why is LDA so Popular?

Serves as building block for models that incorporate more complex features (e.g. dynamics and correlations in topics).

Applications outside of text mining (e.g. genetics and network detection).

Highly successful illustration of the power of Bayesian network models.

Inference algorithms are easy to implement, and produce interpretable topics.

Conclusion

In this lecture, we've moved to consider text as a statistical object.

Words are generated by latent topics that we'd like to infer to determine a document's content.

Next lecture discusses inference for LDA.