

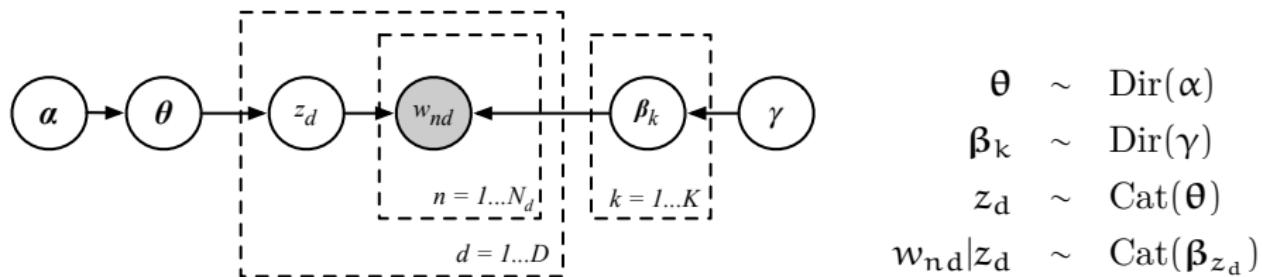
# Latent Dirichlet Allocation for Topic Modeling

Ayush Tewari

December 2nd, 2025

Adapted from Carl Edward Rasmussen

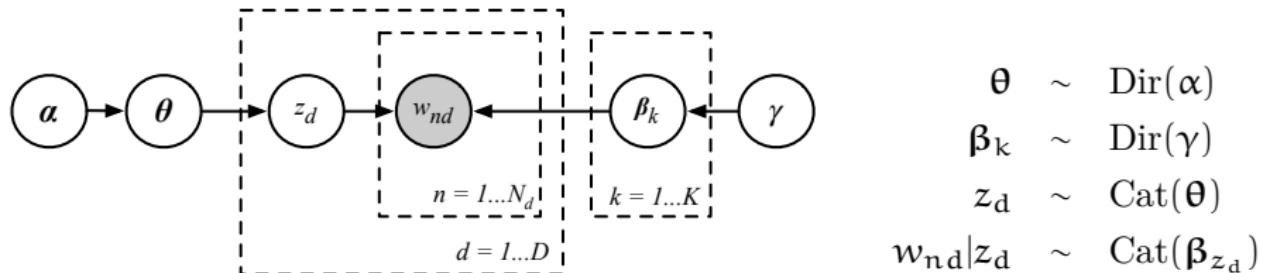
# Limitations of the mixture of categoricals model



A generative view of the mixture of categoricals model

- 1 Draw a distribution  $\Theta$  over  $K$  topics from a  $\text{Dirichlet}(\alpha)$ .

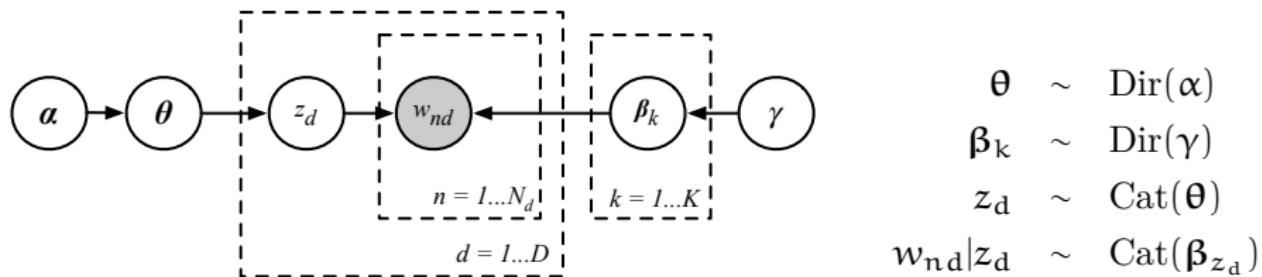
# Limitations of the mixture of categoricals model



A generative view of the mixture of categoricals model

- 1 Draw a distribution  $\theta$  over  $K$  topics from a  $\text{Dirichlet}(\alpha)$ .
- 2 For each topic  $k$ , draw a distribution  $\beta_k$  over words from a  $\text{Dirichlet}(\gamma)$ .

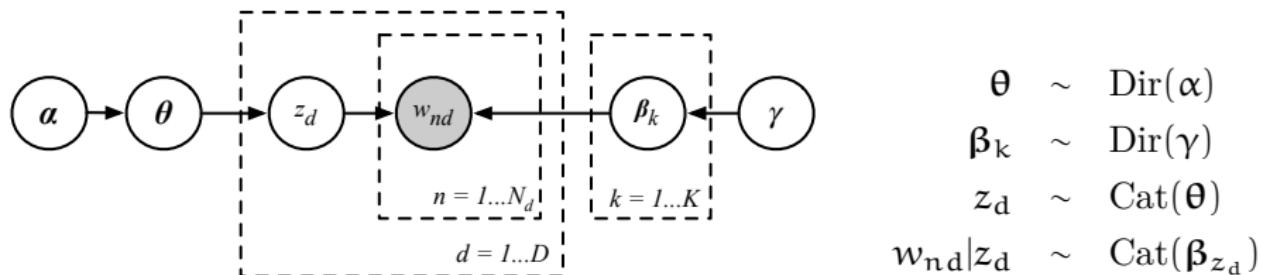
# Limitations of the mixture of categoricals model



A generative view of the mixture of categoricals model

- 1 Draw a distribution  $\theta$  over  $K$  topics from a  $\text{Dirichlet}(\alpha)$ .
- 2 For each topic  $k$ , draw a distribution  $\beta_k$  over words from a  $\text{Dirichlet}(\gamma)$ .
- 3 For each document  $d$ , draw a topic  $z_d$  from a  $\text{Categorical}(\theta)$

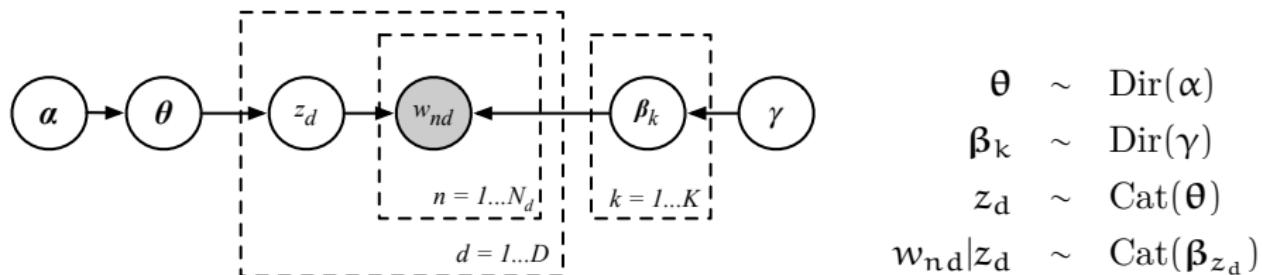
# Limitations of the mixture of categoricals model



A generative view of the mixture of categoricals model

- 1 Draw a distribution  $\theta$  over  $K$  topics from a  $\text{Dirichlet}(\alpha)$ .
- 2 For each topic  $k$ , draw a distribution  $\beta_k$  over words from a  $\text{Dirichlet}(\gamma)$ .
- 3 For each document  $d$ , draw a topic  $z_d$  from a  $\text{Categorical}(\theta)$
- 4 For each document  $d$ , draw  $N_d$  words  $w_{nd}$  from a  $\text{Categorical}(\beta_{z_d})$

# Limitations of the mixture of categoricals model



A generative view of the mixture of categoricals model

- 1 Draw a distribution  $\theta$  over  $K$  topics from a  $\text{Dirichlet}(\alpha)$ .
- 2 For each topic  $k$ , draw a distribution  $\beta_k$  over words from a  $\text{Dirichlet}(\gamma)$ .
- 3 For each document  $d$ , draw a topic  $z_d$  from a  $\text{Categorical}(\theta)$
- 4 For each document  $d$ , draw  $N_d$  words  $w_{nd}$  from a  $\text{Categorical}(\beta_{z_d})$

Limitations:

- All words in each document are drawn from one specific topic distribution.
- This works if each document is exclusively about one topic, but if some documents span more than one topic, then “blurred” topics must be learnt.

# NIPS dataset: LDA topics 1 to 7 out of 20.

network	network	model	problem	neuron	network	cell
unit	node	data	constraint	cell	neural	model
training	representation	distribution	distance	input	system	visual
weight	input	probability	cluster	model	model	direction
input	unit	parameter	point	synaptic	control	motion
hidden	learning	set	algorithm	firing	output	field
output	activation	gaussian	tangent	response	recurrent	eye
learning	nodes	error	energy	activity	input	unit
layer	pattern	method	clustering	potential	signal	cortex
error	level	likelihood	optimization	current	controller	orientation
set	string	prediction	cost	synapses	forward	map
neural	structure	function	graph	membrane	error	receptive
net	grammar	mean	method	pattern	dynamic	neuron
number	symbol	density	neural	output	problem	input
performance	recurrent	prior	transformation	inhibitory	training	head
pattern	system	estimate	matching	effect	nonlinear	spatial
problem	connectionist	estimation	code	system	prediction	velocity
trained	sequence	order	expert	entropy	set	activity
result	context	bayesian	set	network	algorithm	cortical

# NIPS dataset: LDA topics 8 to 14 out of 20.

circuit	learning	speech	classifier	network	data	function
chip	algorithm	word	classification	neuron	memory	linear
network	error	recognition	pattern	dynamic	performance	vector
neural	gradient	system	training	system	genetic	input
analog	weight	training	character	neural	system	space
output	function	network	set	pattern	set	matrix
neuron	convergence	hmm	vector	phase	features	component
current	vector	speaker	class	point	model	dimensional
input	rate	context	algorithm	equation	problem	point
system	parameter	model	recognition	model	task	data
vlsi	optimal	set	data	function	patient	basis
weight	problem	mlp	performance	field	human	output
implementation	method	neural	error	attractor	target	set
voltage	order	acoustic	number	connection	similarity	approximation
processor	descent	phoneme	digit	parameter	algorithm	order
bit	equation	output	feature	oscillation	number	method
hardware	term	input	network	fixed	population	gaussian
data	result	letter	neural	oscillator	probability	network
digital	noise	performance	nearest	states	item	algorithm
transistor	solution	segment	problem	activity	result	dimension

# NIPS dataset: LDA topics 15 to 20 out of 20.

function	learning	model	image	rules	signal
network	action	object	images	algorithm	frequency
bound	task	movement	system	learning	noise
neural	function	motor	features	tree	spike
threshold	reinforcement	point	feature	rule	information
theorem	algorithm	view	recognition	examples	filter
result	control	position	pixel	set	channel
number	system	field	network	neural	auditory
size	path	arm	object	prediction	temporal
weight	robot	trajectory	visual	concept	model
probability	policy	learning	map	knowledge	sound
set	problem	control	neural	trees	rate
proof	step	dynamic	vision	information	train
net	environment	hand	layer	query	system
input	optimal	joint	level	label	processing
class	goal	surface	information	structure	analysis
dimension	method	subject	set	model	peak
case	states	data	segmentation	method	response
complexity	space	human	task	data	correlation
distribution	sutton	inverse	location	system	neuron

# Latent Dirichlet Allocation (LDA)

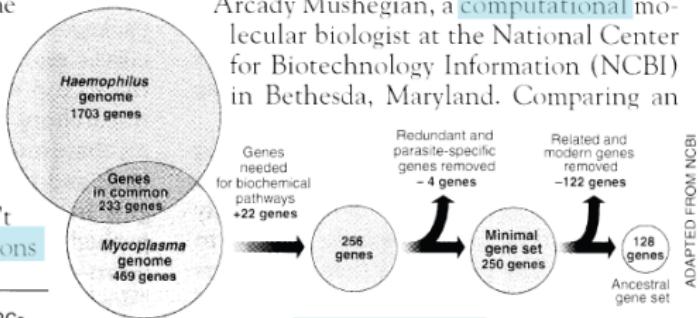
## Seeking Life's Bare (Genetic) Necessities

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

Although the numbers don't match precisely, those predictions

“are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



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

ADAPTED FROM NCBI

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

# Latent Dirichlet Allocation (LDA)

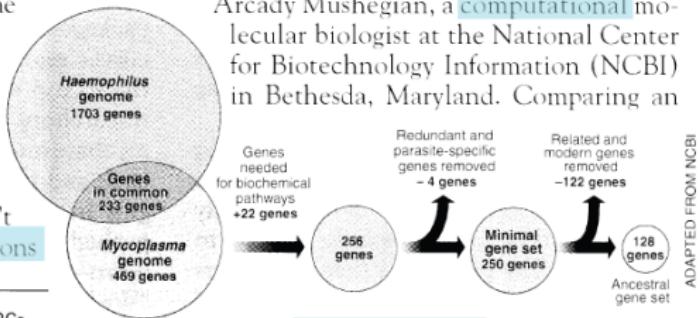
## Seeking Life's Bare (Genetic) Necessities

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

Although the numbers don't match precisely, those predictions

“are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



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

ADAPTED FROM NCBI

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

# Generative model for LDA

## Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>6</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that just today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

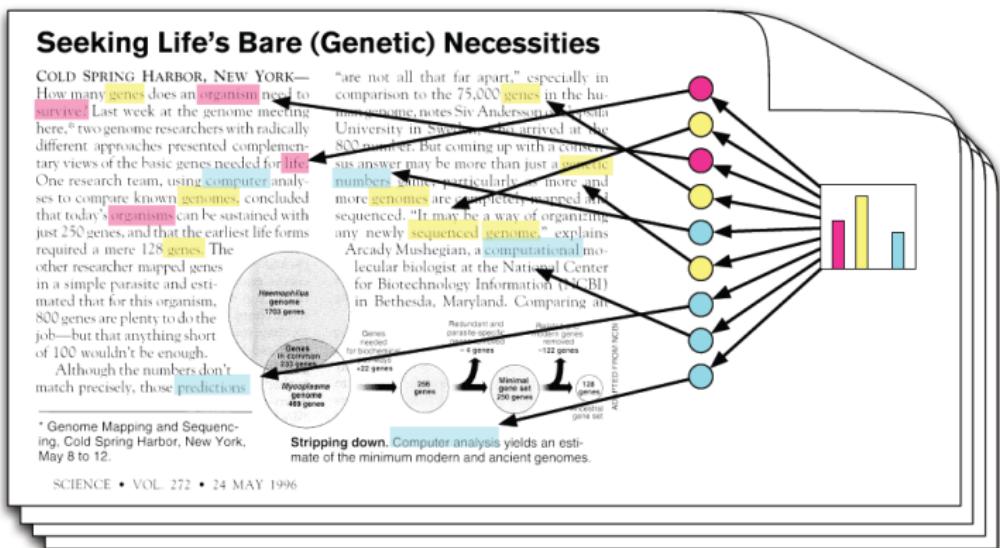
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Umeå University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a numbers game, particularly as more and more genomes are completely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

the genomes to each other, he says, can reveal which genes are shared and which are unique to a particular species.

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

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments



- Each *topic* is a distribution over words.

# Generative model for LDA

## Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>6</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

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

Diagram illustrating the process of genome comparison and gene set reduction:

- Neomphalus genome: 1703 genes
- Mycoplasma genome: 485 genes
- Genes needed for life: 22 genes
- 22 genes → 256 genes
- 256 genes → Maximum gene set: 250 genes
- 250 genes → 128 genes
- 128 genes → Ancient life set

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

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

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments



- Each *topic* is a distribution over words.
- Each *document* is a mixture of corpus-wide topics.

# Generative model for LDA

## Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>6</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that just 125 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

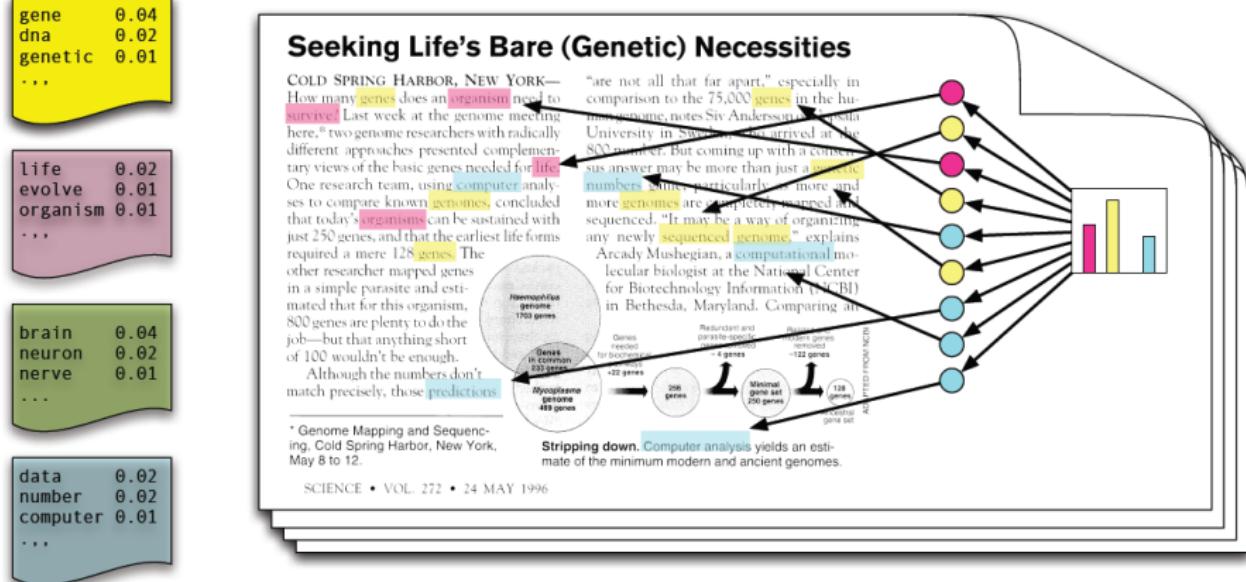
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Umeå University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are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

Diagram illustrating the process of stripping down genomes to find the minimum set of genes required for life. It starts with the Neamphius genome (1703 genes) and the Mycoplasma genome (485 genes), both with 932 genes in common. By removing redundant and parasite-specific genes, 22 genes are removed, leaving 4 genes. Then, by removing other genes, 122 genes are removed, leaving 250 genes. Finally, by removing additional genes, 238 genes are removed, leaving a minimal gene set of 250 genes.

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

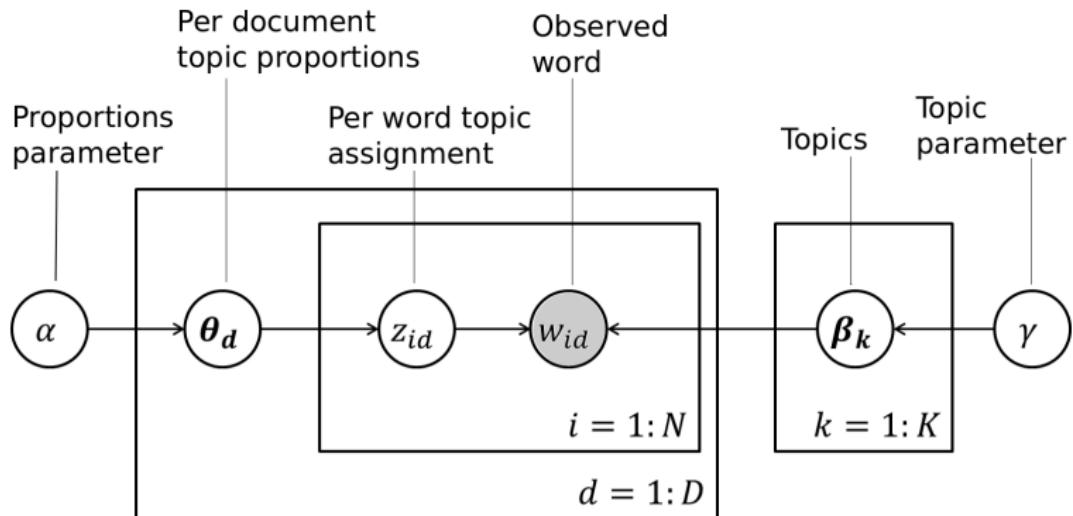
SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments



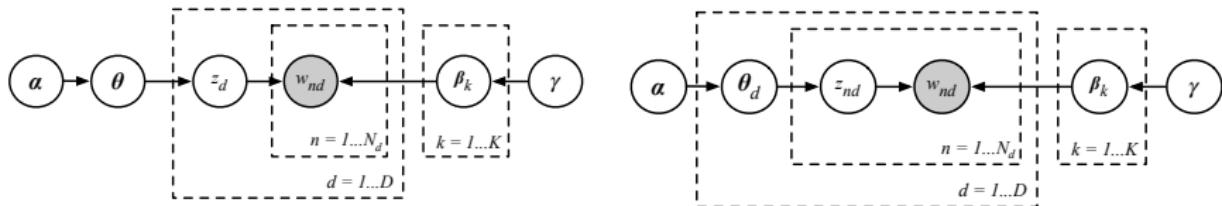
- Each *topic* is a distribution over words.
- Each *document* is a mixture of corpus-wide topics.
- Each *word* is drawn from one of those topics.

# The LDA graphical model



- Nodes are random variables; edges indicate dependence.
- Shaded nodes indicate *observed* variables.

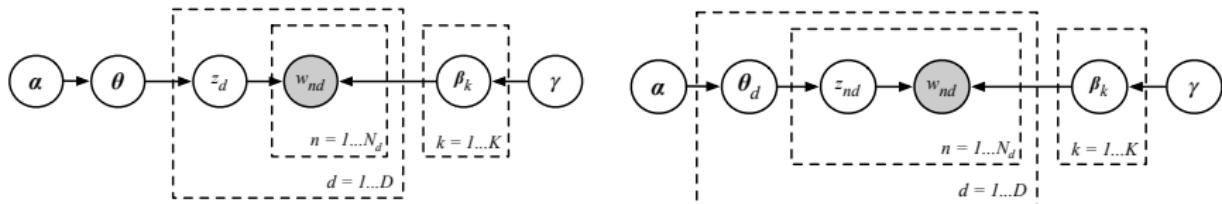
# The difference between mixture of categoricals and LDA



A generative view of LDA

- 1 For each document  $d$  draw a distribution  $\theta_d$  over topics from a  $\text{Dirichlet}(\alpha)$ .

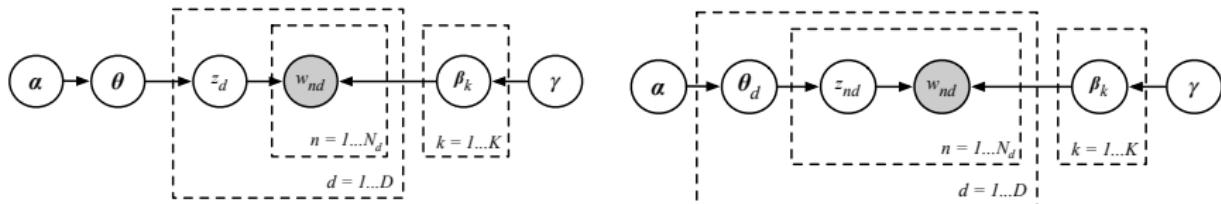
# The difference between mixture of categoricals and LDA



## A generative view of LDA

- 1 For each document  $d$  draw a distribution  $\theta_d$  over topics from a  $\text{Dirichlet}(\alpha)$ .
- 2 For each topic  $k$  draw a distribution  $\beta_k$  over words from a  $\text{Dirichlet}(\gamma)$ .

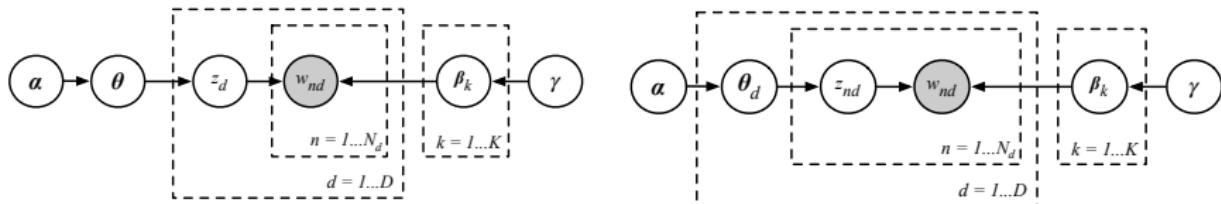
# The difference between mixture of categoricals and LDA



## A generative view of LDA

- 1 For each document  $d$  draw a distribution  $\theta_d$  over topics from a  $\text{Dirichlet}(\alpha)$ .
- 2 For each topic  $k$  draw a distribution  $\beta_k$  over words from a  $\text{Dirichlet}(\gamma)$ .
- 3 Draw a topic  $z_{n,d}$  for the  $n$ -th word in document  $d$  from a  $\text{Categorical}(\theta_d)$

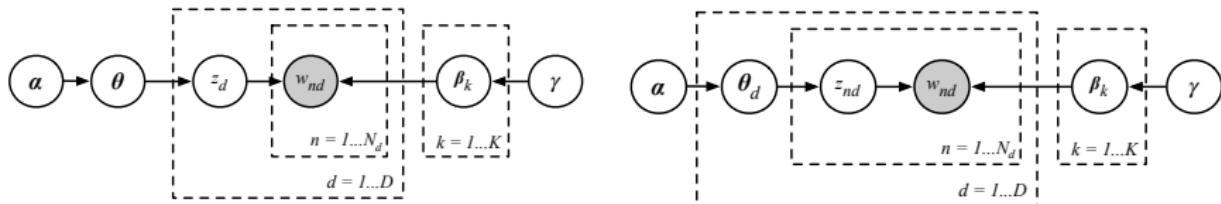
# The difference between mixture of categoricals and LDA



## A generative view of LDA

- 1 For each document  $d$  draw a distribution  $\theta_d$  over topics from a  $\text{Dirichlet}(\alpha)$ .
- 2 For each topic  $k$  draw a distribution  $\beta_k$  over words from a  $\text{Dirichlet}(\gamma)$ .
- 3 Draw a topic  $z_{nd}$  for the  $n$ -th word in document  $d$  from a  $\text{Categorical}(\theta_d)$
- 4 Draw word  $w_{nd}$  from a  $\text{Categorical}(\beta_{z_{nd}})$

## The difference between mixture of categoricals and LDA



## A generative view of LDA

- ① For each document  $d$  draw a distribution  $\theta_d$  over topics from a  $\text{Dirichlet}(\alpha)$ .
- ② For each topic  $k$  draw a distribution  $\beta_k$  over words from a  $\text{Dirichlet}(\gamma)$ .
- ③ Draw a topic  $z_{nd}$  for the  $n$ -th word in document  $d$  from a  $\text{Categorical}(\theta_d)$
- ④ Draw word  $w_{nd}$  from a  $\text{Categorical}(\beta_{z_{nd}})$

Differences with the mixture of categoricals model:

- In LDA, every word in a document can be drawn from a different topic,
- and every document has its own distribution over topics  $\theta_d$ .

# The posterior distribution

Topics



- In reality, we only observe the documents.

Documents

## Seeking Life's Bare (Genetic) Necessities

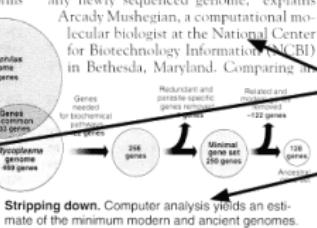
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>10</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

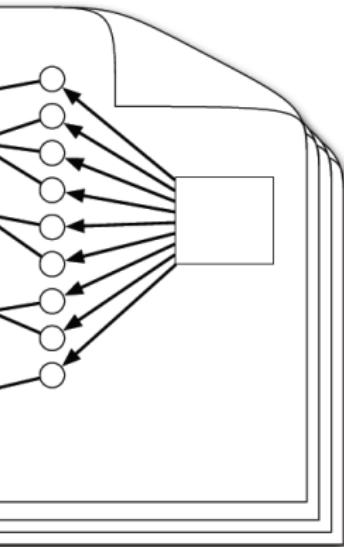
<sup>10</sup> Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden. We arrived at the SGD number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completed, mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

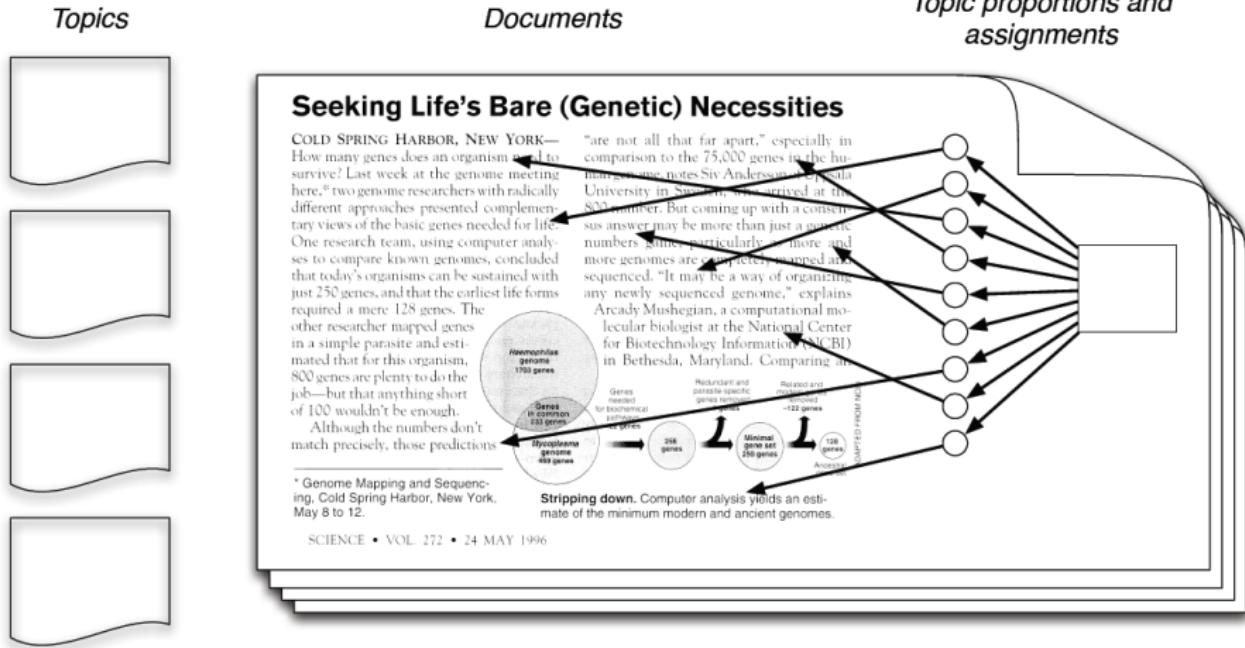
Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all



Topic proportions and assignments



# The posterior distribution



- In reality, we only observe the documents.
- The other structure are *hidden* variables.

from David Blei

# The posterior distribution

Topics

Documents

Topic proportions and assignments

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,<sup>6</sup> two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions

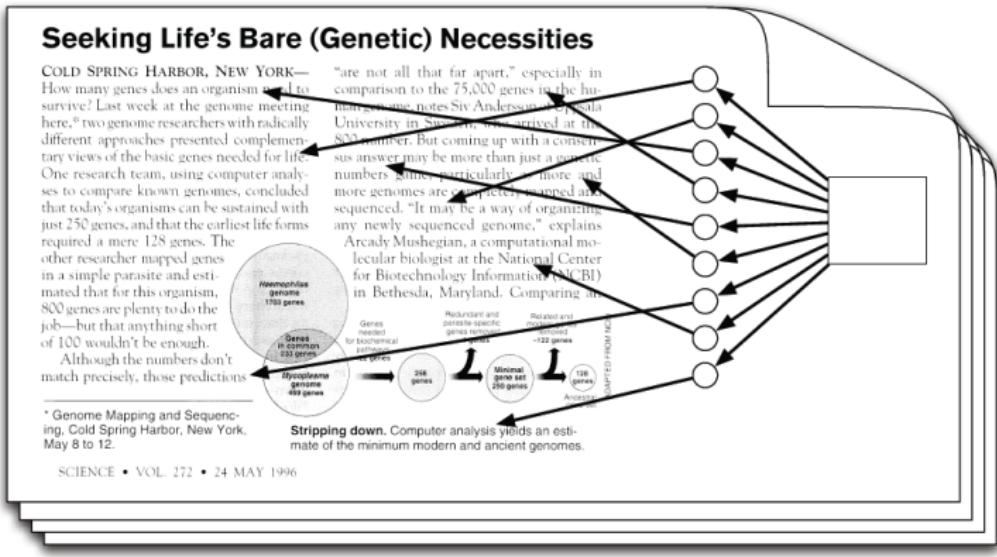
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden. We arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completed, mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains

Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

<sup>6</sup> Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

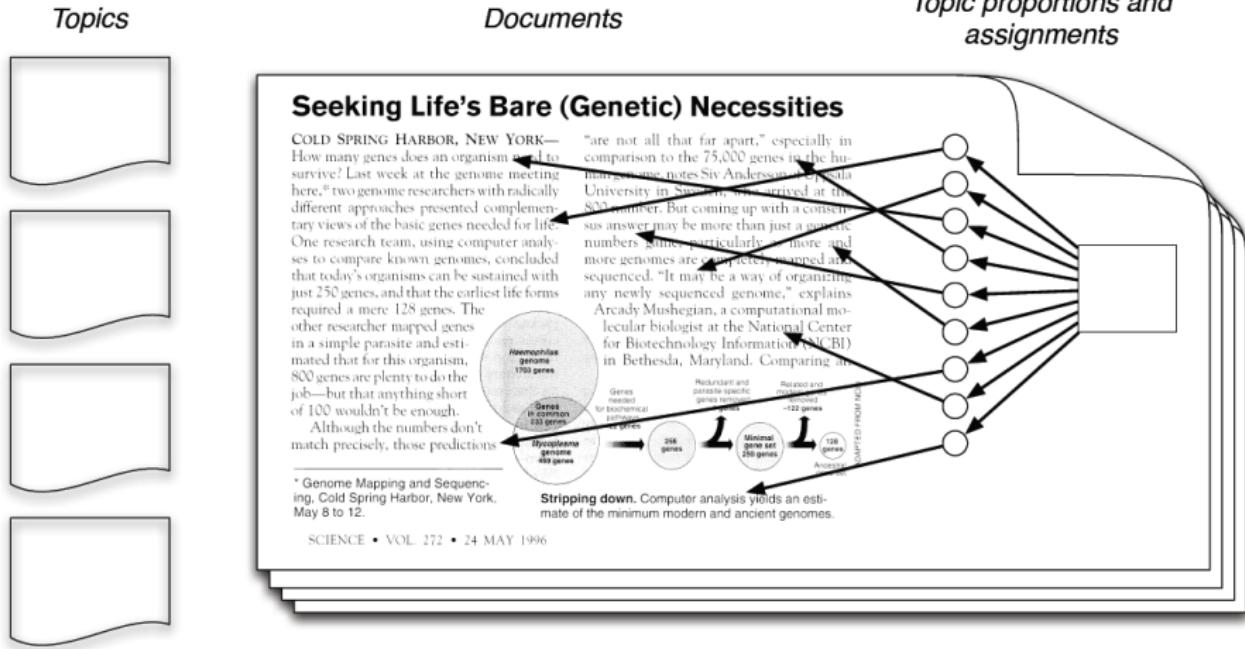
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996



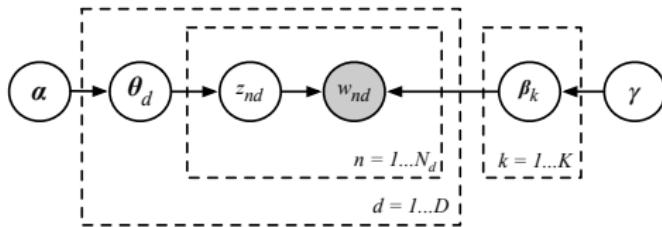
- Our goal is to *infer* the hidden variables.

# The posterior distribution



- Our goal is to *infer* the hidden variables.
- This means computing their distribution conditioned on the documents  $p(\text{topics, proportions, assignments} | \text{documents})$

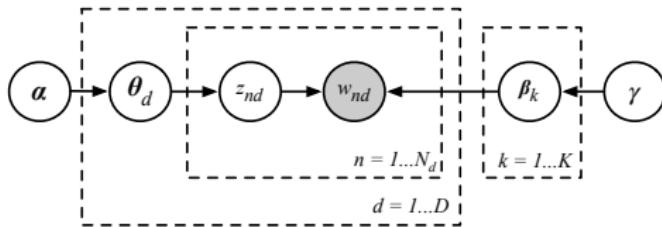
# The LDA inference problem



$$p(\beta_{1:K}, \theta_{1:D}, \{z_{nd}\}, \{w_{nd}\} | \gamma, \alpha)$$

$$= \prod_{k=1}^K p(\beta_k | \gamma) \prod_{d=1}^D \left[ p(\theta_d | \alpha) \prod_{n=1}^{N_d} [p(z_{nd} | \theta_d) p(w_{nd} | \beta_{1:K}, z_{nd})] \right]$$

# The LDA inference problem

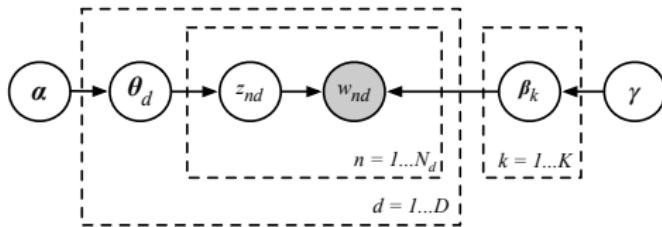


$$p(\beta_{1:K}, \theta_{1:D}, \{z_{nd}\}, \{w_{nd}\} | \gamma, \alpha)$$

$$= \prod_{k=1}^K p(\beta_k | \gamma) \prod_{d=1}^D \left[ p(\theta_d | \alpha) \prod_{n=1}^{N_d} [p(z_{nd} | \theta_d) p(w_{nd} | \beta_{1:K}, z_{nd})] \right]$$

Learning involves computing the posterior over the parameters,  $\beta_{1:K}$  and  $\theta_{1:D}$  given the words  $\{w_{nd}\}$ , but this requires we marginalize out the latent  $\{z_{nd}\}$ .

# The LDA inference problem



$$p(\beta_{1:K}, \theta_{1:D}, \{z_{nd}\}, \{w_{nd}\} | \gamma, \alpha)$$

$$= \prod_{k=1}^K p(\beta_k | \gamma) \prod_{d=1}^D \left[ p(\theta_d | \alpha) \prod_{n=1}^{N_d} [p(z_{nd} | \theta_d) p(w_{nd} | \beta_{1:K}, z_{nd})] \right]$$

Learning involves computing the posterior over the parameters,  $\beta_{1:K}$  and  $\theta_{1:D}$  given the words  $\{w_{nd}\}$ , but this requires we marginalize out the latent  $\{z_{nd}\}$ .

How many configurations are there?

This computation is *intractable*.

# The intractability of LDA

The evidence (normalising constant of the posterior):

$$p(\{w_{nd}\}) = \int \int \prod_{d=1}^D \prod_{k=1}^K \prod_{n=1}^{N_d} \sum_{z_{nd}} p(z_{nd}|\theta_d) p(\theta_d|\alpha) p(w_{nd}|\beta_{1:k}, z_{nd}) p(\beta_k|\gamma) d\beta_k d\theta_d$$

# The intractability of LDA

The evidence (normalising constant of the posterior):

$$p(\{w_{nd}\}) = \int \int \prod_{d=1}^D \prod_{k=1}^K \prod_{n=1}^{N_d} \sum_{z_{nd}} p(z_{nd}|\theta_d) p(\theta_d|\alpha) p(w_{nd}|\beta_{1:k}, z_{nd}) p(\beta_k|\gamma) d\beta_k d\theta_d$$

We need to average over all possible set of values of all  $z_{nd}$ . If every document had  $N$  words, this means  $K^N$  configurations per document.

# Gibbs to the Rescue

The posterior is *intractable* because there are **too many** possible latent  $\{z_{nd}\}$ .

# Gibbs to the Rescue

The posterior is *intractable* because there are **too many** possible latent  $\{z_{nd}\}$ .  
Sigh, ... if only we knew the  $\{z_{nd}\}$  ...?

# Gibbs to the Rescue

The posterior is *intractable* because there are **too many** possible latent  $\{z_{nd}\}$ .

Sigh, ... if only we knew the  $\{z_{nd}\}$  ...?

Which might remind us of Gibbs sampling ... could we sample each latent variable given the values of the other ones?

# Refresher on Beta and Dirichlet

If we had a  $p(\pi) = \text{Beta}(\alpha, \beta)$  prior on a **binomial** probability  $\pi$ , and observed  $k$  successes and  $n - k$  failures, then the posterior probability

$$p(\pi|n, k) = \text{Beta}(\alpha + k, \beta + n - k),$$

and the predictive probability of success at the next experiment

$$p(\text{success}|n, k) = \int p(\text{success}|\pi)p(\pi|n, k)d\pi = \mathbb{E}[\pi|n, k] = \frac{\alpha + k}{\alpha + \beta + n}.$$

---

# Refresher on Beta and Dirichlet

If we had a  $p(\pi) = \text{Beta}(\alpha, \beta)$  prior on a **binomial** probability  $\pi$ , and observed  $k$  successes and  $n - k$  failures, then the posterior probability

$$p(\pi|n, k) = \text{Beta}(\alpha + k, \beta + n - k),$$

and the predictive probability of success at the next experiment

$$p(\text{success}|n, k) = \int p(\text{success}|\pi)p(\pi|n, k)d\pi = \mathbb{E}[\pi|n, k] = \frac{\alpha + k}{\alpha + \beta + n}.$$

---

Analogously, if we had a prior  $p(\pi) = \text{Dir}(\alpha_1, \dots, \alpha_k)$  on the parameter  $\pi$  of a **multinomial**, and  $c_1, \dots, c_k$  observed counts of each value, the posterior is

$$p(\pi|c_1, \dots, c_k) = \text{Dir}(\alpha_1 + c_1, \dots, \alpha_k + c_k),$$

and the predictive probability that the next item takes value  $j$  is:

$$p(j|c_1, \dots, c_k) = \int p(j|\pi)p(\pi|c_1, \dots, c_k)d\pi = \mathbb{E}[\pi|c_1, \dots, c_k] = \frac{\alpha_j + c_j}{\sum_i \alpha_i + c_i}.$$

# Collapsed Gibbs sampler for LDA

In the LDA model, we can integrate out the parameters of the multinomial distributions,  $\theta_d$  and  $\beta$ , and just keep the latent counts  $z_{nd}$ . Sampling these  $z_{nd}$  in turn is called a *collapsed* Gibbs sampler.

# Collapsed Gibbs sampler for LDA

In the LDA model, we can integrate out the parameters of the multinomial distributions,  $\theta_d$  and  $\beta$ , and just keep the latent counts  $z_{nd}$ . Sampling these  $z_{nd}$  in turn is called a *collapsed* Gibbs sampler.

Recall, that the predictive distribution for a symmetric Dirichlet is given by

$$p_i = \frac{\alpha + c_i}{\sum_j \alpha + c_j}.$$

# Collapsed Gibbs sampler for LDA

In the LDA model, we can integrate out the parameters of the multinomial distributions,  $\theta_d$  and  $\beta$ , and just keep the latent counts  $z_{nd}$ . Sampling these  $z_{nd}$  in turn is called a *collapsed* Gibbs sampler.

Recall, that the predictive distribution for a symmetric Dirichlet is given by

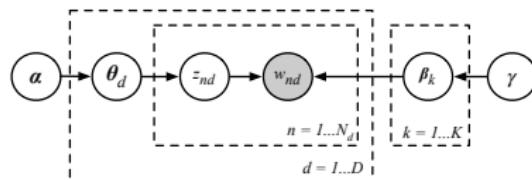
$$p_i = \frac{\alpha + c_i}{\sum_j \alpha + c_j}.$$

Now, for Gibbs sampling, we need the predictive distribution for a single  $z_{nd}$  given all other  $z_{nd}$ , ie, given all the counts except for the word  $n$  in document  $d$ . The Gibbs update contains two parts, one from the topic distribution and one from the word distribution:

$$p(z_{nd} = k | \{z_{-nd}\}, \{w\}, \gamma, \alpha) \propto \frac{\alpha + c_{-nd}^k}{\sum_{j=1}^K (\alpha + c_{-nd}^j)} \frac{\gamma + \tilde{c}_{-w_{nd}}^k}{\sum_{m=1}^M (\gamma + \tilde{c}_{-m}^k)}$$

where  $c_{-nd}^k$  is the count of words from document  $d$ , excluding  $n$ , assigned to topic  $k$ , and  $\tilde{c}_{-m}^k$  is the number of times word  $m$  was generated from topic  $k$  (again, excluding the observation  $nd$ ).

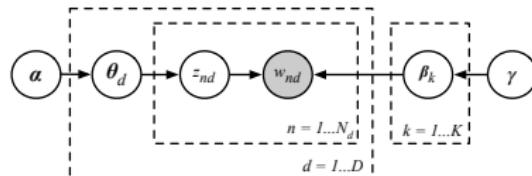
# Derivation of the collapsed Gibbs sampler



The probability of everything:

$$p(\beta_{1:K}, \theta_{1:D}, \{z_{nd}\}, \{w_{nd}\} | \gamma, \alpha) = \prod_{k=1}^K p(\beta_k | \gamma) \prod_{d=1}^D \left[ p(\theta_d | \alpha) \prod_{n=1}^{N_d} [p(z_{nd} | \theta_d) p(w_{nd} | \beta_{1:K}, z_{nd})] \right]$$

# Derivation of the collapsed Gibbs sampler



The probability of everything:

$$p(\beta_{1:K}, \theta_{1:D}, \{z_{nd}\}, \{w_{nd}\} | \gamma, \alpha) = \prod_{k=1}^K p(\beta_k | \gamma) \prod_{d=1}^D \left[ p(\theta_d | \alpha) \prod_{n=1}^{N_d} [p(z_{nd} | \theta_d) p(w_{nd} | \beta_{1:K}, z_{nd})] \right]$$

What we want for Gibbs sampling is:

$$\begin{aligned} p(z_{nd} = k | \{z_{-nd}\}, \{w\}, \gamma, \alpha) &\propto p(z_{nd} = k | \{z_{-nd}\}, \alpha) p(w_{nd} | z_{nd} = k, \{w_{-nd}\}, \{z_{-nd}\}, \gamma) \\ &= \frac{\alpha + c_{-nd}^k}{\sum_{j=1}^K (\alpha + c_{-nd}^j)} \frac{\gamma + \tilde{c}_{-w_{nd}}^k}{\sum_{m=1}^M (\gamma + \tilde{c}_{-m}^k)} \end{aligned}$$

$$\text{where } c_{-nd}^j \stackrel{\text{def}}{=} \sum_{n' \neq n} I(z_{n'd} = j) \text{ and } \tilde{c}_{-m}^k \stackrel{\text{def}}{=} \sum_{(n', d') \neq (n, d)} I(w_{n'd'} = m) I(z_{n'd'} = k).$$