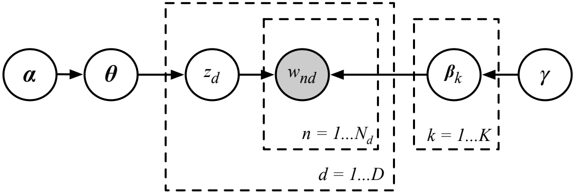
Latent Dirichlet Allocation for Topic Modeling

Carl Edward Rasmussen

November 18th, 2016

Limitations of the mixture of categoricals model



$$\begin{aligned} \theta &\sim \mathrm{Dir}(\alpha) \\ \beta_k &\sim \mathrm{Dir}(\gamma) \\ z_d &\sim \mathrm{Cat}(\theta) \\ w_{nd}|z_d &\sim \mathrm{Cat}(\beta_{z_d}) \end{aligned}$$

A generative view of the mixture of categoricals model

1. Draw a distribution θ over K topics from a Dirichlet(α).
2. For each topic k, draw a distribution β_k over words from a Dirichlet(γ).
3. For each document d, draw a topic z_d from a Categorical(θ)
4. For each document d, draw N_d words w_{nd} from a Categorical(β_{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	neural	objective	neural	adaptive	stimulus
generalization	order	expert	entropy	function	memory	activity
result	context	bayesian	set	network	algorithm	cortical

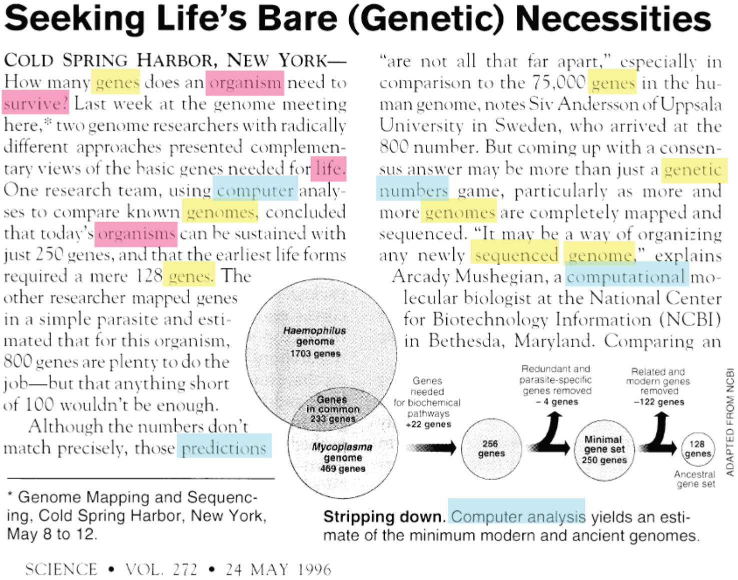
NIPS dataset: LDA topics 8 to 14 out of 20.

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

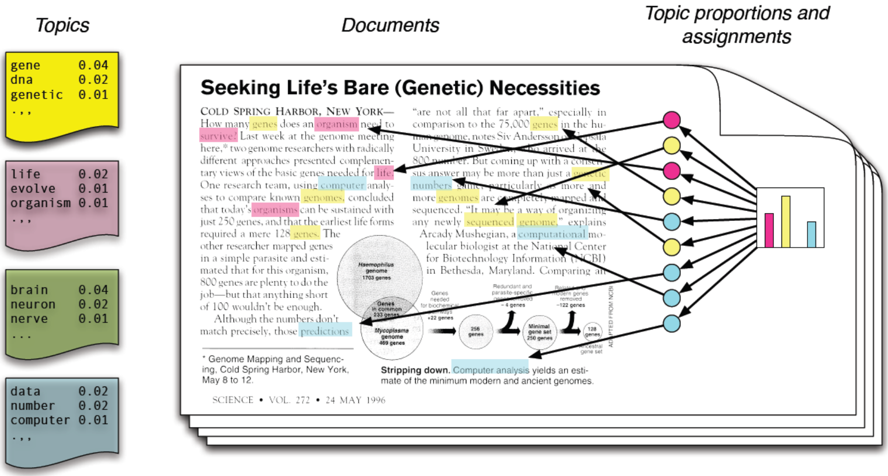
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

Generative model for LDA



```
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
...
```

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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed − 4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

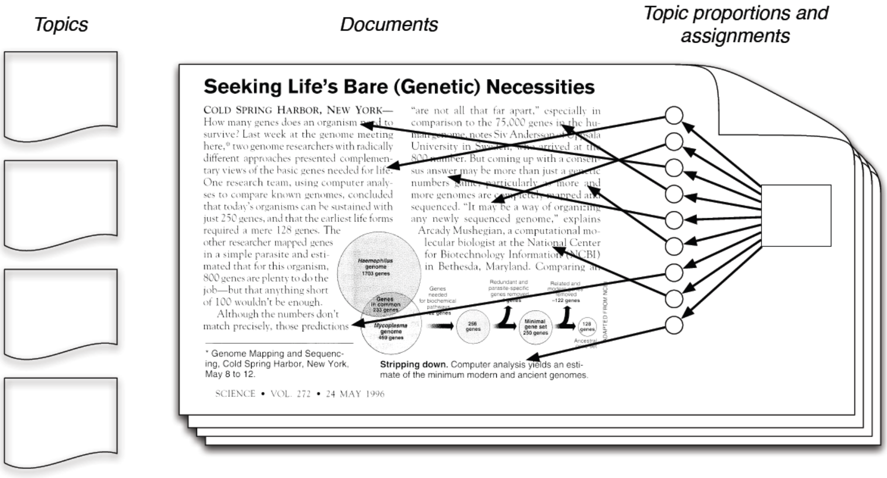
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- 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.

from David Blei

The posterior distribution



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways 22 genes
256 genes
Redundant and parasite-specific genes removed
genes
Minimal gene set 250 genes
Related and modern genes removed ~122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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

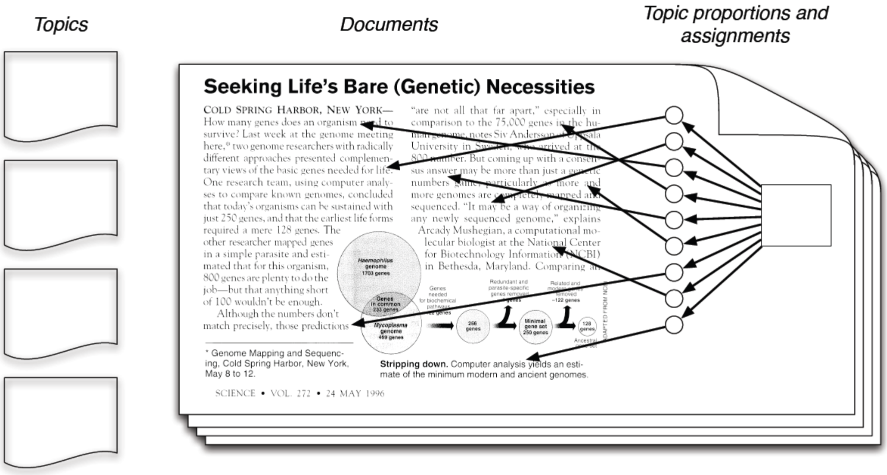
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- 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
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed -4 genes
Minimal gene set 250 genes
Related and modern genes removed -122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

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.

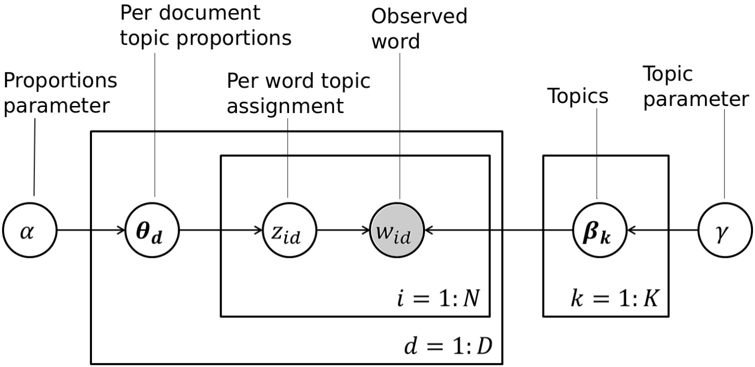
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Our goal is to *infer* the hidden variables.
- This means computing their distribution conditioned on the documents

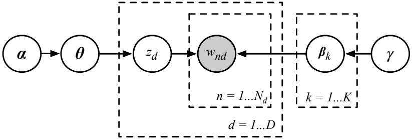
$$p(\text{topics, proportions, assignments}|\text{documents})$$

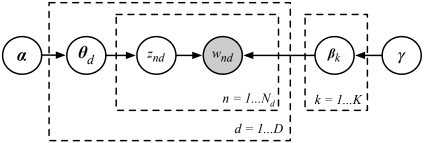
The LDA graphical model


Per document topic proportions
Observed word
Proportions parameter
Per word topic assignment
Topics
Topic parameter
α
$\boldsymbol{\theta}_{\boldsymbol{d}}$
z_{id}
w_{id}
$\boldsymbol{\beta}_{\boldsymbol{k}}$
γ
$i = 1{:}N$
$k = 1{:}K$
$d = 1{:}D$

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

The difference between mixture of categoricals and LDA





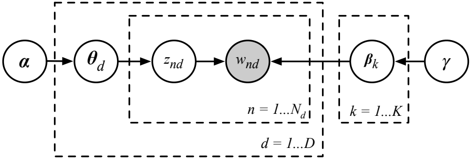
A generative view of LDA

1. For each document d draw a distribution θ_d over topics from a Dirichlet(α).
2. For each topic k draw a distribution β_k over words from a Dirichlet(γ).
3. Draw a topic z_{nd} for the n-th word in document d from a Categorical(θ_d)
4. Draw word w_{nd} from a 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 θ_d.

The LDA inference problem



"Always write down the probability of everything." (Steve Gull)

$$p(\boldsymbol{\beta}_{1:K}, \theta_{1:D}, \{z_{nd}\}, \{w_{nd}\} | \gamma, \alpha) \\ = \prod_{k=1}^{K} p(\boldsymbol{\beta}_k|\gamma) \prod_{d=1}^{D} \Big[p(\theta_d|\alpha) \prod_{n=1}^{N_d} \big[p(z_{nd}|\theta_d) p(w_{nd}|\boldsymbol{\beta}_{1:K}, z_{nd}) \big] \Big]$$

Learning involves computing the posterior over the parameters, $\boldsymbol{\beta}_{1:K}$ and $\theta_{1:D}$ given the words $\{w_{nd}\}$, but this requires the 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):

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

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 $\{z_{nd}\}$.

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 π, 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 $p(\boldsymbol{\pi}) = \text{Dir}(\alpha_1, \ldots, \alpha_k)$ on the parameter $\boldsymbol{\pi}$ of a multinomial, and $c_1, \ldots, c_k$ observed counts of each value, the posterior is

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

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

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

Collapsed Gibbs sampler for LDA

In the LDA model, we can integrate out the parameters of the multinomial distributions, θ_d and β, 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

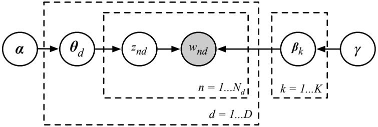
$$p_i = \frac{\alpha + c_i}{\sum_j \alpha + c_j}.$$

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:

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

where c^k_{-nd} is the count of words from document d, excluding n, assigned to topic k, and $\tilde{c}^k_{-m}$ 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:

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

What we want for Gibbs sampling is:

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

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

Per word Perplexity

In text modeling, performance is often given in terms of per word *perplexity*. The perplexity for a document is given by

$$\exp(-\ell/n),$$

where ℓ is the log joint probability over the words in the document, and n is the number of words. Note, that the average is done in the log space.
A perplexity of g corresponds to the uncertainty associated with a die with g sides, which generates each new word. Example:

$$p(w_1, w_2, w_3, w_4) = \frac{1}{6}\frac{1}{6}\frac{1}{6}\frac{1}{6} \tag{1}$$

$$\frac{1}{n}\log p(w_1, \ldots, w_4) = \frac{1}{4}\log(\frac{1}{6})^4 = -\log 6 \tag{2}$$

$$\text{perplexity} = \exp(-\frac{1}{n}\log p(w_1, \ldots, w_4)) = 6 \tag{3}$$