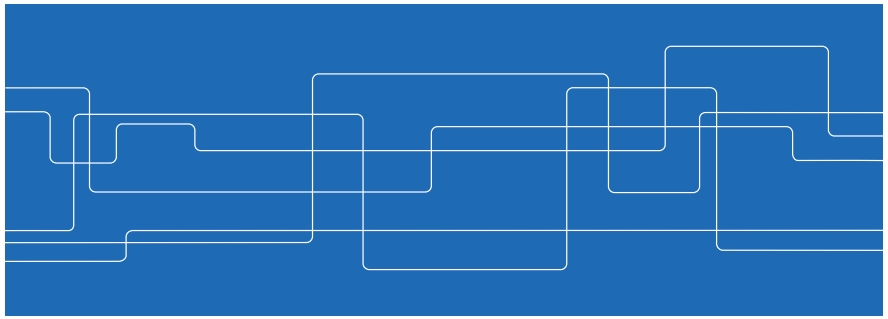
DD2434 Machine Learning, Advanced Course

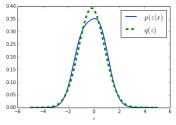
Lecture 12: Sampling

Hedvig Kjellström
hedvig@kth.se
https://www.kth.se/social/course/DD2434/



Recap from Lecture 6: Inference – in general, approximation is needed

In Lecture 6, deterministic approximation methods

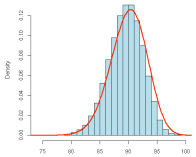


Analytic approximations to the exact posterior p(latent | obs), i.e. fitting some known parametric function or assume some independencies

\+ : Fast since analytic/closed-form solution

– : always an approximation to the true posterior

Here, stochastic approximation methods



Monte Carlo sampling from the exact posterior p(latent | obs)

\+ : Given ∞ samples, converges to exact solution

– : slow in many cases, sometimes hard to know if sampling independent samples from true posterior

Today

Monte Carlo (MC) sampling (Bishop 11.1)
- Standard Monte Carlo sampling
- Rejection sampling
- Importance sampling

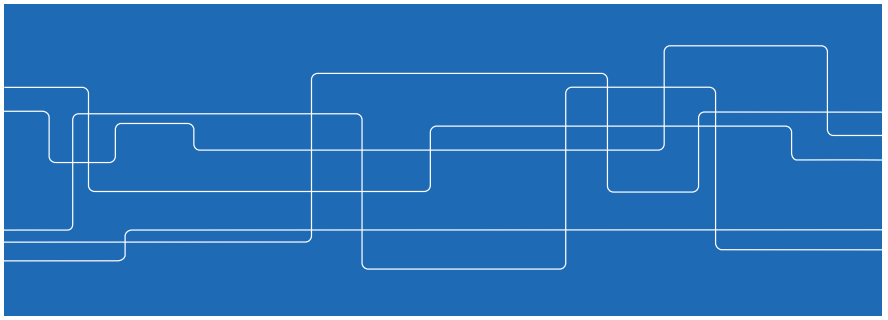
Markov chain Monte Carlo (MCMC) sampling (Bishop 11.2)

Gibbs sampling (Bishop 11.3)

Some intuitions about Gibbs sampling in LDA (Griffiths)

Monte Carlo (MC) Sampling

Bishop Section 11.1



The Monte Carlo Principle

Start off with **discrete** state space z

Imagine that we can **sample** $z^{(l)}$ from the pdf $p(z)$ but that we do not know its functional form

Might want to estimate for example:

$$E[z] = \sum z\, p(z)$$

$p(z)$ can be approximated by a histogram over $z^{(l)}$:

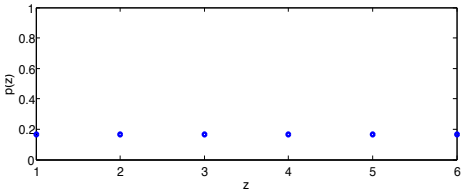
$$\hat{q}(z) = \frac{1}{L}\sum_{l=1}^{L} \delta_{z^{(l)}=z}$$

Example: Dice Roll

The probability of outcomes of dice rolls: $p(z) = \frac{1}{6}$

Exact solution:

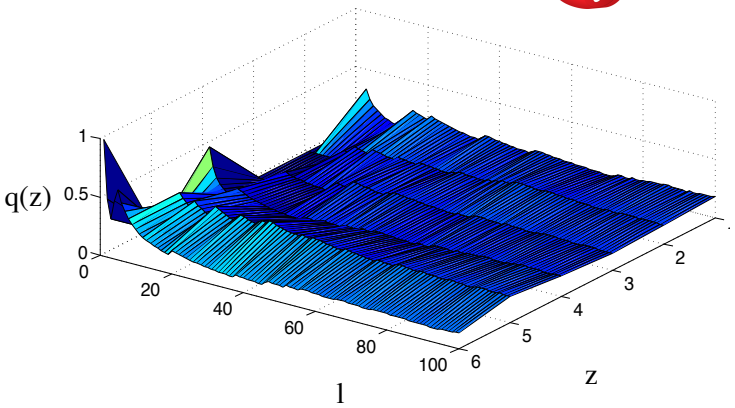
What would happen if the dice was bad?



Monte Carlo approximation:
Roll a dice a number of times, might get

$$z^{(1)} = 6 \quad z^{(2)} = 4 \quad z^{(3)} = 1 \quad z^{(4)} = 6 \quad z^{(5)} = 6$$

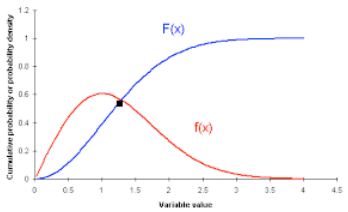
Example: Dice Roll



Monte Carlo Sampling – Inverse Probability Transform

Cumulative distribution function F of distribution f (that we want to sample from)

A uniformly distributed random variable $U \sim U(0,1)$ will render $F^{-1}(U) \sim F$



$f(z)$ does not have to be an analytic function, can also be a histogram like $\hat{q}(z)$!

Importance Sampling

We very often (in Bayesian methods for example) want to approximate integrals of the form

$$E[f] = \int f(x)p(x)dx$$

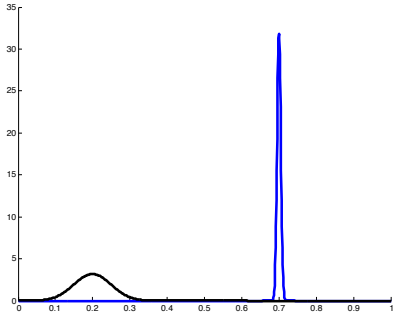
Monte Carlo sampling approach is to draw samples x^s from $p(x)$ and approximating the integral with a sum

$$E[f] = \int f(x)p(x)dx = \frac{1}{S}\sum_{s=1}^{S} f(x^s)$$

Importance Sampling

Discuss with your neighbor (5 min):

But what if $p(x)$ and $f(x)$ look like this, what happens with the estimation?



Importance Sampling

In these cases, a good idea is to introduce **proposal** $q(x)$ to sample from:

$$E[f] = \int f(x)\frac{p(x)}{q(x)}q(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} w_s f(x^s)$$

where $w_s \equiv \frac{p(x^s)}{q(x^s)}$

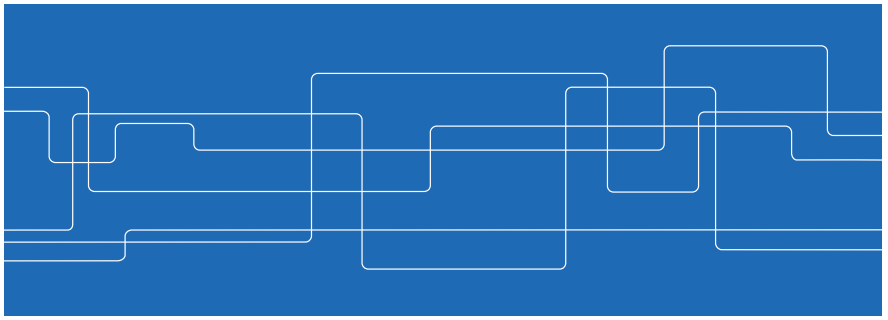
Reasons:

- $q(x)$ is smoother / less spiky than $p(x)$
- $q(x)$ is of a nicer analytical form than $p(x)$

In general, good to keep $q(x) \propto p(x)$ approximately

Markov Chain Monte Carlo (MCMC) Sampling

Bishop Section 11.2



Intuition behind MCMC

Standard MC and Importance sampling do not work well in high dimensions

High dimensional space but actual model has lower (VC) dimension => exploit correlation!

Instead of drawing independent samples x^s draw chains of correlated samples – perform random walk in the data where the number of visits to x is proportional to target density $p(x)$

Random walk = Markov chain

What is a Markov Chain?

Definition: a *stochastic process* in which future states are independent of past states given the present state

Stochastic process: a *consecutive* set of *random* (not deterministic) quantities defined on some known state space Θ.

- think of Θ as our parameter space.
- *consecutive* implies a time component, indexed by t.

Consider a draw of $\boldsymbol{\theta}^{(t)}$ to be a state at iteration t. The next draw $\boldsymbol{\theta}^{(t+1)}$ is dependent only on the current draw $\boldsymbol{\theta}^{(t)}$, and not on any past draws.

This satisfies the **Markov property**:

$$p(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(t)}) = p(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})$$

So our Markov chain is a bunch of draws of $\boldsymbol{\theta}$ that are each slightly dependent on the previous one. The chain wanders around the parameter space, remembering only where it has been in the last period.

What are the rules governing how the chain jumps from one state to another at each period?

The jumping rules are governed by a **transition kernel**, which is a mechanism that describes the probability of moving to some other state based on the current state.

Transition Kernel

For discrete state space (k possible states): a $k \times k$ matrix of transition probabilities.

Example: Suppose $k = 3$. The 3×3 transition matrix $\mathbf{P}$ would be

$p(\boldsymbol{\theta}_{\mathbf{A}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{A}}^{(t)})$	$p(\boldsymbol{\theta}_{\mathbf{B}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{A}}^{(t)})$	$p(\boldsymbol{\theta}_{\mathbf{C}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{A}}^{(t)})$
$p(\boldsymbol{\theta}_{\mathbf{A}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{B}}^{(t)})$	$p(\boldsymbol{\theta}_{\mathbf{B}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{B}}^{(t)})$	$p(\boldsymbol{\theta}_{\mathbf{C}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{B}}^{(t)})$
$p(\boldsymbol{\theta}_{\mathbf{A}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{C}}^{(t)})$	$p(\boldsymbol{\theta}_{\mathbf{B}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{C}}^{(t)})$	$p(\boldsymbol{\theta}_{\mathbf{C}}^{(t+1)}\|\boldsymbol{\theta}_{\mathbf{C}}^{(t)})$

where the subscripts index the 3 possible values that $\boldsymbol{\theta}$ can take.

The rows sum to one and define a conditional PMF, conditional on the current state. The columns are the marginal probabilities of being in a certain state in the next period.

For continuous state space (infinite possible states), the transition kernel is a bunch of conditional PDFs: $f(\boldsymbol{\theta}_{\mathbf{j}}^{(t+1)}|\boldsymbol{\theta}_{\mathbf{i}}^{(t)})$

Stationary (Limiting) Distribution

Define a stationary distribution π to be some distribution $\prod$ such that $\pi = \pi \mathbf{P}$.

For all the MCMC algorithms we use in Bayesian statistics, the Markov chain will typically **converge** to π regardless of our starting points.

So if we can devise a Markov chain whose stationary distribution π is our desired posterior distribution $p(\boldsymbol{\theta}|y)$, then we can run this chain to get draws that are approximately from $p(\boldsymbol{\theta}|y)$ once the chain has converged.

Monte Carlo Integration on the Markov Chain

Once we have a Markov chain that has converged to the stationary distribution, then the draws in our chain appear to be like draws from $p(\boldsymbol{\theta}|y)$, so it seems like we should be able to use Monte Carlo Integration methods to find quantities of interest.

One problem: our draws are not independent, which we required for Monte Carlo Integration to work (remember SLLN).

Luckily, we have the **Ergodic Theorem**.

Ergodic Theorem

Let $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$ be M values from a Markov chain that is *aperiodic*, *irreducible*, and *positive recurrent* (then the chain is ergodic), and $E[g(\boldsymbol{\theta})] < \infty$.

Then with probability 1,

$$\frac{1}{M}\sum_{i=1}^{M} g(\boldsymbol{\theta}_i) \to \int_{\Theta} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

as $M \to \infty$, where π is the stationary distribution.

This is the Markov chain analog to the SLLN, and it allows us to ignore the dependence between draws of the Markov chain when we calculate quantities of interest from the draws.

But what does it mean for a chain to be *aperiodic*, *irreducible*, and *positive recurrent*, and therefore ergodic?

So Really, What is MCMC?

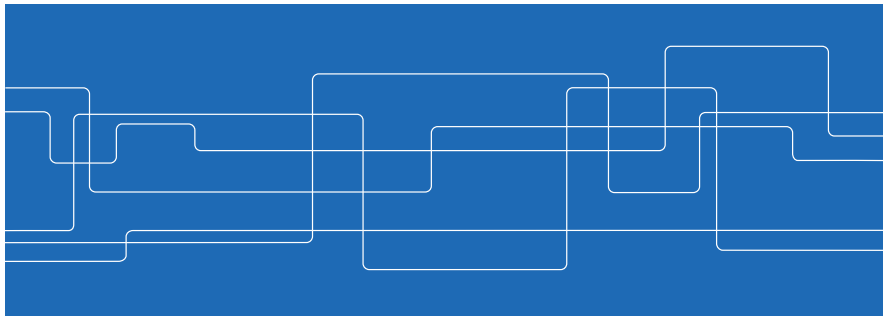
MCMC is a class of methods in which we can simulate draws that are slightly dependent and are approximately from a (posterior) distribution.

We then take those draws and calculate quantities of interest for the (posterior) distribution.

In Bayesian statistics, there are generally two MCMC algorithms that we use: the Gibbs Sampler and the Metropolis-Hastings algorithm.

Gibbs Sampling

Bishop Section 11.3



Gibbs Sampling

Suppose we have a joint distribution $p(\theta_1, \ldots, \theta_k)$ that we want to sample from (for example, a posterior distribution).

We can use the Gibbs sampler to sample from the joint distribution if we knew the **full conditional** distributions for each parameter.

For each parameter, the **full conditional** distribution is the distribution of the parameter conditional on the known information and all the other parameters: $p(\theta_j|\theta_{-j}, y)$

Gibbs Sampler Steps

Let's suppose that we are interested in sampling from the posterior $p(\boldsymbol{\theta}|\mathbf{y})$, where $\boldsymbol{\theta}$ is a vector of three parameters, $\theta_1, \theta_2, \theta_3$.

The steps to a Gibbs Sampler (and the analogous steps in the MCMC process) are

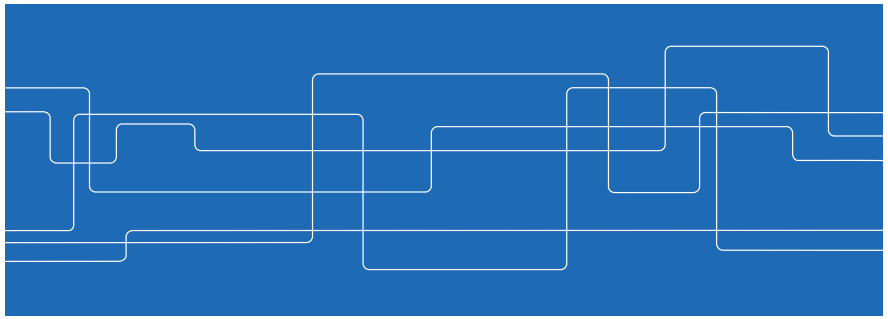
1. Pick a vector of starting values $\boldsymbol{\theta}^{(0)}$. (Defining a starting distribution $\Pi^{(0)}$ and drawing $\boldsymbol{\theta}^{(0)}$ from it.)
2. Start with any θ (order does not matter, but I'll start with θ_1 for convenience). Draw a value $\theta_1^{(1)}$ from the full conditional $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, \mathbf{y})$.

3. Draw a value $\theta_2^{(1)}$ (again order does not matter) from the full conditional $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \mathbf{y})$. Note that we must use the updated value of $\theta_1^{(1)}$.
4. Draw a value $\theta_3^{(1)}$ from the full conditional $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, \mathbf{y})$ using both updated values. (Steps 2-4 are analogous to multiplying $\Pi^{(0)}$ and $\mathbf{P}$ to get $\Pi^{(1)}$ and then drawing $\boldsymbol{\theta}^{(1)}$ from $\Pi^{(1)}$.)
5. Draw $\boldsymbol{\theta}^{(2)}$ using $\boldsymbol{\theta}^{(1)}$ and continually using the most updated values.
6. Repeat until we get M draws, with each draw being a vector $\boldsymbol{\theta}^{(t)}$.
7. Optional burn-in and/or thinning.

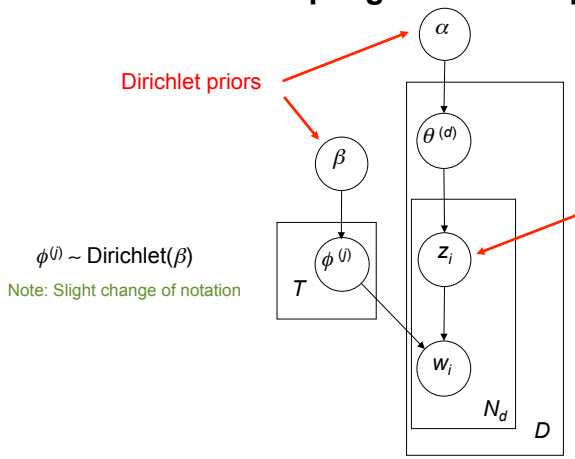
Our result is a Markov chain with a bunch of draws of $\boldsymbol{\theta}$ that are approximately from our posterior. We can do Monte Carlo Integration on those draws to get quantities of interest.

Some Intuitions about Gibbs Sampling in LDA

Griffiths



Gibbs sampling in LDA: Graphical model



Slide from Griffiths

Gibbs sampling in LDA: Intuition

For details to accomplish Task 2.6, see the paper by Griffith
For details to accomplish Task 2.7, see the original LDA paper, cited in Griffith

Sample from joint distribution over words, documents, topics
Sample *i* denoted [w_i, d_i, z_i]. We observe w_i, d_i, while z_i are hidden/latent
Gibbs sampling task – to find topic assignments z_i for each observed [w_i, d_i].

Once we have (sampled version of) distribution over (*w*, *d*, *z*), we can take the marginal over *w* which gives θ, and the marginal over *d* which gives Φ

Gibbs sampling in LDA: Example

T=2 **N_d=10** **M=5**

iteration 1

i	w_i	d_i	z_i
1	MATHEMATICS	1	2
2	KNOWLEDGE	1	2
3	RESEARCH	1	1
4	WORK	1	2
5	MATHEMATICS	1	1
6	RESEARCH	1	2
7	WORK	1	2
8	SCIENTIFIC	1	1
9	MATHEMATICS	1	2
10	WORK	1	1
11	SCIENTIFIC	2	1
12	KNOWLEDGE	2	1
.	.	.	.
.	.	.	.
.	.	.	.
50	JOY	5	2

Slide from Griffiths

Gibbs sampling in LDA: Example

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

Slide from Griffiths

Gibbs sampling in LDA: Example

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_w n_{-i,j}^{(w)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{\sum_k n_{-i,k}^{(d_i)} + T\alpha}$$

Slide from Griffiths

Gibbs sampling in LDA: Example

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	?
2	KNOWLEDGE	1	2	
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_w n_{-i,j}^{(w)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{\sum_k n_{-i,k}^{(d_i)} + T\alpha}$$

Slide from Griffiths

Gibbs sampling in LDA: Example

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	?
3	RESEARCH	1	1	
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_w n_{-i,j}^{(w)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{\sum_k n_{-i,k}^{(d_i)} + T\alpha}$$

Slide from Griffiths

Gibbs sampling in LDA: Example

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	?
4	WORK	1	2	
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_w n_{-i,j}^{(w)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{\sum_k n_{-i,k}^{(d_i)} + T\alpha}$$

Slide from Griffiths

Gibbs sampling in LDA: Example

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	?
5	MATHEMATICS	1	1	
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_w n_{-i,j}^{(w)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{\sum_k n_{-i,k}^{(d_i)} + T\alpha}$$

Slide from Griffiths

Gibbs sampling in LDA: Example

			iteration 1	iteration 2
i	w_i	d_i	z_i	z_i
1	MATHEMATICS	1	2	2
2	KNOWLEDGE	1	2	1
3	RESEARCH	1	1	1
4	WORK	1	2	2
5	MATHEMATICS	1	1	?
6	RESEARCH	1	2	
7	WORK	1	2	
8	SCIENTIFIC	1	1	
9	MATHEMATICS	1	2	
10	WORK	1	1	
11	SCIENTIFIC	2	1	
12	KNOWLEDGE	2	1	
.	.	.	.	
.	.	.	.	
.	.	.	.	
50	JOY	5	2	

$$P(z_i = j \mid \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_w n_{-i,j}^{(w)} + W\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{\sum_k n_{-i,k}^{(d_i)} + T\alpha}$$

Slide from Griffiths

What is next?

Continue with Assignment 2, deadline **December 16**.

Paper assignments for project groups are published tonight, deadline **January 18**.

Next on the schedule

Fri 4 Dec 15:15-17:00 E3
Lecture 13: The Structure of a Scientific Paper
Hedvig Kjellström
Readings: Allen, Duvenaud et al.

Bring Duvenaud et al. on paper (or pdf) for reference!