10-708 Probabilistic Graphical Models

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Topic Modeling
+
Variational Inference

Matt Gormley
Lecture 16
Mar. 19, 2021

Reminders

- Project Proposal
  - Due: Wed, Mar. 31 at 11:59pm
- Homework 4: MCMC
  - Out: Wed, Mar. 24
  - Due: Wed, Apr. 7 at 11:59pm

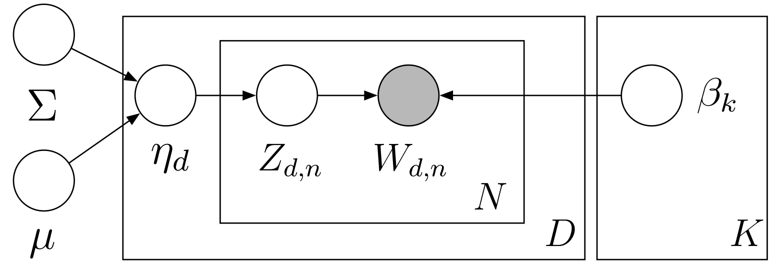
Outline

- **Applications of Topic Modeling**
- **Review: Latent Dirichlet Allocation (LDA)**
    1. Beta-Bernoulli
    2. Dirichlet-Multinomial
    3. Dirichlet-Multinomial Mixture Model
    4. LDA
- **Bayesian Inference for Parameter Estimation**
    - Exact inference
    - EM
    - Monte Carlo EM
    - Gibbs sampler
    - Collapsed Gibbs sampler
- **Extensions of LDA**
    - Correlated topic models
    - Dynamic topic models
    - Polylingual topic models
    - Supervised LDA

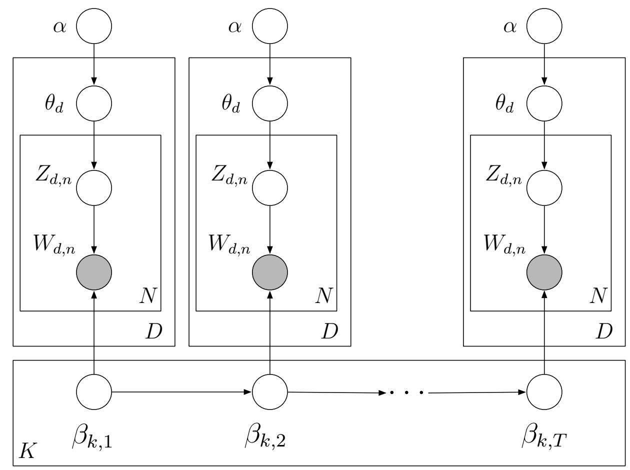
EXTENSIONS OF LDA

Extensions to the LDA Model

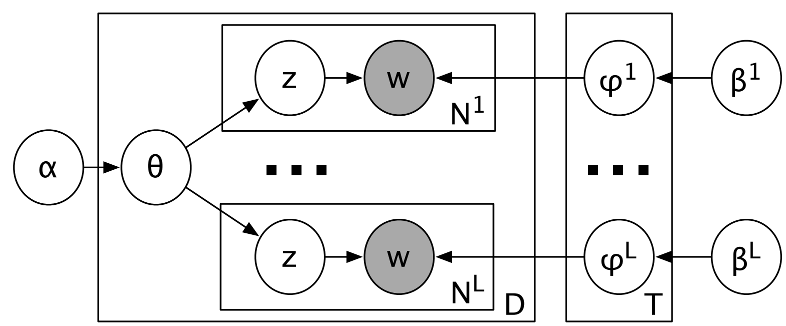
- Correlated topic models
  - Logistic normal prior over topic assignments



- Dynamic topic models
  - Learns topic changes over time



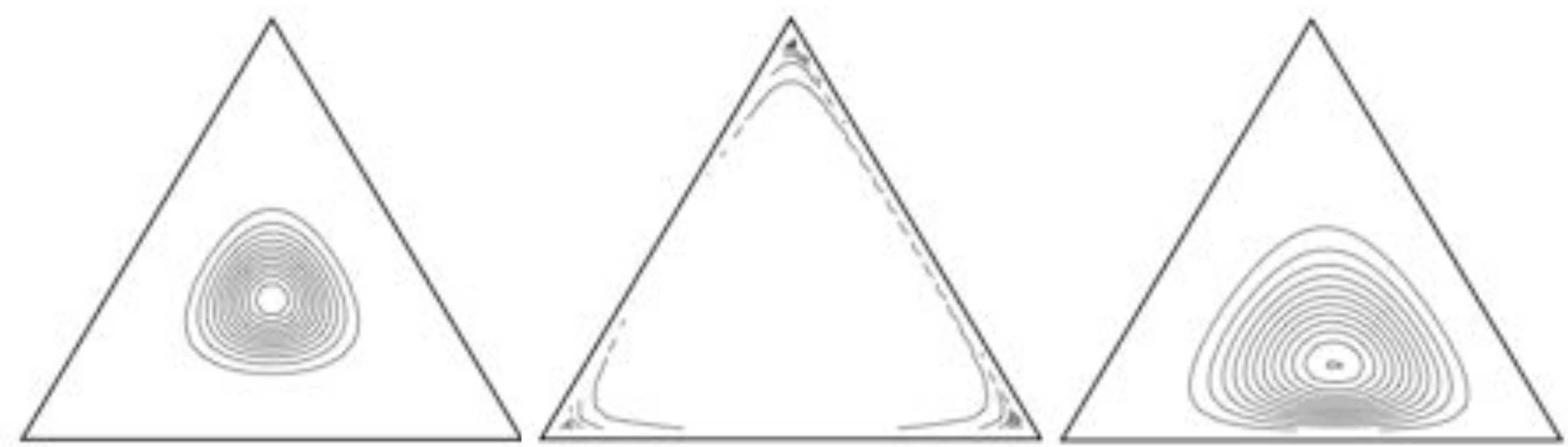
- Polylingual topic models
  - Learns topics aligned across multiple languages



…

(Blei & Lafferty, 2004)

Correlated Topic Models

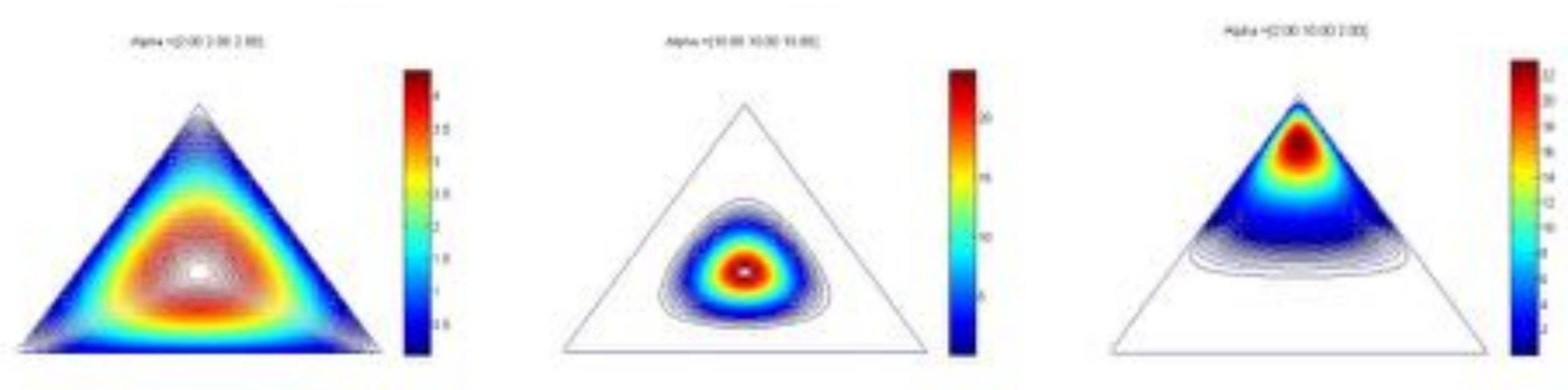


- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models

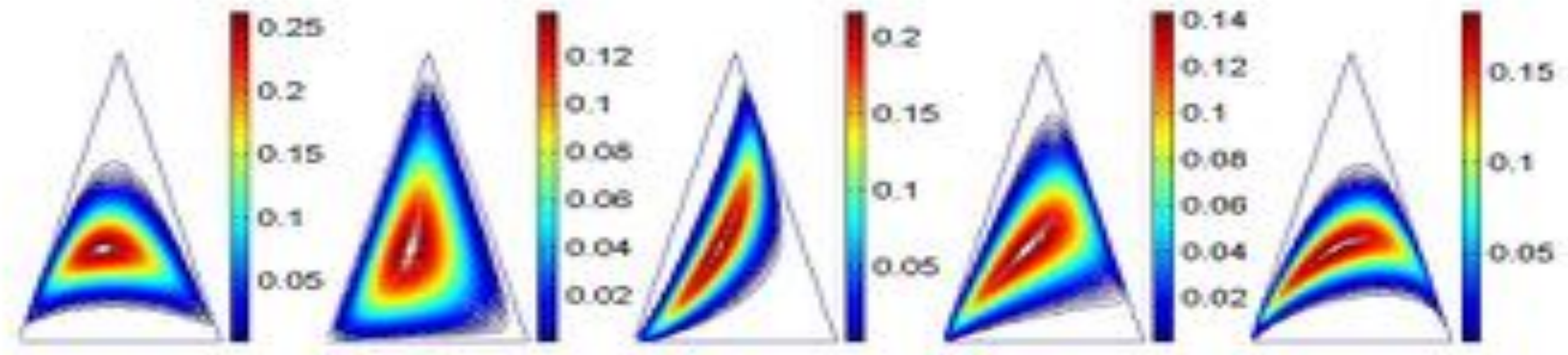


- The Dirichlet is a distribution on the simplex, positive vectors that sum to 1.
- It assumes that components are nearly independent.
- In real data, an article about *fossil fuels* is more likely to also be about *geology* than about *genetics*.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models



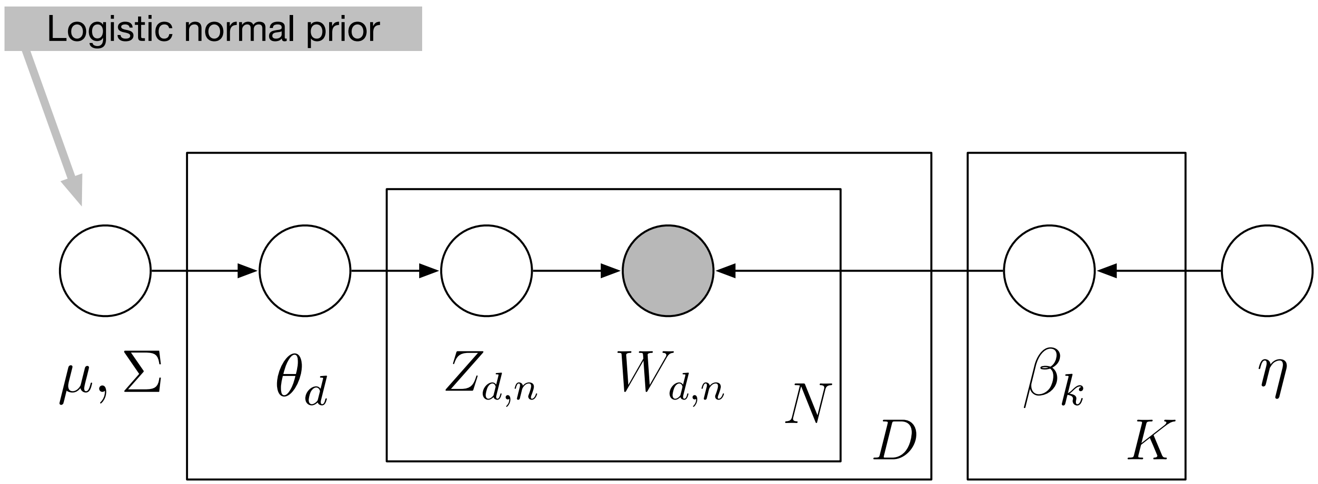
- The **logistic normal** is a distribution on the simplex that can model dependence between components (Aitchison, 1980).
- The log of the parameters of the multinomial are drawn from a multivariate Gaussian distribution,

$$
\begin{aligned}
X &\sim \mathcal{N}_K(\mu, \Sigma) \\
\theta_i &\propto \exp\{x_i\}.
\end{aligned}
$$

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models

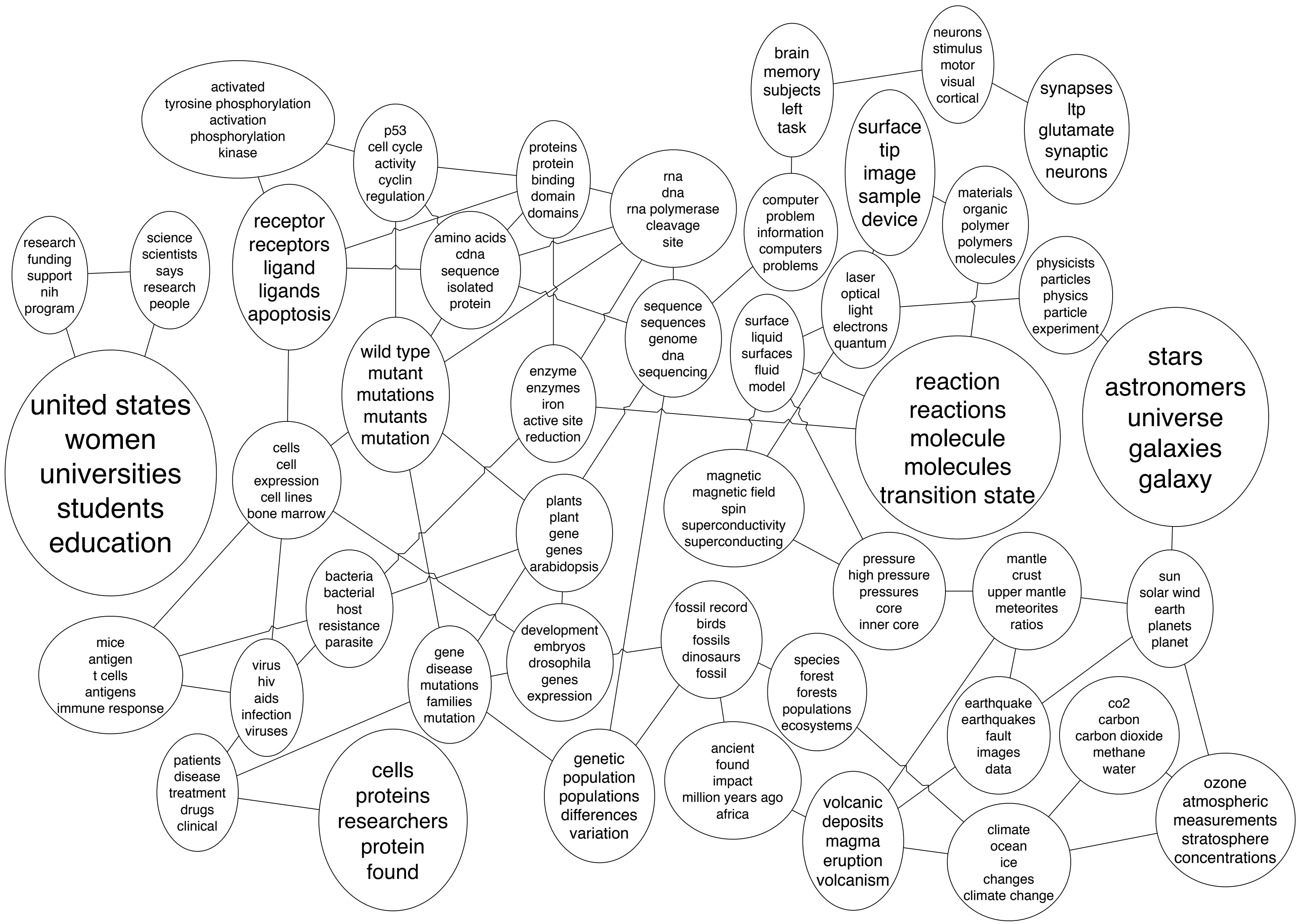


- Draw topic proportions from a logistic normal
- This allows topic occurrences to exhibit correlation.
- Provides a “map” of topics and how they are related
- Provides a better fit to text data, but computation is more complex

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2004)

Correlated Topic Models



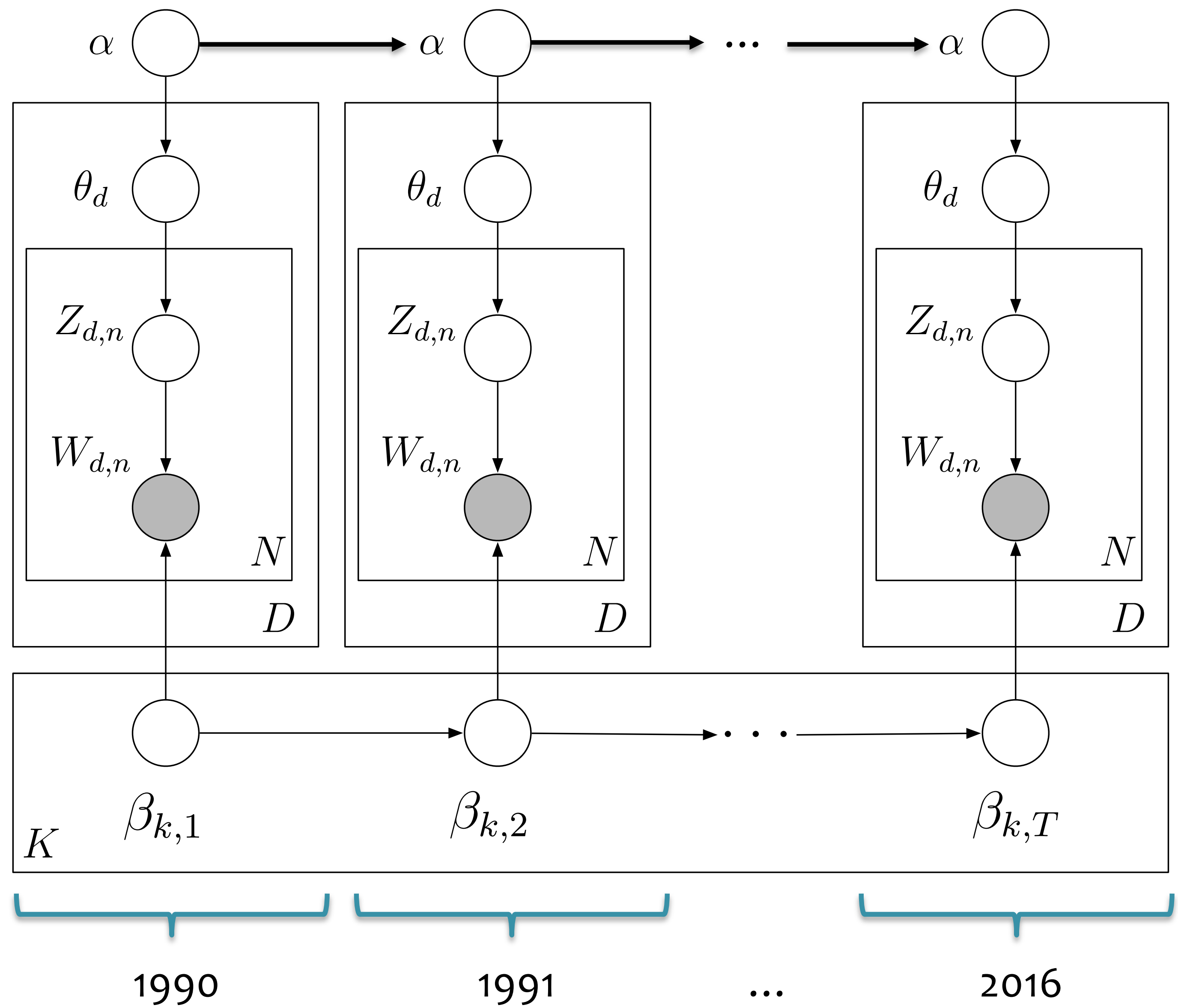
Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2006)

Dynamic Topic Models

High-level idea:

- Divide the documents up by year
- Start with a separate topic model for each year
- Then add a dependence of each year on the previous one



(Blei & Lafferty, 2006)

Dynamic Topic Models

1789

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors...

Inaugural addresses →

2009

AMONG the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order...

- LDA assumes that the order of documents does not matter.
- Not appropriate for sequential corpora (e.g., that span hundreds of years)
- Further, we may want to track how language changes over time.
- Dynamic topic models let the topics *drift* in a sequence.

Slide from David Blei, MLSS 2012

(Blei & Lafferty, 2006)

Dynamic Topic Models

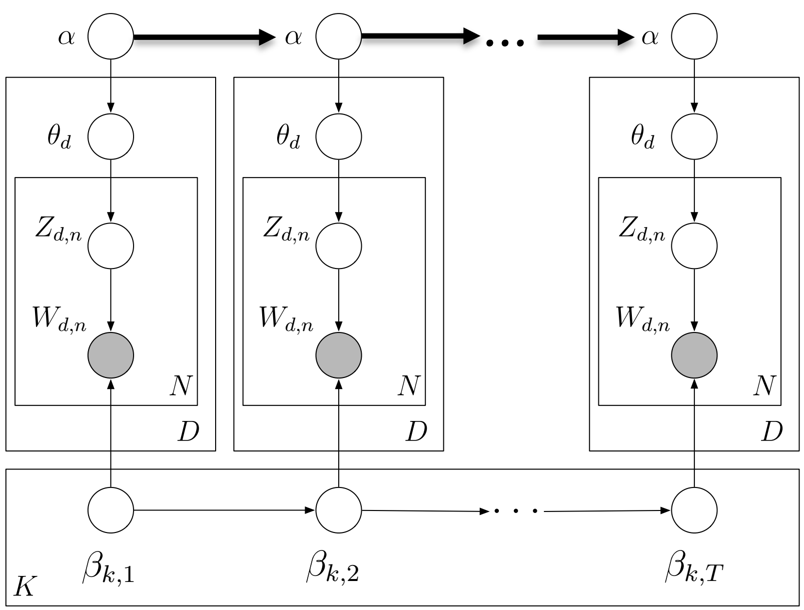
Generative Story

1. Draw topics $\beta_t \mid \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t \mid \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
   - (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
   - (b) For each word:
     - i. Draw $Z \sim \mathit{Mult}(\pi(\eta))$.
     - ii. Draw $W_{t,d,n} \sim \mathit{Mult}(\pi(\beta_{t,z}))$.

Logistic-normal priors

The pi function maps from the natural parameters to the mean parameters:

$$\pi(\beta_{k,t})_w = \frac{\exp(\beta_{k,t,w})}{\sum_w \exp(\beta_{k,t,w})}.$$



(Blei & Lafferty, 2006)

Dynamic Topic Models

Top ten most likely words in a "drifting" topic shown at 10-year increments

1880	1890	1900	1910	1920	1930	1940
electric	electric	apparatus	air	apparatus	tube	air
machine	power	steam	water	tube	apparatus	tube
power	company	power	engineering	air	glass	apparatus
engine	steam	engine	apparatus	pressure	air	glass
steam	electrical	engineering	room	water	mercury	laboratory
two	machine	water	laboratory	glass	laboratory	rubber
machines	two	construction	engineer	gas	pressure	pressure
iron	system	engineer	made	made	made	small
battery	motor	room	gas	laboratory	gas	mercury
wire	engine	feet	tube	mercury	small	gas

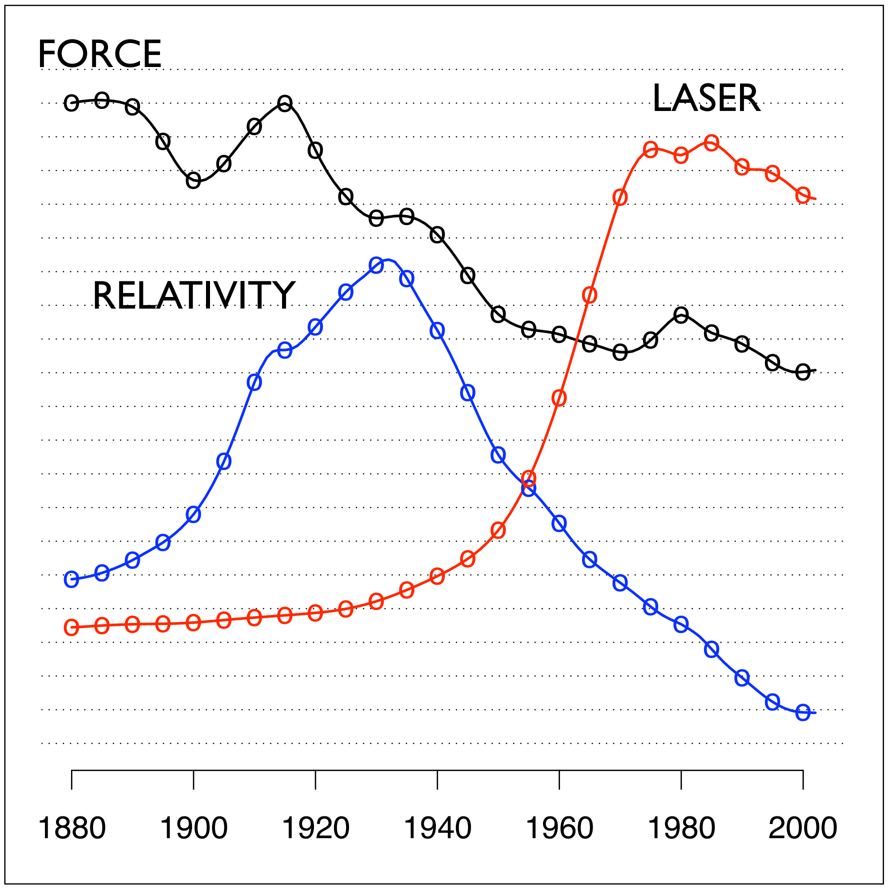
1950	1960	1970	1980	1990	2000
tube	tube	air	high	materials	devices
apparatus	system	heat	power	high	device
glass	temperature	power	design	power	materials
air	air	system	heat	current	current
chamber	heat	temperature	system	applications	gate
instrument	chamber	chamber	systems	technology	high
small	power	high	devices	devices	light
laboratory	high	flow	instruments	design	silicon
pressure	instrument	tube	control	device	material
rubber	control	design	large	heat	technology

(Blei & Lafferty, 2006)

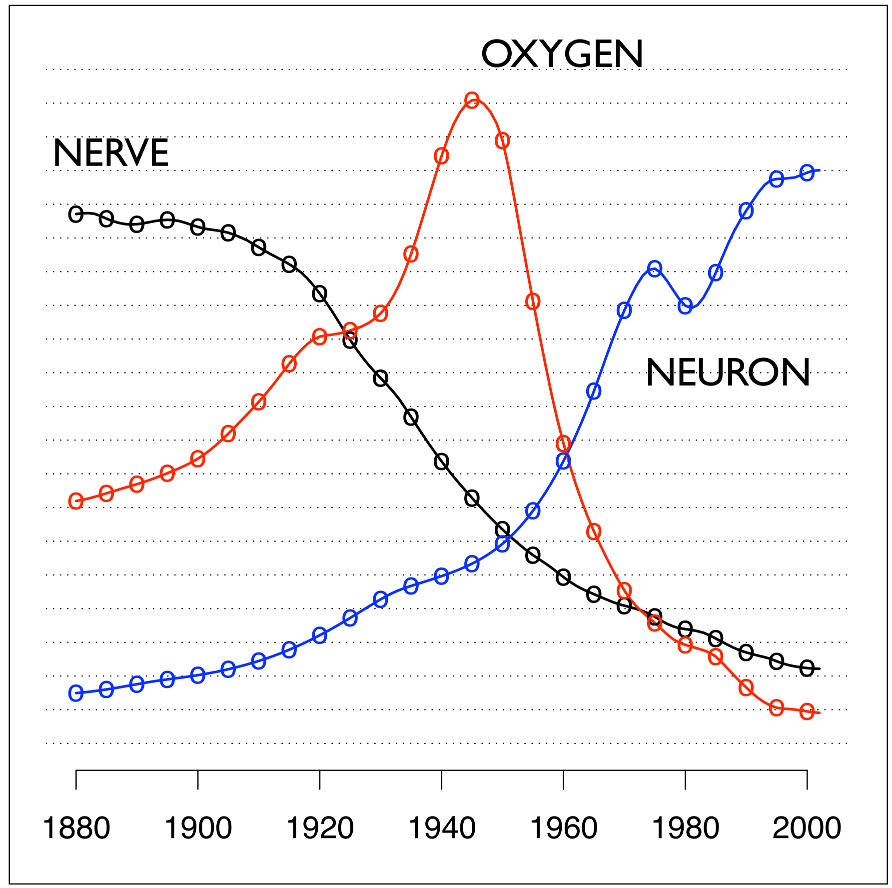
Dynamic Topic Models

Posterior estimate of **word frequency as a function of year** for three words each in two separate topics:

"Theoretical Physics"



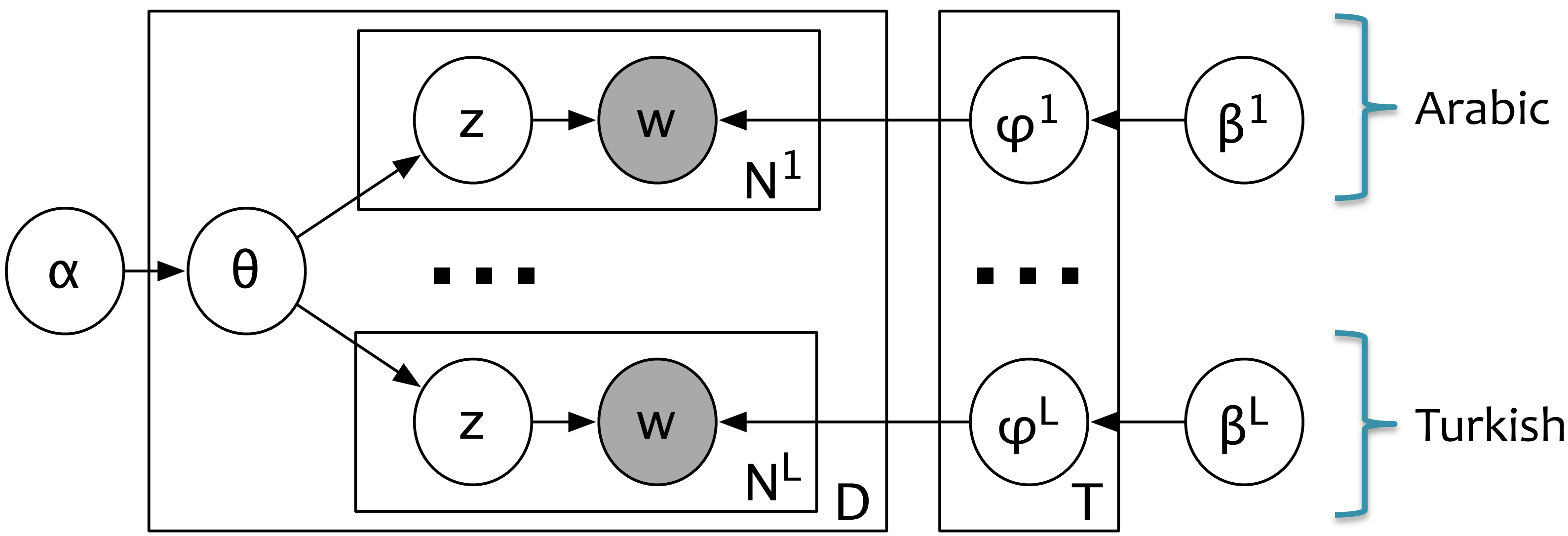
"Neuroscience"



(Mimno et al., 2009)

Polylingual Topic Models

- **Data Setting:** Comparable versions of each document exist in multiple languages (e.g. the Wikipedia article for "Barak Obama" in twelve languages)
- **Model:** Very similar to LDA, except that the topic assignments, z, and words, w, are sampled separately for each language.



(Mimno et al., 2009)

Polylingual Topic Models

Topic 1 (twelve languages)

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضایی ماموریت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

(Mimno et al., 2009)

Polylingual Topic Models

Topic 2 (twelve languages)

CY	sbaen madrid el la josé sbaeneg
DE	de spanischer spanischen spanien madrid la
EL	ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN	**de spanish spain la madrid y**
FA	ترین de اسپانیا اسپانیایی کوبا مادرید
FI	espanja de espanjan madrid la real
FR	espagnol espagne madrid espagnole juan y
HE	ספרד ספרדית דה מדריד הספרדית קובה
IT	de spagna spagnolo spagnola madrid el
PL	de hiszpański hiszpanii la juan y
RU	де мадрид испании испания испанский de
TR	ispanya ispanyol madrid la küba real

(Mimno et al., 2009)

Polylingual Topic Models

Topic 3 (twelve languages)

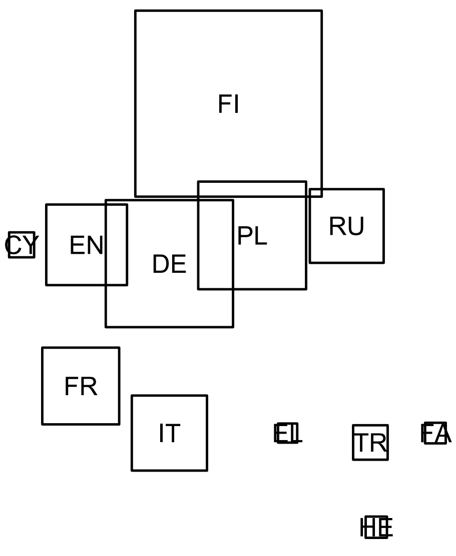
CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

(Mimno et al., 2009)

Polylingual Topic Models

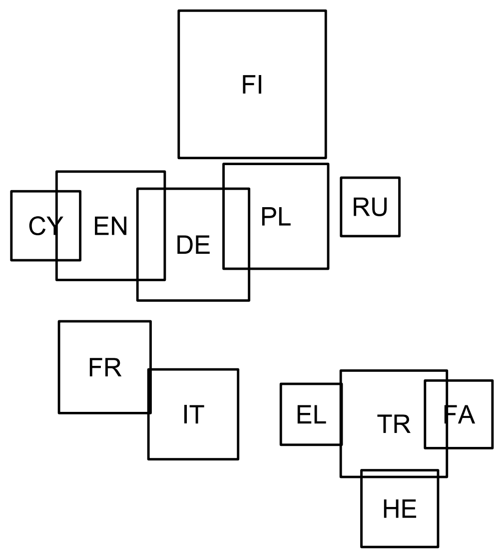
Size of each square represents proportion of tokens assigned to the specified topic.

world
ski
km
won



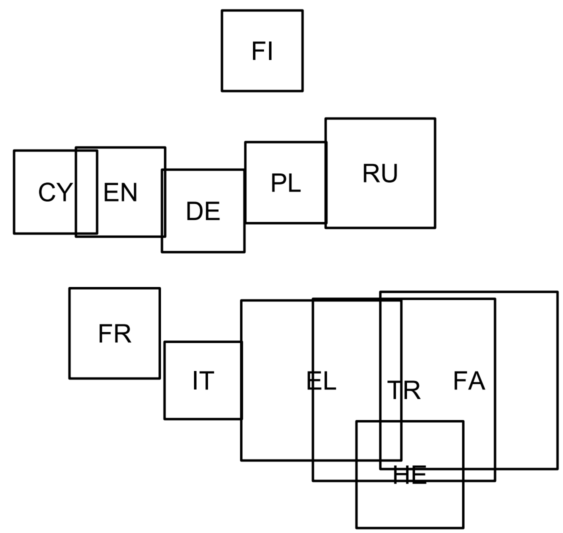
Analysis: mostly Nordic countries

actor
role
television
actress



Analysis: uniform across countries

ottoman
empire
khan
byzantine



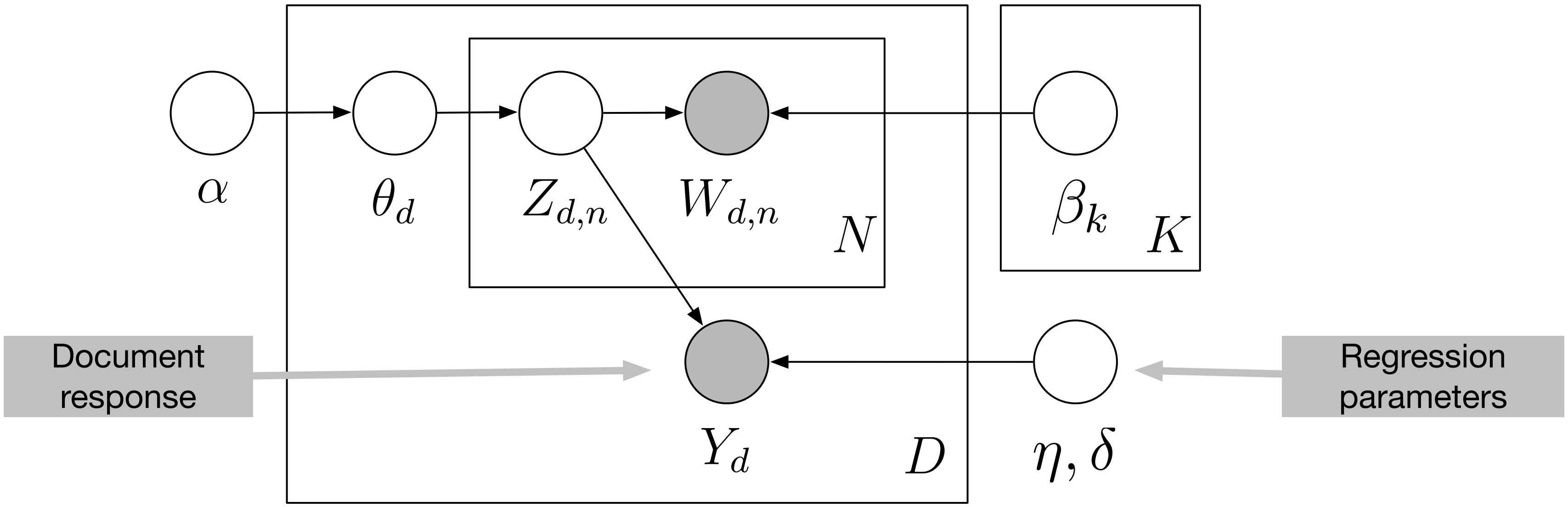
Analysis: mostly countries near Istanbul

Supervised LDA

- LDA is an unsupervised model. How can we build a topic model that is good at the task we care about?
- Many data are paired with **response variables**.
  - User reviews paired with a number of stars
  - Web pages paired with a number of “likes”
  - Documents paired with links to other documents
  - Images paired with a category
- **Supervised LDA** are topic models of documents and responses. They are fit to find topics predictive of the response.

Slide from David Blei, MLSS 2012

Supervised LDA



1. Draw topic proportions $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$.
2. For each word
   - Draw topic assignment $z_n \mid \theta \sim \mathrm{Mult}(\theta)$.
   - Draw word $w_n \mid z_n, \beta_{1:K} \sim \mathrm{Mult}(\beta_{z_n})$.
3. Draw response variable $y \mid z_{1:N}, \eta, \sigma^2 \sim \mathrm{N}\left(\eta^\top \bar{z}, \sigma^2\right)$, where

$$\bar{z} = (1/N) \sum_{n=1}^{N} z_n.$$

Slide from David Blei, MLSS 2012

Summary: Topic Modeling

- **The Task of Topic Modeling**
  - Topic modeling enables the **analysis of large** (possibly unannotated) **corpora**
  - Applicable to more than just bags of words
  - Extrinsic evaluations are often appropriate for these unsupervised methods
- **Constructing Models**
  - LDA is comprised of **simple building blocks** (Dirichlet, Multinomial)
  - LDA itself can act as a building block **for other models**
- **Approximate Inference**
  - Many different approaches to inference (and learning) can be applied to the same model

What if we don't know the number of topics, K, ahead of time?

Solution: Bayesian Nonparametrics

- New modeling constructs:
  - Chinese Restaurant Process (Dirichlet Process)
  - Indian Buffet Process
- e.g. an **infinite number of topics** in a finite amount of space

Summary: Approximate Inference

- Markov Chain Monte Carlo (MCMC)
  - Metropolis-Hastings, Gibbs sampling, Hamiltonion MCMC, slice sampling, etc.
- Variational inference
  - Minimizes KL(q||p) where q is a simpler graphical model than the original p
- Loopy Belief Propagation
  - Belief propagation applied to general (loopy) graphs
- Expectation propagation
  - Approximates belief states with moments of simpler distributions
- Spectral methods
  - Uses tensor decompositions (e.g. SVD)

Slice Sampling, Hamiltonian Monte Carlo

MCMC (AUXILIARY VARIABLE METHODS)

Auxiliary variables

The point of MCMC is to marginalize out variables, but one can introduce more variables:

$$\int f(x)P(x)\,\mathrm{d}x = \int f(x)P(x,v)\,\mathrm{d}x\,\mathrm{d}v$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(x^{(s)}), \quad x,v \sim P(x,v)$$

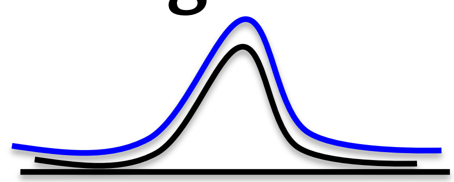
We might want to do this if

- $P(x|v)$ and $P(v|x)$ are simple
- $P(x,v)$ is otherwise easier to navigate

Slice Sampling

- Motivation:
  - Want **samples** from $p(x)$ and don't know the normalizer Z
  - Choosing a proposal at the correct **scale** is difficult
- Properties:
  - *Similar to Gibbs Sampling*: **one-dimensional** transitions in the state space
  - *Similar to Rejection Sampling*: (asymptotically) draws samples from the **region under the curve**

$\tilde{p}(x)$

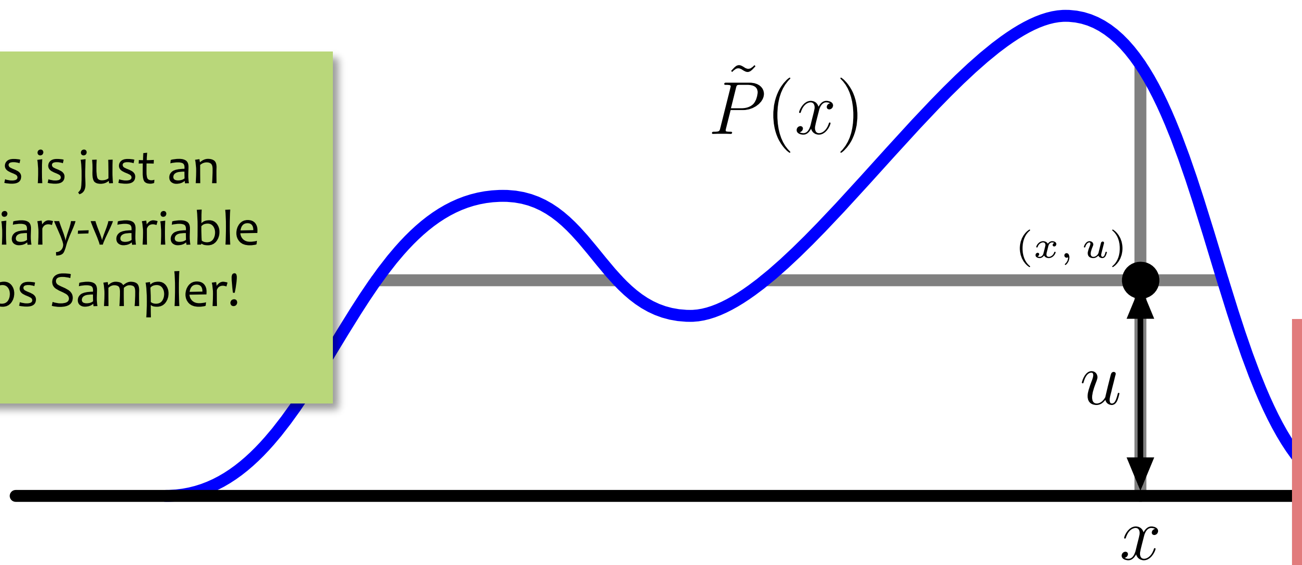


  - An MCMC method with an **adaptive proposal**

Slice sampling idea

Sample point uniformly under curve $\tilde{P}(x) \propto P(x)$

This is just an auxiliary-variable Gibbs Sampler!



Problem: Sampling from the conditional $p(x \mid u)$ might be infeasible.

$$p(u|x) = \text{Uniform}[0, \tilde{P}(x)]$$

$$p(x|u) \propto \begin{cases} 1 & \tilde{P}(x) \geq u \\ 0 & \text{otherwise} \end{cases} = \text{“Uniform on the slice”}$$

Slice Sampling

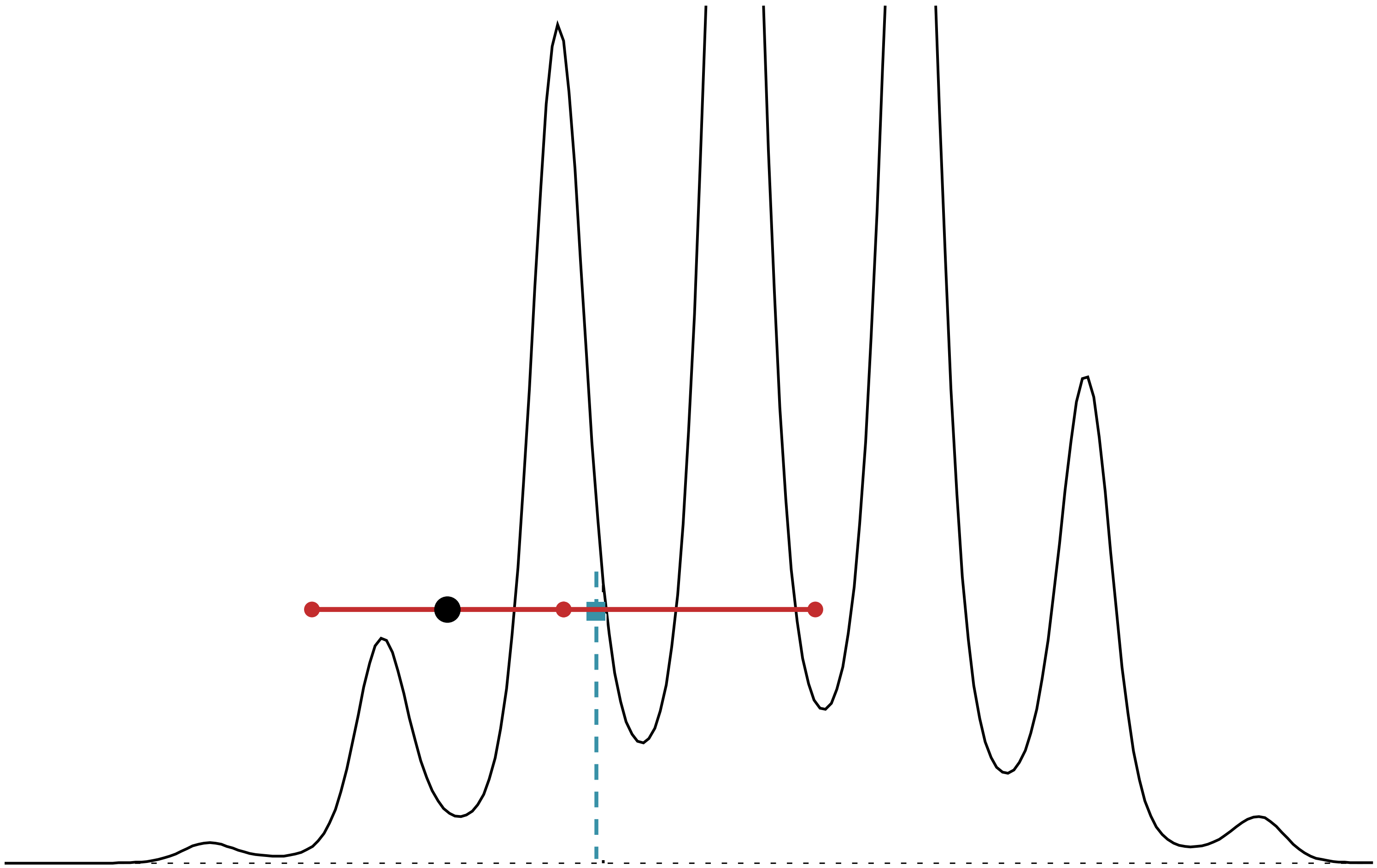


Figure adapted from MacKay Ch. 29

Slice Sampling

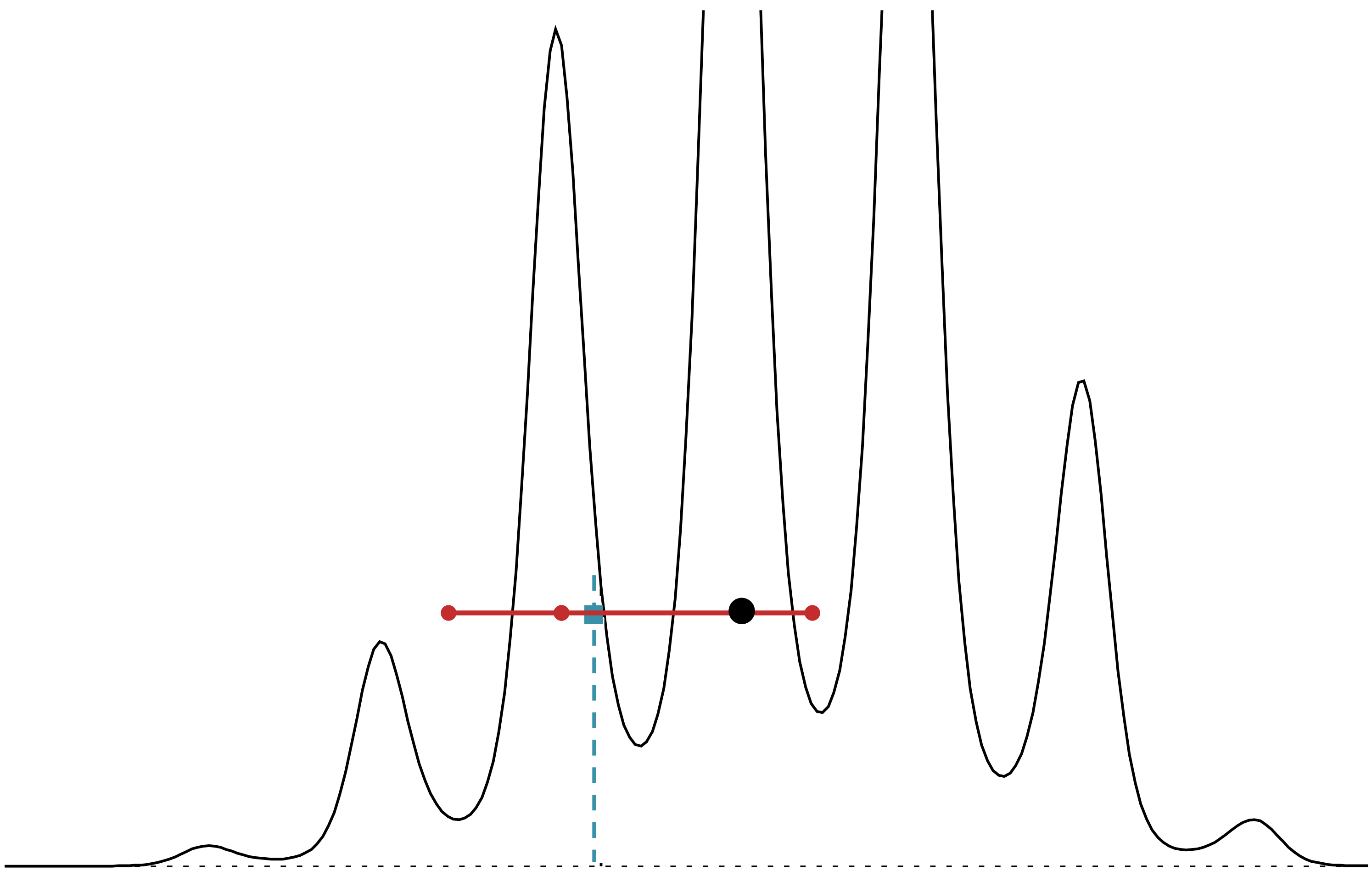
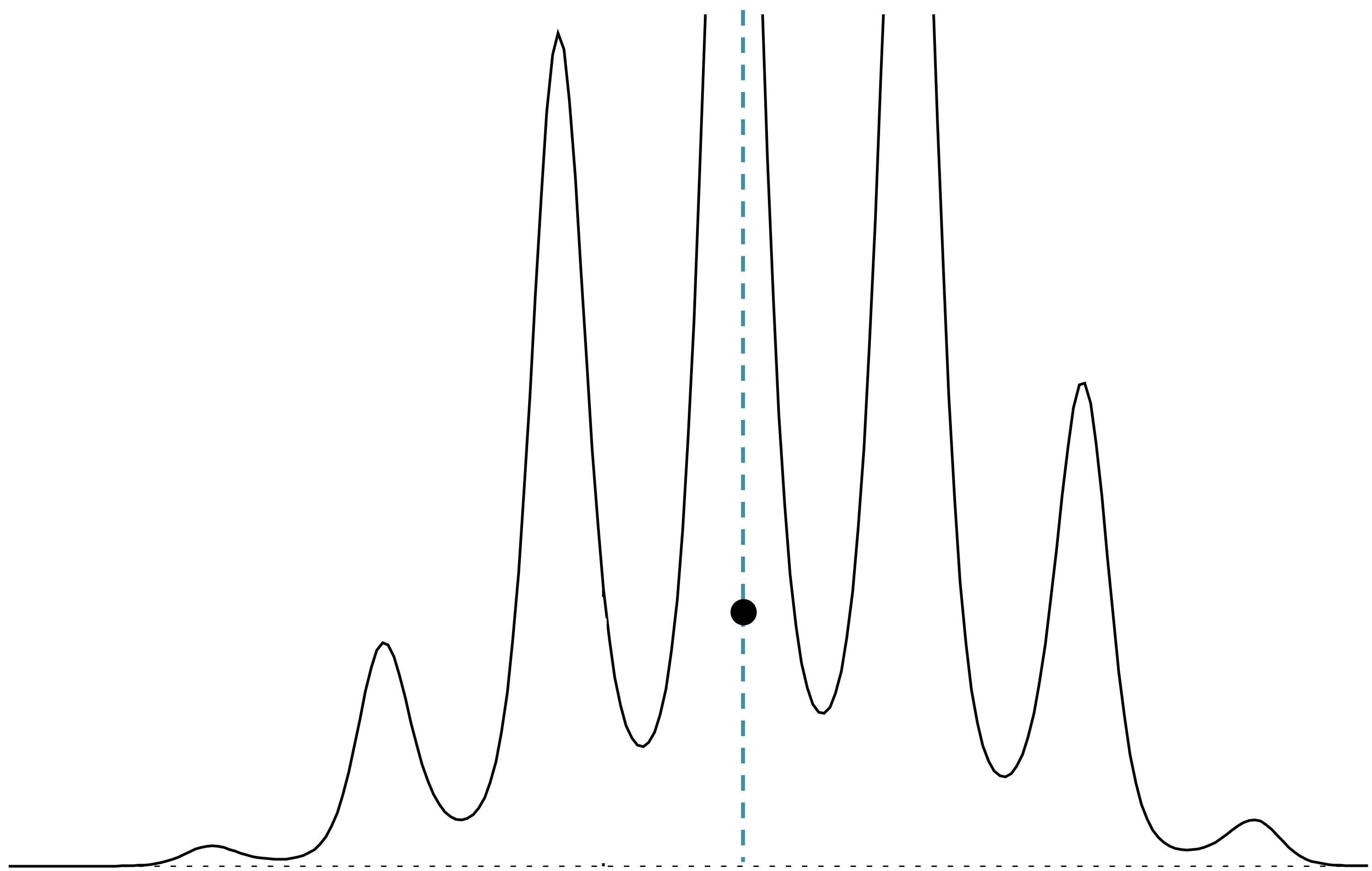


Figure adapted from MacKay Ch. 29

Slice Sampling



Slice Sampling

Algorithm:

Goal: sample (x, u) given $(u^{(t)}, x^{(t)})$.

Part 1: Stepping Out

Sample interval (x_l, x_r) enclosing $x^{(t)}$.

Expand until endpoints are "outside" region under curve.

Part 2: Sample x (Shrinking)

Draw x from within the interval (x_l, x_r), then accept or shrink.

Slice Sampling

Algorithm:

Goal: sample (x, u) given $(u^{(t)}, x^{(t)})$.

$u \sim \text{Uniform}(0, p(x^{(t)})$

Part 1: Stepping Out

Sample interval (x_l, x_r) enclosing $x^{(t)}$.

$r \sim \text{Uniform}(u, w)$

$(x_l, x_r) = (x^{(t)} - r, x^{(t)} + w - r)$

Expand until endpoints are "outside" region under curve.

$\text{while}(\tilde{p}(x_l) > u)\{x_l = x_l - w\}$

$\text{while}(\tilde{p}(x_r) > u)\{x_r = x_r + w\}$

Part 2: Sample x (Shrinking)

Draw x from within the interval (x_l, x_r), then accept or shrink.

Slice Sampling

Algorithm:

Goal: sample (x, u) given $(u^{(t)}, x^{(t)})$.

$u \sim \text{Uniform}(0, p(x^{(t)}))$

Part 1: Stepping Out

Sample interval (x_l, x_r) enclosing $x^{(t)}$.

$r \sim \text{Uniform}(u, w)$

$(x_l, x_r) = (x^{(t)} - r, x^{(t)} + w - r)$

Expand until endpoints are "outside" region under curve.

$\text{while}(\tilde{p}(x_l) > u)\{x_l = x_l - w\}$

$\text{while}(\tilde{p}(x_r) > u)\{x_r = x_r + w\}$

Part 2: Sample x (Shrinking)

while(true) {

Draw x from within the interval (x_l, x_r), then accept or shrink.

$x \sim \text{Uniform}(x_l, x_r)$

$\text{if}(\tilde{p}(x) > u)\{\text{break}\}$

$\text{else if}(x > x^{(t)})\{x_r = x\}$

$\text{else}\{x_l = x\}$

}

$x^{(t+1)} = x, \;\; u^{(t+1)} = u$

Slice Sampling

Multivariate Distributions

- Resample each variable x_i one-at-a-time (just like Gibbs Sampling)
- Does not require sampling from

$$p(x_i|\{x_j\}_{j \neq i})$$

- Only need to evaluate a quantity proportional to the conditional

$$p(x_i|\{x_j\}_{j \neq i}) \propto \tilde{p}(x_i|\{x_j\}_{j \neq i})$$

Hamiltonian Monte Carlo

- Suppose we have a distribution of the form:

$$p(\boldsymbol{x}) = \exp\{-E(\boldsymbol{x})\}/Z$$

where $\boldsymbol{x} \in \mathcal{R}^N$

- We could use **random-walk M-H** to draw samples, but it seems a shame to **discard gradient information** $\nabla_{\boldsymbol{x}} E(\boldsymbol{x})$

- If we can evaluate it, the gradient tells us where to look for **high-probability regions**!

Background: Hamiltonian Dynamics

Applications:

- Following the motion of atoms in a fluid through time
- Integrating the motion of a solar system over time
- Considering the evolution of a galaxy (i.e. the motion of its stars)
- "molecular dynamics"
- "N-body simulations"

Properties:

- Total energy of the system H(x,p) stays constant
- Dynamics are reversible

Important for detailed balance

Background: Hamiltonian Dynamics

Let $\boldsymbol{x} \in \mathcal{R}^N$ be a position

$\boldsymbol{p} \in \mathcal{R}^N$ be a momentum

Potential energy: $E(\boldsymbol{x})$

Kinetic energy: $K(\boldsymbol{p}) = \boldsymbol{p}^T \boldsymbol{p}/2$

Total energy: $H(\boldsymbol{x}, \boldsymbol{p}) = E(\boldsymbol{x}) + K(\boldsymbol{p})$

Hamiltonian function

Given a starting position $x^{(1)}$ and a starting momentum $p^{(1)}$ we can simulate the Hamiltonian dynamics of the system via:

1. Euler's method
2. Leapfrog method
3. etc.

Background: Hamiltonian Dynamics

Parameters to tune:

1. Step size, ϵ
2. Number of iterations, L

Leapfrog Algorithm:

$$\text{for } \tau \text{ in } 1 \ldots L\text{:}$$

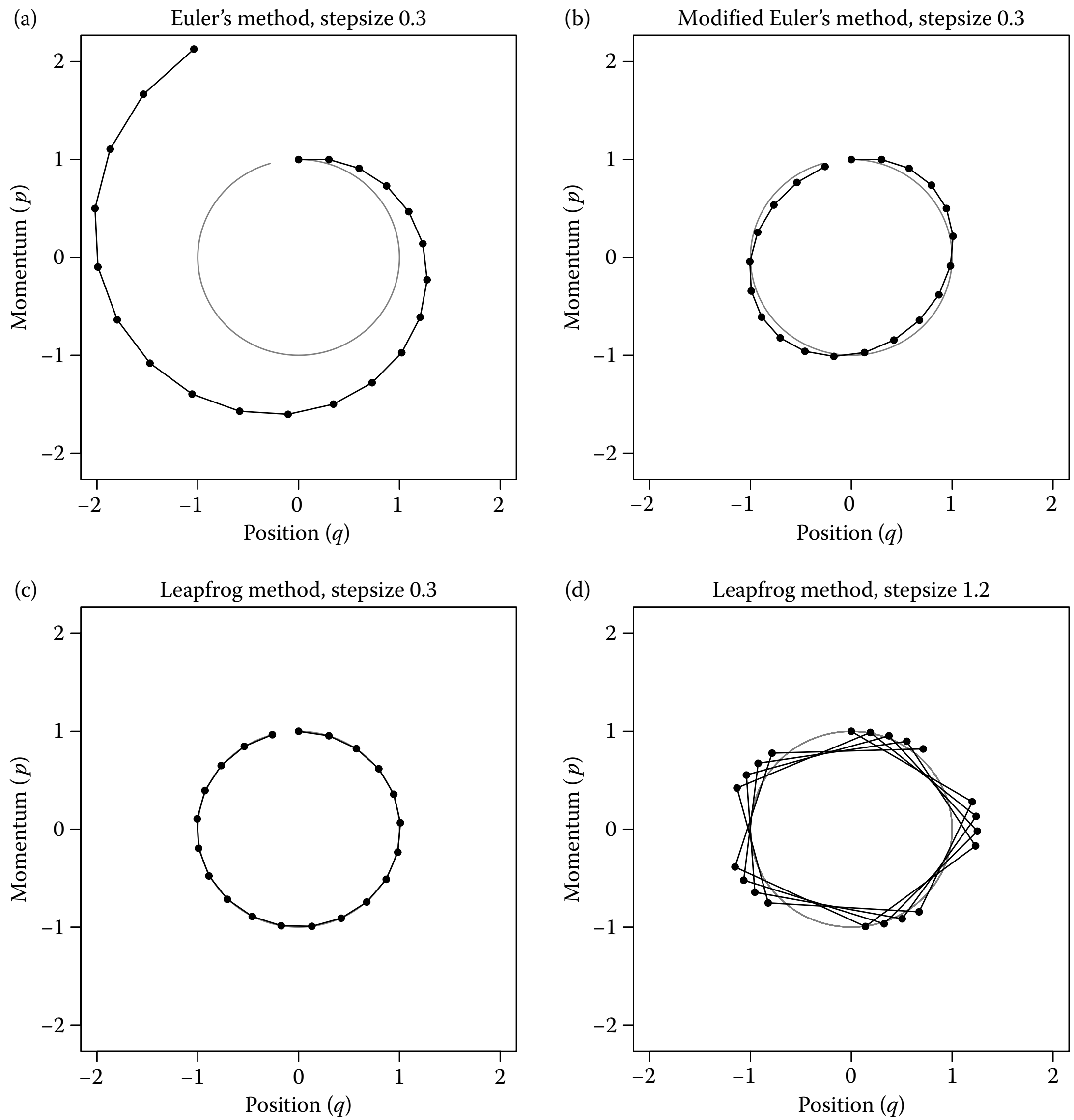
$$\boldsymbol{p} = \boldsymbol{p} - \frac{\epsilon}{2} \nabla_{\boldsymbol{x}} E(\boldsymbol{x})$$

$$\boldsymbol{x} = \boldsymbol{x} + \epsilon \boldsymbol{p}$$

$$\boldsymbol{p} = \boldsymbol{p} - \frac{\epsilon}{2} \nabla_{\boldsymbol{x}} E(\boldsymbol{x})$$

Figure from Neal (2011)

Background: Hamiltonian Dynamics



Hamiltonian Monte Carlo

Preliminaries

Goal: $p(\boldsymbol{x}) = \exp\{-E(\boldsymbol{x})\}/Z$ where $\boldsymbol{x} \in \mathcal{R}^N$

Define:

$$K(\boldsymbol{p}) = \boldsymbol{p}^T \boldsymbol{p}/2$$

$$H(\boldsymbol{x}, \boldsymbol{p}) = E(\boldsymbol{x}) + K(\boldsymbol{p})$$

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{p}) &= \exp\{-H(\boldsymbol{x}, \boldsymbol{p})\}/Z_H \\ &= \exp\{-E(\boldsymbol{x}\} \exp\{-K(\boldsymbol{p})\}/Z_H \end{aligned}$$

Note:
Since p(x,p) is separable…

$$\Rightarrow \sum_{\boldsymbol{p}} p(\boldsymbol{x}, \boldsymbol{p}) = \exp\{-E(\boldsymbol{x}\}/Z$$ Target dist.

$$\Rightarrow \sum_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{p}) = \exp\{-K(\boldsymbol{x}\}/Z_K$$ Gaussian

Whiteboard

- Hamiltonian Monte Carlo algorithm (aka. Hybrid Monte Carlo)

Figure from Neal (2011)

Hamiltonian Monte Carlo

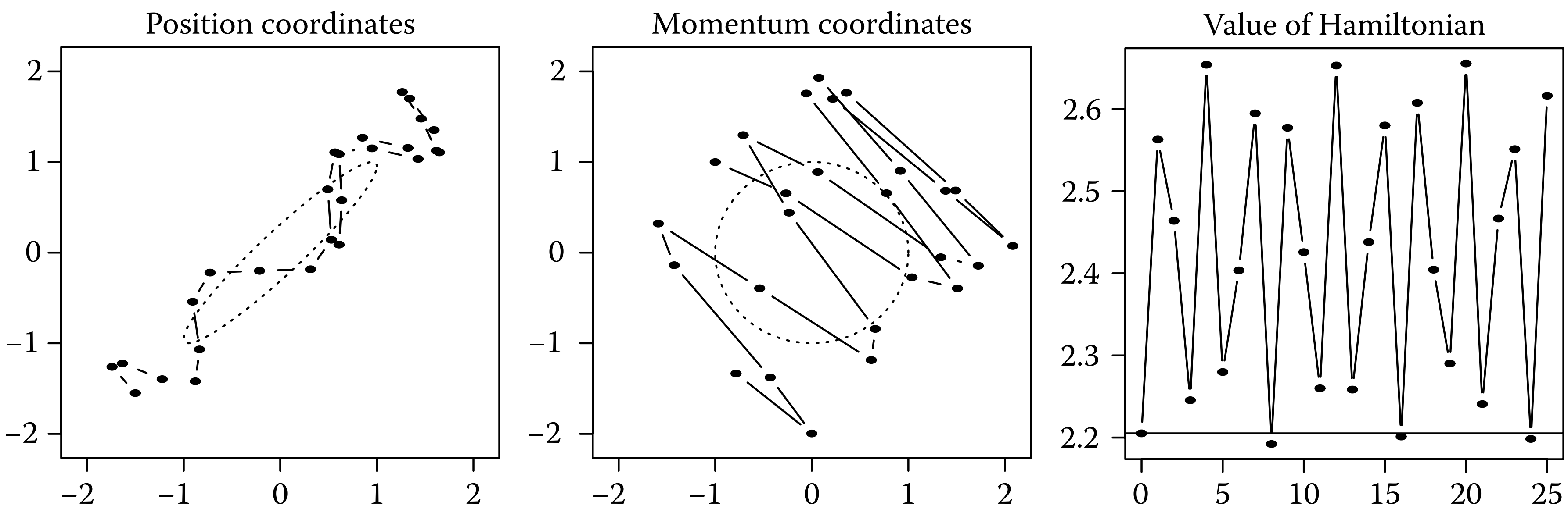
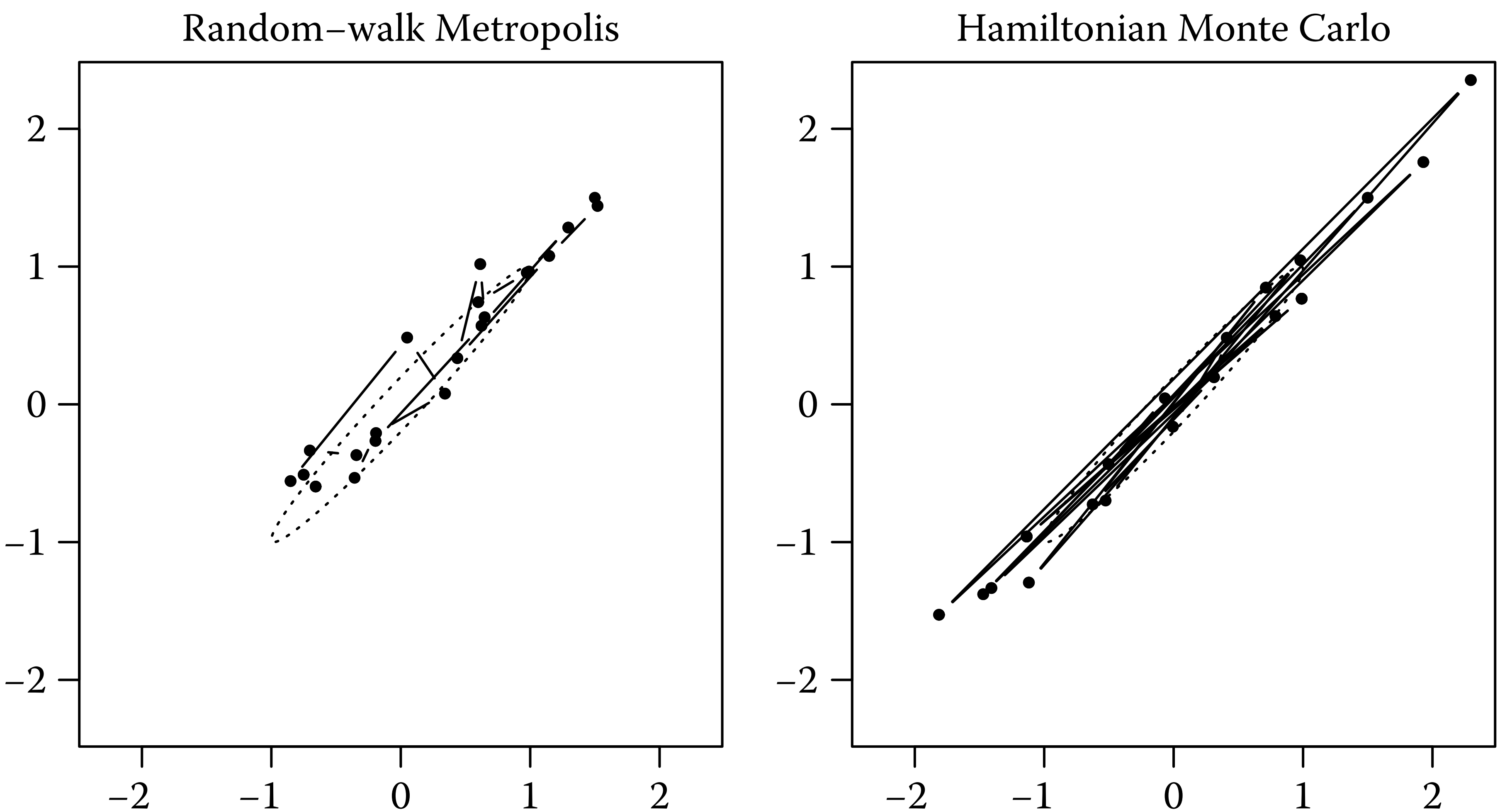


Figure from Neal (2011)

M-H vs. HMC

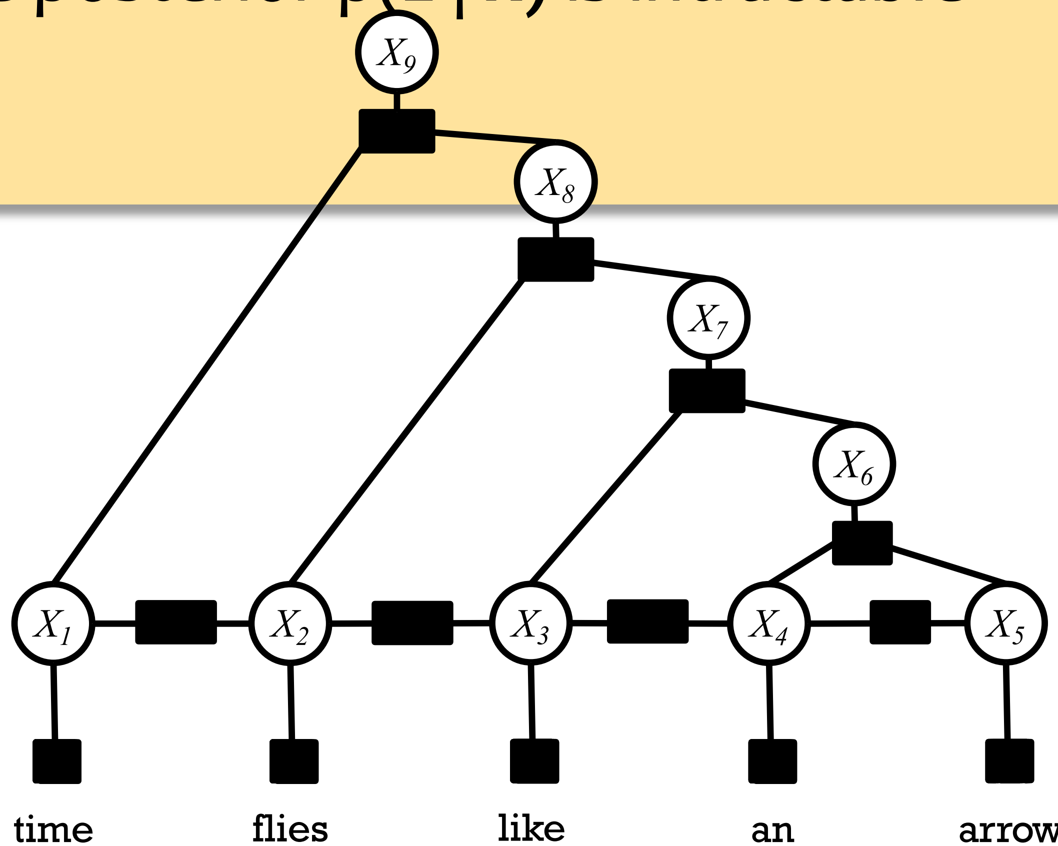


HIGH-LEVEL INTRO TO VARIATIONAL INFERENCE

Variational Inference

Problem:

– For observed variables **x** and latent variables **z**, estimating the posterior p(**z** | **x**) is intractable

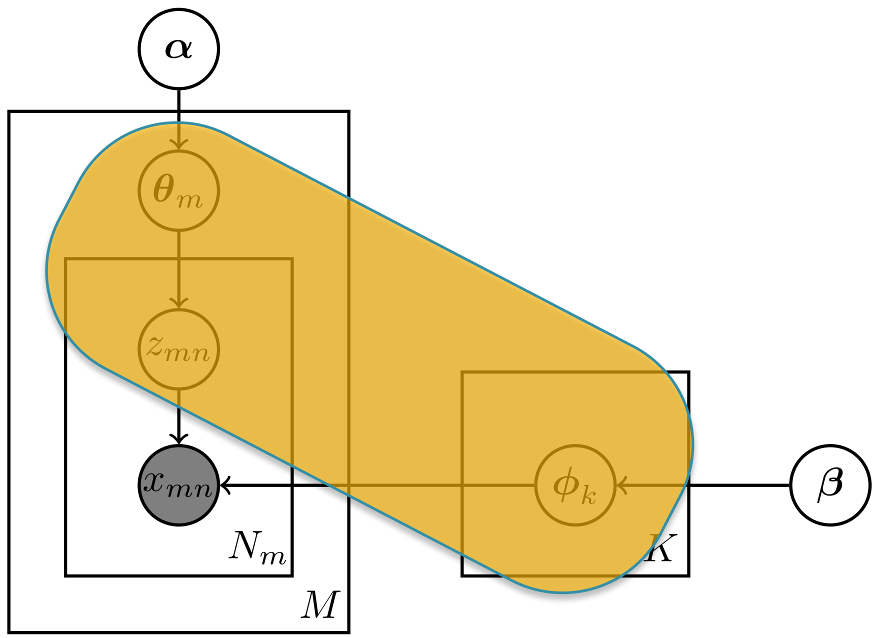


Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Problem:

- For observed variables **x** and latent variables **z**, estimating the posterior p(**z** | **x**) is intractable
- For training data **x** and parameters **z**, estimating the posterior p(**z** | **x**) is intractable



Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Problem:

– For observed variables **x** and latent variables **z**, estimating the posterior p(**z** | **x**) is intractable

– For training data **x** and parameters **z**, estimating the posterior p(**z** | **x**) is intractable

Solution:

– Approximate p(**z** | **x**) with a simpler q(**z**)

– Typically q(**z**) has more independence assumptions than p(**z** | **x**) – fine b/c q(**z**) is tuned for a specific **x**

– **Key idea:** pick a single q(**z**) from some family Q that best approximates p(**z** | **x**)

Narrative adapted from Jason Eisner's High-Level Explanation of VI: https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Terminology:

- $q(\mathbf{z})$: the **variational approximation**
- Q: the **variational family**
- Usually $q_\theta(\mathbf{z})$ is parameterized by some θ called **variational parameters**
- Usually $p_\alpha(\mathbf{z} \mid \mathbf{x})$ is parameterized by some fixed α – we'll call them the parameters

Example Algorithms:

- mean-field variational inference
- loopy belief propagation
- tree-reweighted belief propagation
- expectation propagation

Narrative adapted from Jason Eisner's High-Level Explanation of VI:
https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

Is this trivial?

- Note: We are not defining a new distribution simple $q_\theta(\mathbf{z} \mid \mathbf{x})$, there is one simple $q_\theta(\mathbf{z})$ for each $p_\alpha(\mathbf{z} \mid \mathbf{x})$
- Consider the MCMC equivalent of this:
  - you could draw samples $z^{(i)} \sim p(\mathbf{z} \mid \mathbf{x})$
  - then train some simple $q_\theta(\mathbf{z})$ on $z^{(1)}, z^{(2)}, \ldots, z^{(N)}$
  - hope that the sample adequately represents the posterior for the given $\mathbf{x}$
- How is VI different from this?
  - VI doesn't require sampling
  - VI is fast and deterministic
  - Why? b/c we choose an objective function (KL divergence) that defines which q_θ best approximates p_α, and exploit the special structure of q_θ to optimize it

Narrative adapted from Jason Eisner's High-Level Explanation of VI: https://www.cs.jhu.edu/~jason/tutorials/variational.html

Variational Inference

V.I. offers a new design decision

- Choose the distribution $p_\alpha(\mathbf{z} \mid \mathbf{x})$ that you really want, i.e. don't just simpify it to make it computationally convenient
- Then design a the structure of another distribution $q_\theta(\mathbf{z})$ such that V.I. is efficient

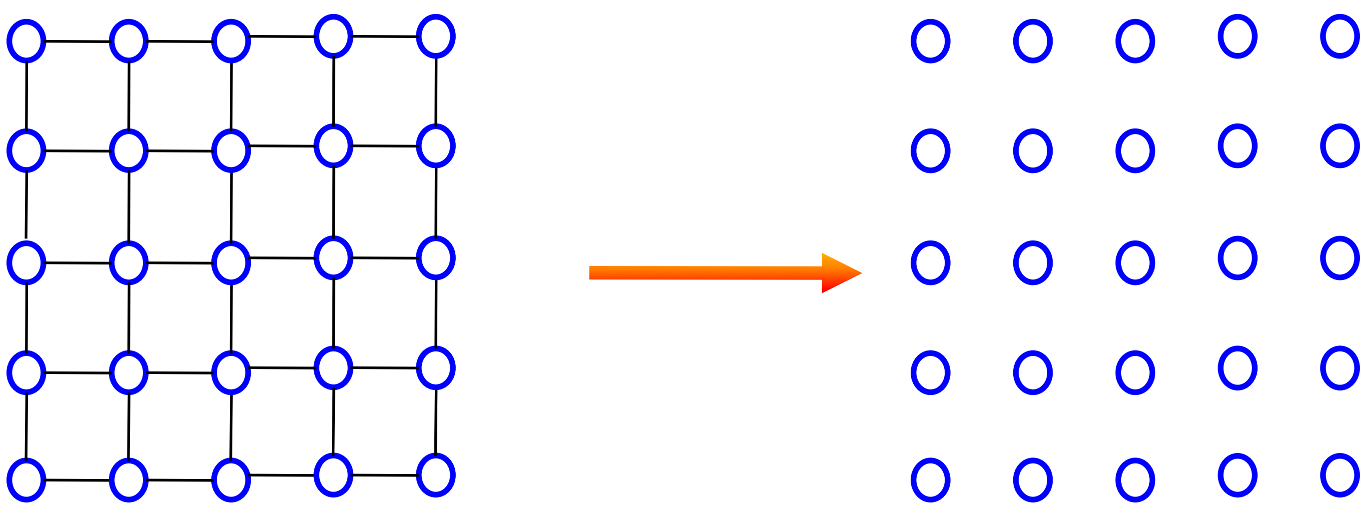
Narrative adapted from Jason Eisner's High-Level Explanation of VI: https://www.cs.jhu.edu/~jason/tutorials/variational.html

EXAMPLES OF VARIATIONAL APPROXIMATIONS

Mean Field for MRFs

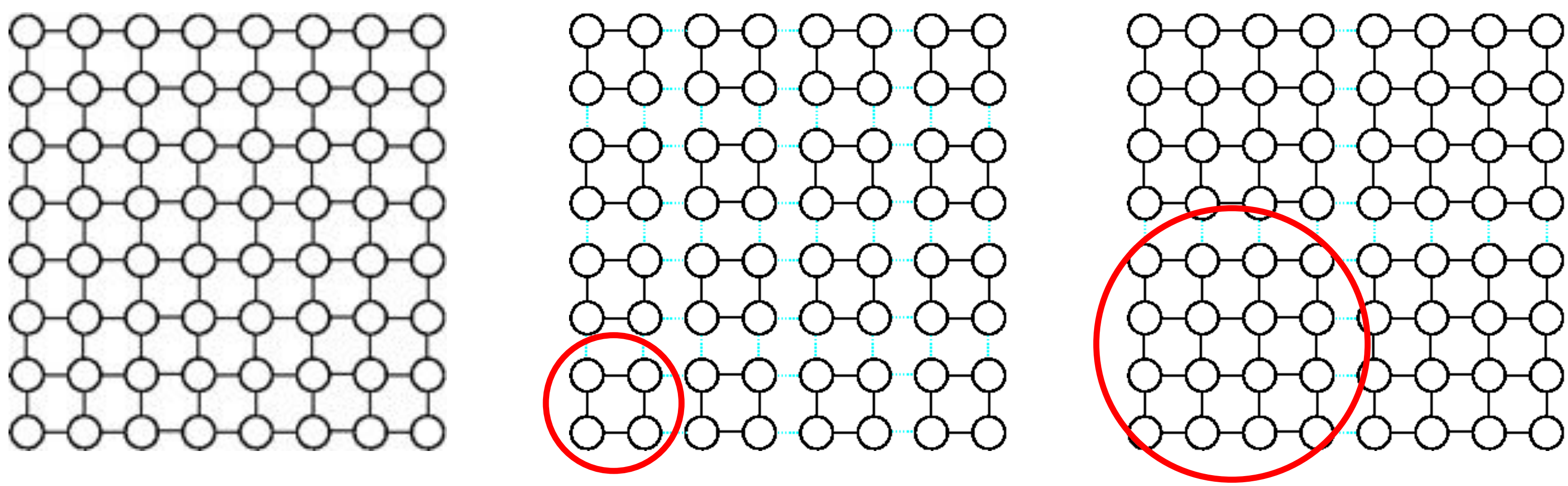
- Mean field approximation for Markov random field (such as the Ising model):

$$q(x) = \prod_{s \in V} q(x_s)$$



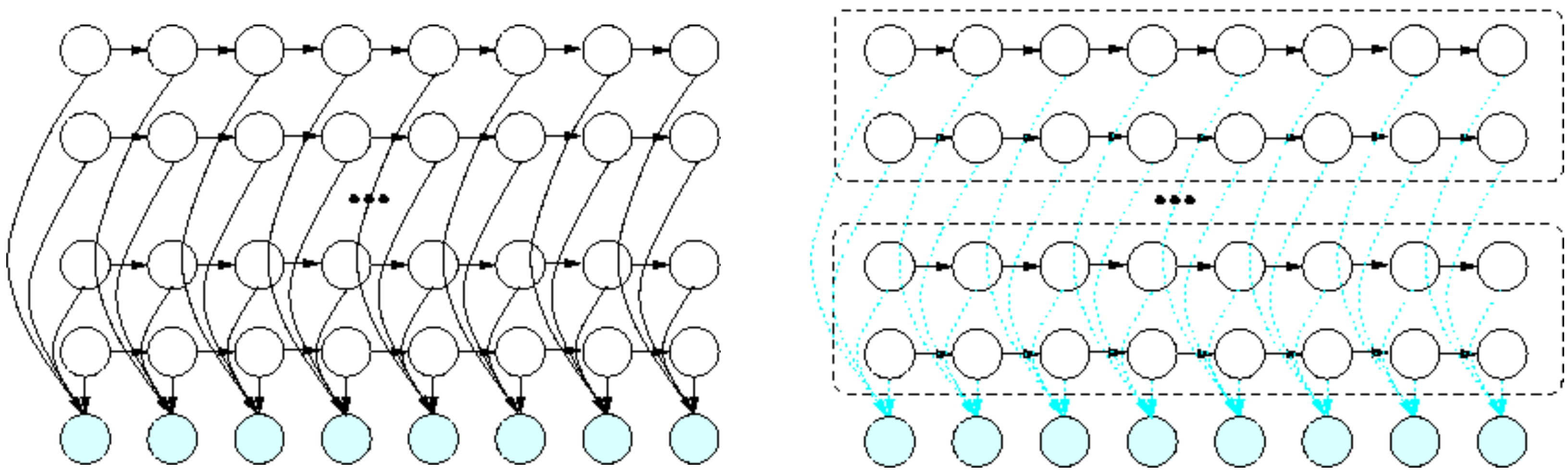
Variational Inference for MRFs

- We can also apply more general forms of mean field approximations (involving clusters) to the Ising model:
- Instead of making all latent variables independent (i.e. naïve mean field, previous figure), clusters of (disjoint) latent variables are independent.



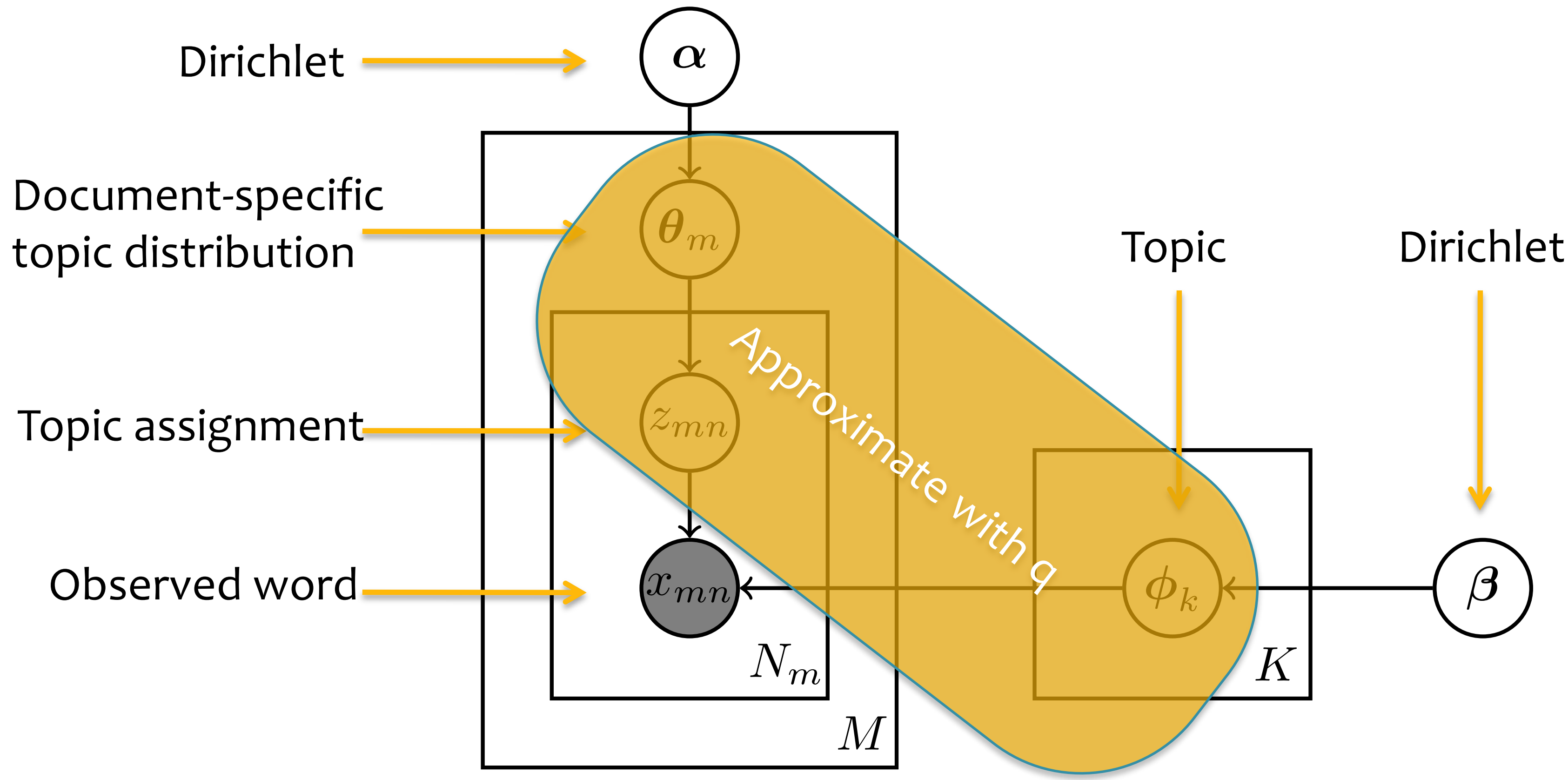
V.I. for Factorial HMM

- For a factorial HMM, we could decompose into chains



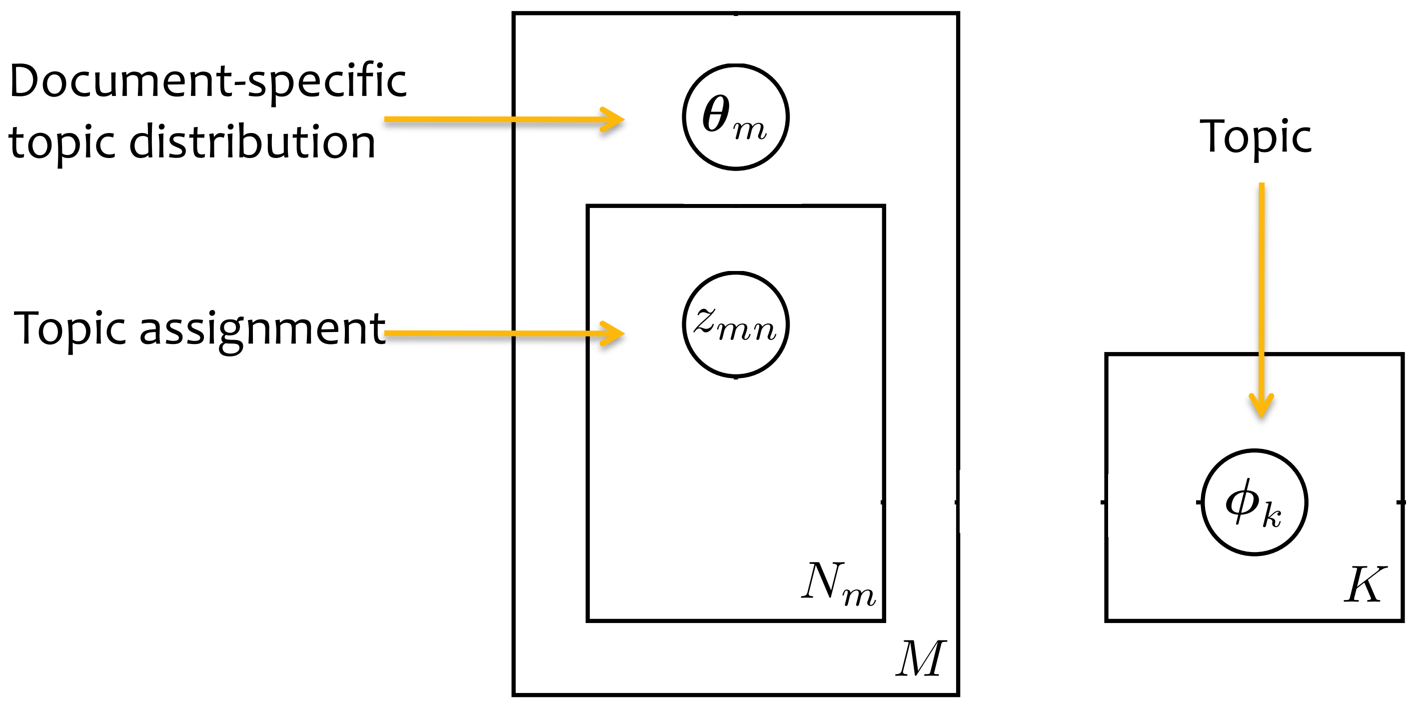
LDA Inference

- Explicit Variational Inference (original distribution)



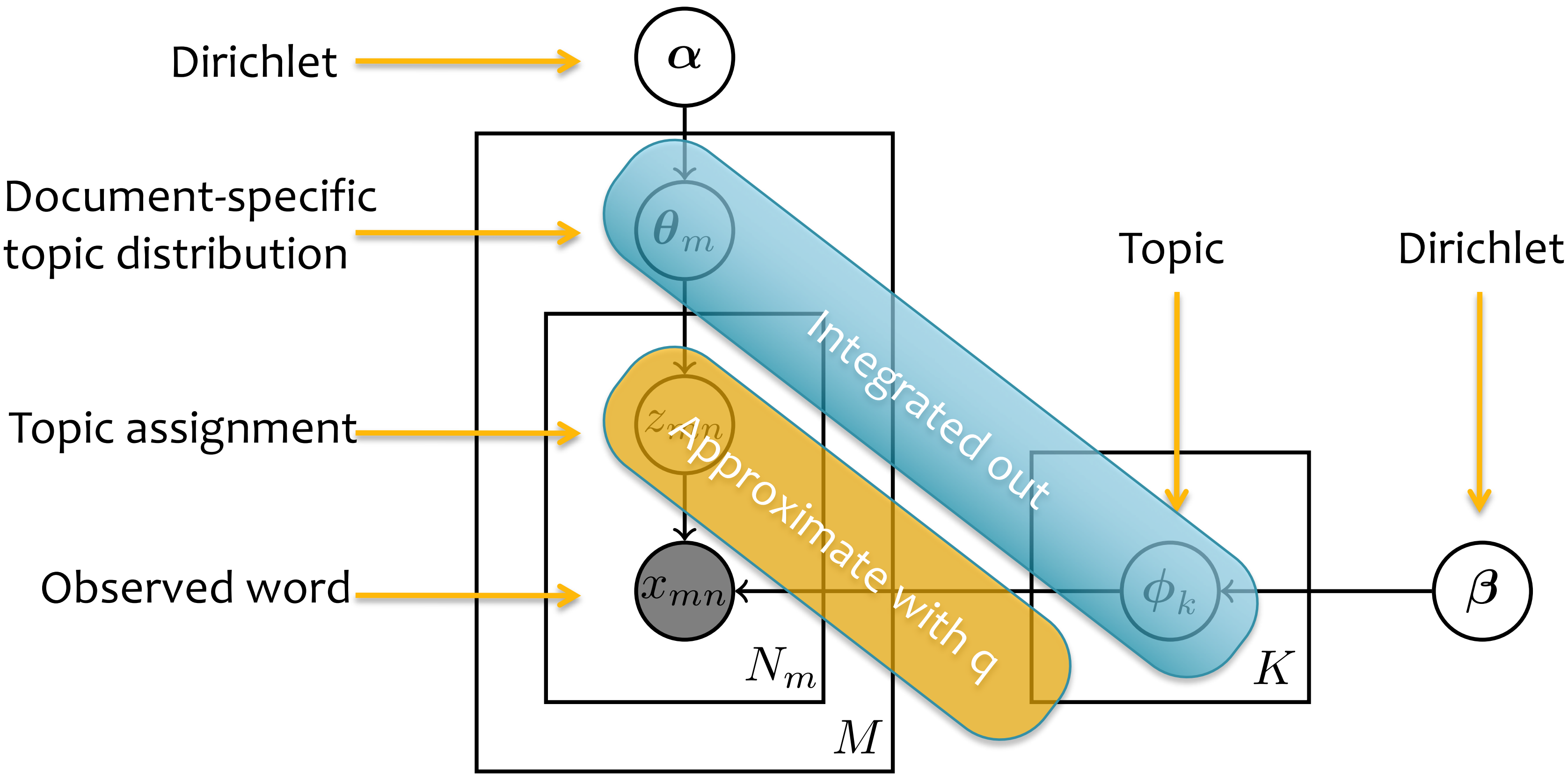
LDA Inference

- Explicit Variational Inference (variational approximation)



LDA Inference

- Collapsed Variational Inference



MEAN FIELD VARIATIONAL INFERENCE

KL Divergence

- *Definition*: for two distributions $q(x)$ and $p(x)$ over $x \in \mathcal{X}$, the **KL Divergence** is:

$$KL(q \,||\, p) = E_{q(x)}[\log q(x)/p(x)]$$

- *Properties*:
  - $KL(q \,||\, p)$ measures the **proximity** of two distributions q and p
  - KL is **not** symmetric: $KL(q \,||\, p) \neq KL(p \,||\, q)$
  - KL is minimized when $q(x) = p(x)$ for all $x \in \mathcal{X}$

Variational Inference

Whiteboard

- Background: KL Divergence
- Mean Field Variational Inference (overview)
- Evidence Lower Bound (ELBO)
- ELBO's relation to log p(x)