# 10707
# Deep Learning: Spring 2020

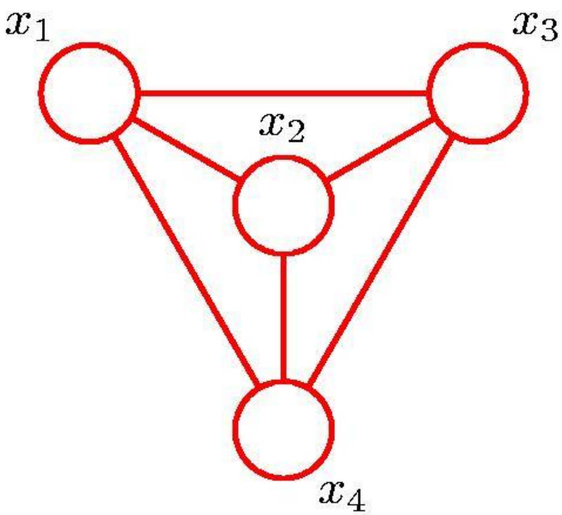
Andrej Risteski

Machine Learning Department

## **Lecture 11:**
## Variational methods

# Graphical Models

Recall: graph contains a set of nodes connected by edges.




In a probabilistic graphical model, each node represents a random variable, links represent "probabilistic dependencies" between random variables.

Graph specifies how joint distribution over all random variables decomposes into a **product** of factors, each factor depending on a subset of the variables.

Two types of graphical models:

- **Bayesian networks**, also known as Directed Graphical Models (the links have a particular directionality indicated by the arrows)
- **Markov Random Fields**, also known as Undirected Graphical Models (the links do not carry arrows and have no directional significance).

# Algorithmic pros/cons of latent-variable models (so far)

**RBM's**

- Hard to draw samples ✗

(In fact, #P-hard provably, even in Ising models)

- Easy to sample posterior distribution over latents ✓

**Directed models**

- Easy to draw samples ✓

- Hard to sample posterior distribution over latents ✗

(In fact, #P-hard even in mixtures)

# Canonical tasks with graphical models

## Inference

Given values for the parameters $\theta$ of the model, *sample/calculate* marginals (e.g. sample $p_\theta(x_1), p_\theta(x_4, x_5), p_\theta(z|x)$, etc.)

## Learning

Find values for the parameters $\theta$ of the model, that give a *high likelihood* for the observed data. (e.g. canonical way is solving maximum likelihood optimization

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p(x_i)$$

*Other methods exist, e.g. method of moments (matching moments of model), but less used in deep learning practice.*

# Canonical tasks with graphical models

**Inference**

Inference is hard in undirected fully-observable models (due to partition function); easy in fully-observable Bayesian nets.

It's easy for RBM's, hard for latent-variable Bayesian nets (again, implicit normalizing factor is hard.)

**Learning**

We will derive "iterative"/"incremental" learning algorithms, using inference algorithms as subroutines.

We will, in particular see how a technique called "variational methods" can be used. (Next time, we see how MCMC methods can be used.)

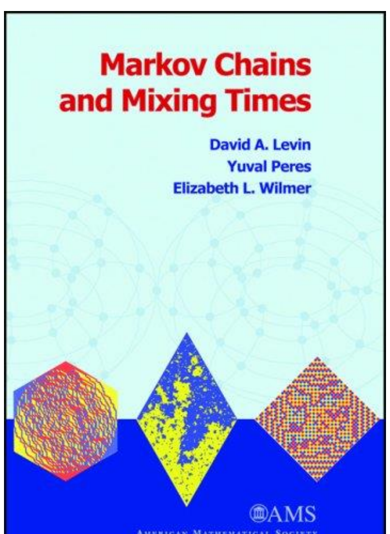
# Inference

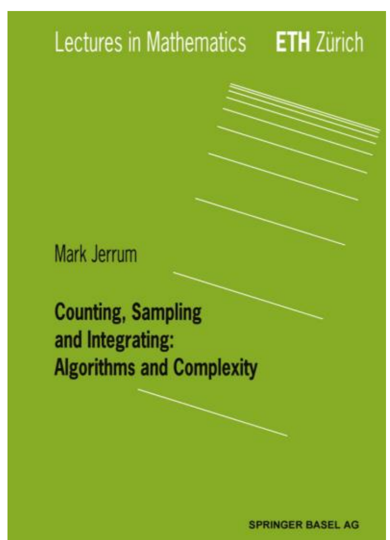
# Algorithmic approaches

When faced with a difficult to calculate probabilistic quantity (partition function, difficult posterior), there are two families of approaches:

**MARKOV CHAIN MONTE CARLO**

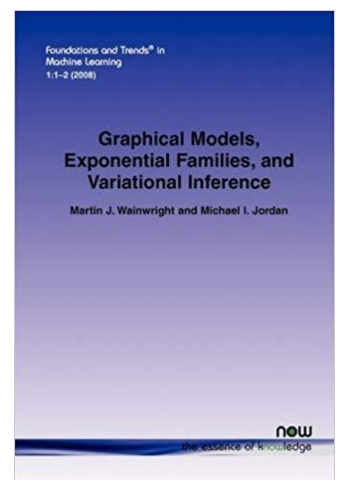
- Random walk w/ equilibrium distribution the one we are trying to sample from.
- Well studied in TCS.




**VARIATIONAL METHODS**

- Based on solving an optimization problem.
- Very popular in practice.
- Comparatively poorly understood



# Part I: Inference in undirected graphical models

Though not always true, calculating marginals is often reducible to calculating partition functions.

**Simple example**: Ising models

$$P_\theta(\mathbf{x}) = \frac{1}{\mathcal{Z}(\theta)} \exp\Big( \sum_{ij\in E} x_i x_j \theta_{ij} + \sum_{i\in V} x_i \theta_i \Big)$$

$$P_\theta(x_k = 1) = \frac{\overbrace{\sum_{\boldsymbol{x}_{-k} \in\{-1,1\}^{n-1}} \exp\big(\theta_k + \sum_{kj\in E} x_k \theta_j + \sum_{ij\in E, i\neq k, j\neq k} x_i x_j \theta_{ij} + \sum_{i\neq k} \theta_i x_i\big)}^{\textit{Partition fn of an appropriately modified Ising model}}}{\underbrace{\sum_{\boldsymbol{x}\in\{-1,1\}^{n}} \exp\big(\sum_{ij} x_i x_j \theta_{ij} + \sum_i \theta_i x_i\big)}_{\textit{Partition fn of original Ising model}}}$$

# Part I: Inference in undirected graphical models

Though not always true, calculating marginals is often reducible to calculating partition functions.

**Simple example**: Ising models

$$P_\theta(\mathbf{x}) = \frac{1}{\mathcal{Z}(\theta)} \exp\Big( \sum_{ij \in E} x_i x_j \theta_{ij} + \sum_{i \in V} x_i \theta_i \Big)$$

Formal term for when sampling reduces to calculating partition functions: *self-reducible problems*.

How do we calculate/approximate the partition function?

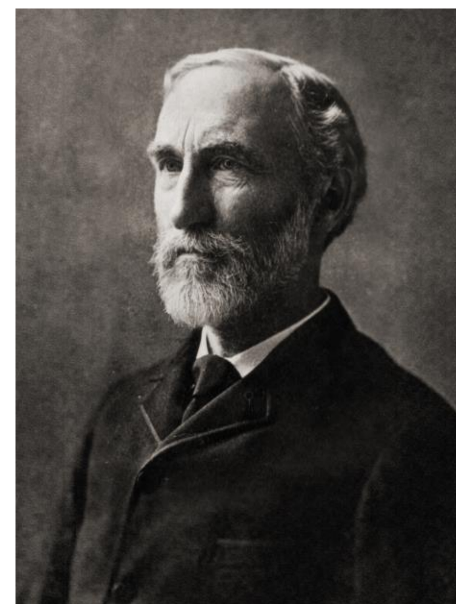
# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log\ Z = \max_{q:\text{ distribution over }\mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

*Find the distribution that has both high entropy, and high expected energy value*

$$H(q) \coloneqq -\sum_{x\in\mathcal{X}} q(x)\log q(x)$$



# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

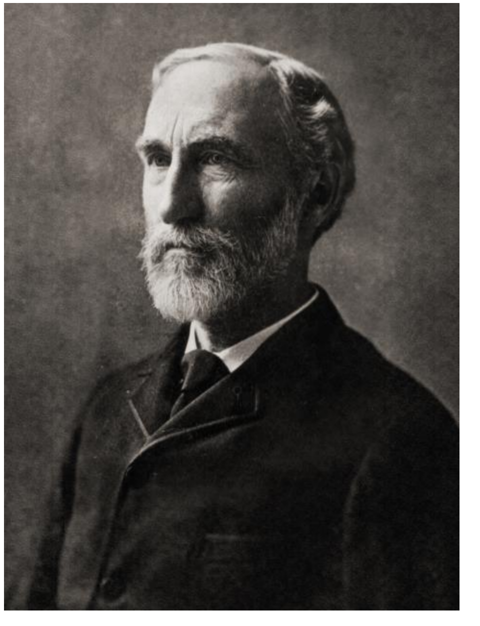
$$\log Z = \max_{q:\ \text{distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

*Proof:* $0 \leq KL(q\,||\,p) = \mathbb{E}_q \log q - \mathbb{E}_q \log p$

$$= -H(q) - \mathbb{E}_{x\sim q}[E(x)] + \log Z$$

$$H(q) + \mathbb{E}_{x\sim q}[E(x)] \leq \log Z$$

Hence, $\max_{q:\ \text{distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)] \leq \log Z$



# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q:\ \text{distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

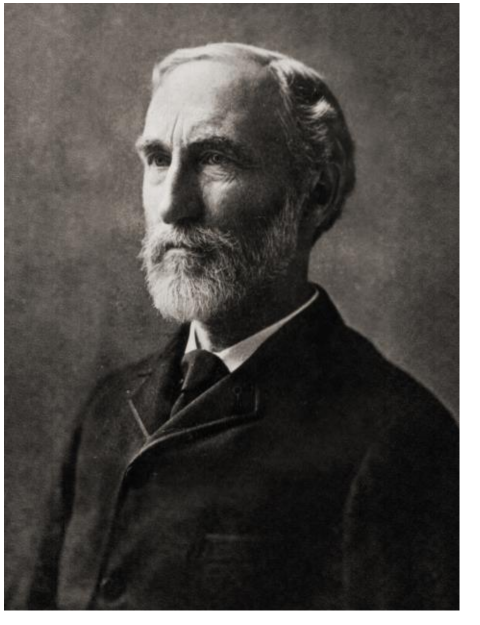
*Proof:* $0 \leq KL(q\,||\,p) = \mathbb{E}_q \log q - \mathbb{E}_q \log p$

$$= -H(q) - \mathbb{E}_{x\sim q}[E(x)] + \log Z$$

$$H(q) + \mathbb{E}_{x\sim q}[E(x)] \leq \log Z$$

Equality is attained if p = q: $KL(q\,||\,p)$=0, so

$$H(q) + \mathbb{E}_{x\sim q}[E(x)] = \log Z$$



# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q\text{: distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

Hence, we've reduced calculating partition function to an optimization problem!

But, there is a **serious issue**: how do we solve an optimization over the set of distributions over $\mathcal{X}$?

Even if $\mathcal{X}$ is a really simple domain, e.g. $\mathcal{X} = \{\pm 1\}^n$, the trivial way to solve the problem would involve introducing a variable $q(x), \forall x \in \{\pm 1\}^n$: there are $2^n$ of them.

*In fact, you can't be clever – there are results showing this can be #P hard even for Ising models!*

# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q\text{: distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

What can we do to try to approximate this expression?

**Inspiration from physics**: solve a *simpler* optimization problem over a *restricted class* of distributions we can explicitly parametrize.

**Typical example:** mean-field approximation

Consider again $\mathcal{X} = \{\pm 1\}^n$. A *product distribution* depends on n parameters only: since $p(x) = \Pi_i p_i(x_i)$, for each $i \in [n]$, we only need to specify $p_i(x_i = 1)$.

# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q:\ \text{distribution over}\ \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

**Typical example:** mean-field approximation

Consider again $\mathcal{X} = \{\pm 1\}^n$. A *product distribution* depends on n parameters only: since $\mathrm{q}(x) = \Pi_i q_i(x_i)$, for each $i \in [n]$, we only need to specify $q_i(x_i) = 1$.

Then, we will solve the optimization problem: $\max_{q=\Pi_i q_i} H(q) + \mathbb{E}_{x\sim q}[E(x)]$

It's clear that the number of parameters is small at least.

If we can take gradient wrt variables $q_i(x_i)$ we can at least do *gradient descent*. Objective in general is *non-convex* though, so technically, this **can fail !!** (But often works ok. )

# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q:\ \text{distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

**Typical example:** mean-field approximation: $\max_{q=\Pi_i q_i} H(q) + \mathbb{E}_{x\sim q}[E(x)]$

Can we take gradients?

(1) *Entropy factorizes*: $H(q_1 q_2) = \sum_{x_1,x_2\in\{\pm1\}} q_1(x_1)q_2(x_2)\log(q_1(x_1)q_2(x_2))$

$$= \sum_{x_1,x_2} q_1(x_1)q_2(x_2)\left(\log q_1(x_1) + \log q_2(x_2)\right) = \sum_{x_2} q_2(x_2)H(q_1) + \sum_{x_1} q_1(x_1)H(q_2)$$

$$= H(q_1) + H(q_2) = q_1(x_1)\log q_1(x_1) + (1 - q_1(x_1))\log\big(1 - q_1(x_1)\big) + H(q_2)$$

# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q:\text{ distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

**Typical example:** mean-field approximation: $\max_{q=\Pi_i q_i} H(q) + \mathbb{E}_{x\sim q}[E(x)]$

Can we solve this?

(2) $\mathbb{E}_{x\sim q}[E(x)]$ is often not a problem:

e.g. for Ising models, $E(x) = \sum_{ij} J_{ij} x_i x_j$ so $\mathbb{E}_q[E(x)] = \sum_{ij} J_{ij} \mathbb{E}[x_i]\mathbb{E}[x_j]$

$\mathbb{E}\big[x_j\big] = q_i(x_i) - \big(1 - q_i(x_i)\big)$, so taking a gradient is simple.

# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q:\ \text{distribution over}\ \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

**Typical example:** mean-field approximation: $\max_{q=\Pi_i q_i} H(q) + \mathbb{E}_{x\sim q}[E(x)]$

Can we solve this?

(2) $\mathbb{E}_{x\sim q}[E(x)]$ is often not a problem:

Even if gradients are not explicit, if $E(x)$ is a sum of terms $\phi_S(x_S)$, we can estimate the expectations, and estimate the gradient from that.
(i.e. use zeroth order optimization method.)

# Variational methods for partition functions

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q:\text{ distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

**Typical example:** Gaussians with mean/covariance mx as parameters.

(1) Entropy of Gaussian has a closed-form formula:

$$\frac{1}{2}\log((2\pi e)^n \det \Sigma)$$

(2) Expectations wrt to Gaussian can be (often) estimated by drawing samples.

# Part II: Inference in latent-variable Bayesian networks

As we noted last time, the difficulty for calculating posteriors in latent-variable models is of a *similar nature*:

$$p(z|x) = \frac{p(z)p(x|z)}{p(x)}$$

*Simple, closed-form*

*Integrates over all possible z's, hard normalizing factor*

Taking inspiration from previous part, we will approximate posteriors $p(z|x)$ in an analogous way:

# Variational methods for posterior distributions

Let $p(z,x)$ be a joint distribution over latent variables and observables.
Then, same as in the ***Gibbs principle*** calculation:

$$KL(q\ (z|x)||\ p(z|x)) = \mathbb{E}_{q\ (z|x)} \log q\ (z|x) - \ \mathbb{E}_{q\ (z|x)} \log p\ (z|x)$$
$$= -H(q\ (z|x)) - \mathbb{E}_{z\sim q}[p(z,x)] + \log p(x)$$

Hence,

$$\text{argmin}_{q(z|x)}\ KL(q\ (z|x)||\ p(z|x)\big) = \text{argmin}\{\ \mathbb{E}_{q\ (z|x)} \log q\ (z|x) - \ \mathbb{E}_{q\ (z|x)} \log p\ (z,x)\}$$

As in the undirected case, if q is a simple distribution (e.g. product distribution, Gaussian), we can optimize this using gradient descent.

# Yet another mean field strategy: coordinate ascent

Consider updating a *single* coordinate of the mean-field distribution, that is keep $q_{-i}\ (z_i|x)$ fixed, and optimize for $q_i\ (z_i|x)$. Rewriting, we have:

$$KL(q\ (z|x)||\ p(z|x))$$

$$= \mathbb{E}_{q\ (z|x)} \log q\ (z|x) - \mathbb{E}_{q\ (z|x)} \log p\ (z,x)$$

$$= \sum_i \mathbb{E}_{q_i\ (z_i|x)} \log q_i\ (z_i|x) - \mathbb{E}_{q_i\ (z_i|x)} \left[ \mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\ (z_i, z_{-i}, x) \right]$$

$$= \mathbb{E}_{q_i\ (z_i|x)} \log q_i\ (z_i|x) - \mathbb{E}_{q_i\ (z_i|x)} [\log \tilde{p}\ (z_i, x)] + C$$

*Renormalize to make it a distribution*

# Yet another mean field strategy: coordinate ascent

Consider updating a *single* coordinate of the mean-field distribution, that is keep $q_{-i}\ (z_i|x)$ fixed, and optimize for $q_i\ (z_i|x)$. Rewriting, we have:

$$KL(q\ (z|x)||\ p(z|x))$$

$$= \mathbb{E}_{q\ (z|x)} \log q\ (z|x) - \mathbb{E}_{q\ (z|x)} \log p\ (z, x)$$

$$= \sum_i \mathbb{E}_{q_i\ (z_i|x)} \log q_i\ (z_i|x) - \mathbb{E}_{q_i\ (z_i|x)} \left[\mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\ (z_i, z_{-i}, x)\right]$$

$$= \mathbb{E}_{q_i\ (z_i|x)} \log q_i\ (z_i|x) - \mathbb{E}_{q_i\ (z_i|x)}[\log \tilde{p}\ (z_i, x)] + C$$

$$= KL(q_i\ (z_i|x)||\ \tilde{p}(z_i, x)) + C$$

Optimum is $q_i\ (z_i|x) = \tilde{p}(z_i, x)$

$$= \frac{\mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\ (z_i, z_{-i}, x)}{\int_{z_i} \mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\ (z_i, z_{-i}, x)}$$

# Yet another mean field strategy: coordinate ascent

Consider updating a *single* coordinate of the mean-field distribution, that is keep $q_{-i}\,(z_i|x)$ fixed, and optimize for $q_i\,(z_i|x)$. Rewriting, we have:

$$KL(q\,(z|x)||\,p(z|x))$$

Iterate!

$$= \mathbb{E}_{q\,(z|x)} \log q\,(z|x) - \mathbb{E}_{q\,(z|x)} \log p\,(z,x)$$

$$= \sum_i \mathbb{E}_{q_i\,(z_i|x)} \log q_i\,(z_i|x) - \mathbb{E}_{q_i\,(z_i|x)} \left[\mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\,(z_i, z_{-i}, x)\right]$$

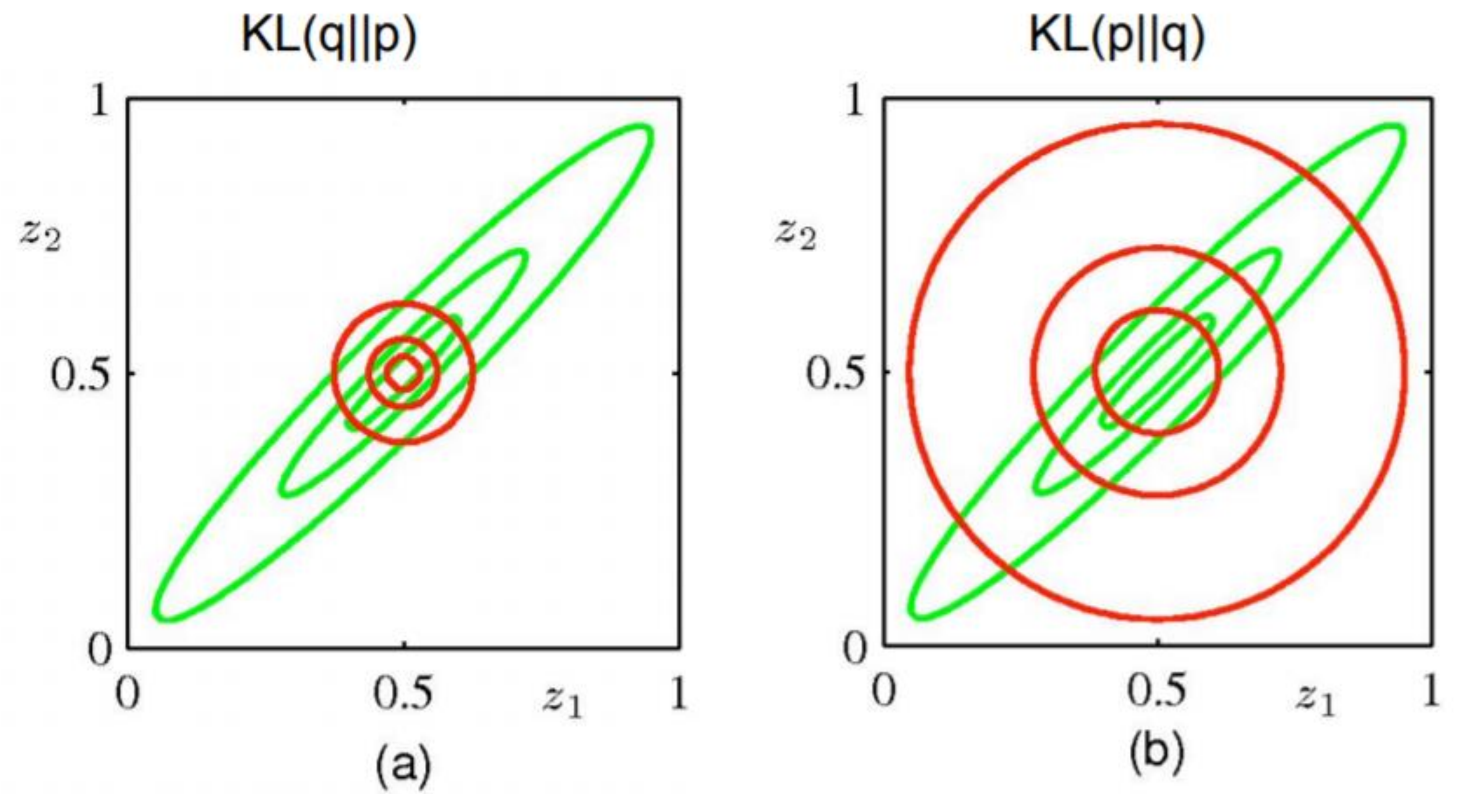
$$= \mathbb{E}_{q_i\,(z_i|x)} \log q_i\,(z_i|x) - \mathbb{E}_{q_i\,(z_i|x)}[\log \tilde{p}\,(z_i,x)] + C$$

$$= KL(q_i\,(z_i|x)||\,\tilde{p}(z_i,x)) + C$$

Optimum is $q_i\,(z_i|x) = \tilde{p}(z_i,x)$

$$= \frac{\mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\,(z_i, z_{-i}, x)}{\int_{z_i} \mathbb{E}_{q_{-i}(z_{-i}|x)} \log p\,(z_i, z_{-i}, x)}$$

# What if we changed the order of p, q?




Approximation is too compact.

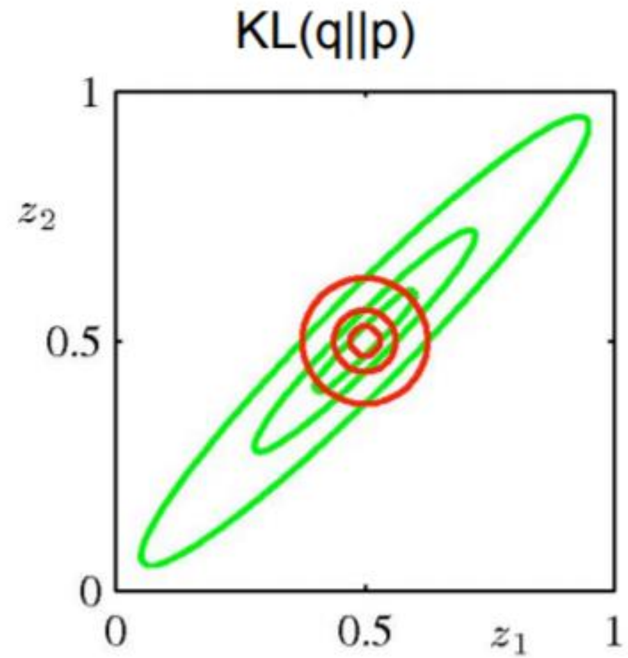
Approximation is too spread.

# What if we changed the order of p, q?

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \frac{p(\mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

There is a large positive contribution to the KL divergence from regions of Z space in which:
- p(Z) is near zero
- unless q(Z) is also close to zero.

Minimizing KL(q||p) leads to distributions q(Z) that avoid regions in which p(Z) is small.



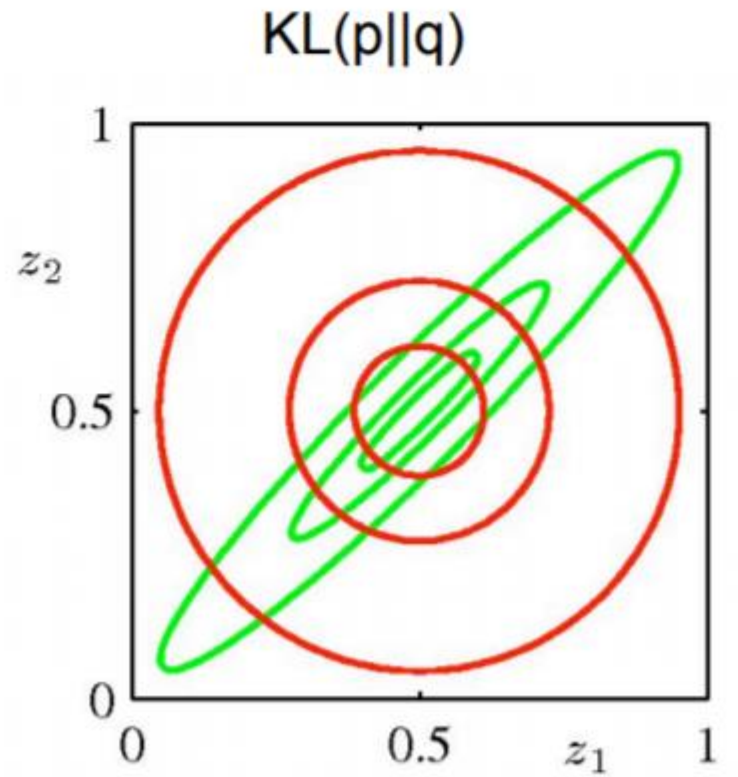
(a)

# What if we changed the order of p, q?

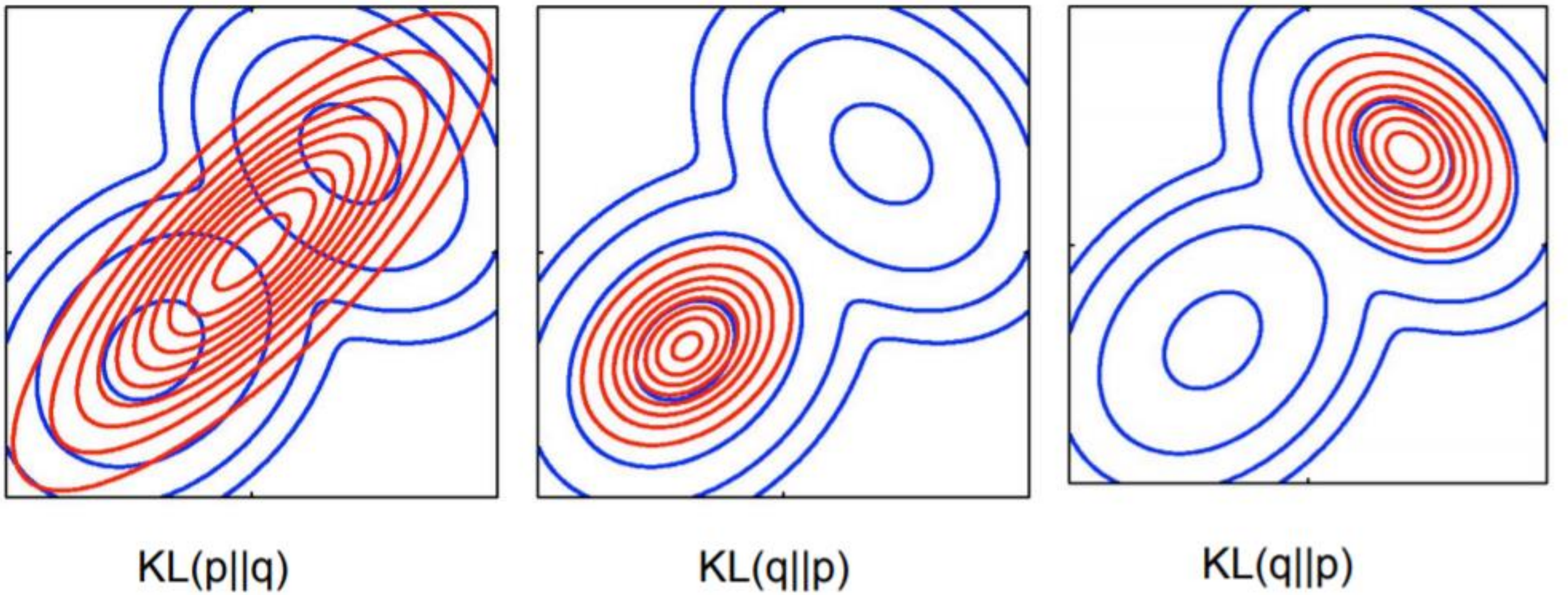
$$\mathrm{KL}(p\|q) = -\int p(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z}.$$

There is a large positive contribution to the KL divergence from regions of Z space in which:
- q(Z) is near zero,
- unless p(Z) is also close to zero.

Minimizing KL(p||q) leads to distributions q(Z) that are nonzero in regions where p(Z) is nonzero.



# Multimodal distributions




Blue contours show bimodal distribution, red contours single Gaussian distribution that best approximates it.

KL(q||p) will tend to find a single mode, whereas KL(p||q) will average across all of the modes.

# Learning

# Learning latent-variable directed graphical models

How should we try to learn the parameters of a graphical model?

The most obvious strategy: maximum likelihood estimation

Given data $x_1, x_2, \ldots, x_n$, solve the optimization problem

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p(x_i)$$

Latent variables: we will use the variational principle again!

$$\log_\theta \; p(x) = \max_{q(z|x):\text{ distribution over } Z} H(q(z|x)) + \mathbb{E}_{q(z|x)}[\log p_\theta(x,z)]$$

Hence, MLE objective can be written as double maximization:

$$\max_{\theta \in \Theta} \max_{\{q_i(z|x_i)\}} \sum_{i=1}^{n} H(q_i(z|x_i)) + \mathbb{E}_{q_i(z|x_i)}[\log p_\theta(x_i,z)]$$

# Variational methods for posterior distributions

**ELBO (Evidence Lower Bound)**: Let $p(z,x)$ be a joint distribution over latent variables and observables. Then:

$$\log\ p(x) = \max_{q(z|x):\text{ distribution over } \mathcal{Z}} H(q|z) + \mathbb{E}_{q(z|x)}[\log p(x,z)]$$

Write, by Bayes rule, $p(z|x) = \frac{p(x,z)}{p(x)}$. Then, the formula above follows by Gibbs variational principle with $\mathrm{E(x)} = \log p(x,z)$. **Argmax** = p(z|x)!

**Gibbs variational principle**: Let $p(x) = \frac{1}{Z}\exp\big(E(x)\big)$ be a distribution over a domain $\mathcal{X}$. Then, Z is the solution to the following optimization problem:

$$\log Z = \max_{q:\text{ distribution over } \mathcal{X}} H(q) + \mathbb{E}_{x\sim q}[E(x)]$$

# Expectation-maximization/ variational inference

The canonical algorithm for learning a single-layer latent-variable Bayesian network is an iterative algorithm as follows.

Consider the max-likelihood objective, rewritten as in the previous slide:

$$\max_{\theta \in \Theta} \max_{\{q_i(z|x_i)\}} \sum_{i=1}^{n} H(q_i(z|x_i)) + \mathbb{E}_{q_i(z|x_i)}[\log p_\theta(x_i, z)]$$

Algorithm maintains iterates $\theta^t, \{q_i^t(z|x_i)\}$, and updates them iteratively

(1) **Expectation (E)-step**:

Keep $\theta^t$ fixed, set $\{q_i^{t+1}(z|x_i)\}$, s.t. they maximize the objective above.

(2) **Maximization (M)-step**:

Keep $\{q_i^t(z|x_i)\}$ fixed, set $\theta^{t+1}$ s.t. it maximizes the objective above.

Clearly, every step cannot make the objective worse!

Does *not* mean it converges to global optimum – could, e.g. get stuck in a local minimum.

# Expectation-maximization/ variational inference

The canonical algorithm for learning a single-layer latent-variable Bayesian network is an iterative algorithm as follows.

Consider the max-likelihood objective, rewritten as in the previous slide:

$$\max_{\theta\in\Theta} \max_{q_i(z|x_i)} \sum_{i=1}^{n} H(q_i(z|x_i)) + \mathbb{E}_{q_i(z|x_i)}[\log p_\theta(x_i, z)]$$

Algorithm maintains iterates $\theta^t, q_i^t(z|x_i)$, and updates them iteratively

(1) **Expectation step**:

Keep $\theta^t$ and set $q_i^{t+1}(z|x_i)$, s.t. they maximize the objective above.

If the class is infinitely rich, the optimum is $q_i^{t+1}(z|x_i) = p_{\theta^t}(z|x_i)$

This is called **expectation-maximization (EM)**.
If class is not infinitely rich, it's called **variational inference**.

# Examples

1: Mixtures of spherical Gaussians

Consider a mixture of K Gaussians with unknown means $p = \sum_{i=1}^{K} \frac{1}{K} \mathcal{N}(\mu_i, I_d)$

Let's try to calculate the E and M steps.

**E-step**: the optimal $q_i^{t+1}(z|x_i)$ is $p_{\theta^t}(z|x_i)$. Can we calculate this?

By Bayes rule, $p_{\theta^t}(z = k|x_i) \propto p(x_i|z = k) \propto e^{-\left|\left|x_i - \mu_k^t\right|\right|^2}$

Writing out the normalizing constant, we have

$$p_{\theta^t}(z = k|x_i) = \frac{e^{-\left|\left|x_i - \mu_k^t\right|\right|^2}}{\sum_{k'} e^{-\left|\left|x_i - \mu_{k'}^t\right|\right|^2}}$$

*"Soft" version of assigning point to nearest cluster*

# Examples

<u>1: Mixtures of spherical Gaussians</u>

Consider a mixture of K Gaussians with unknown means $p = \sum_{i=1}^{K} \frac{1}{K} \mathcal{N}(\mu_i, I_d)$

Let's try to calculate the E and M steps.

**M-step**: given a quess $q_i^t(z|x_i)$ , we can rewrite the maximization for $\theta$ as:

$$\max_{\theta \in \Theta} \sum_{i=1}^{n} H\big(q_i^t(z|x_i)\big) + \mathbb{E}_{q_i^t(z|x_i)}[\log p_\theta(x_i, z)]$$

$$= \mathbb{E}_{q_i^t(z|x_i)}\,[\cancel{\log p(z)} + \log p_\theta(x|z)]$$

*Doesn't depend on $\theta$*

$$\mathbb{E}_{q_i^t(z|x_i)}[\ \log p_\theta(x|z)]$$

# Examples

<u>1: Mixtures of spherical Gaussians</u>

Consider a mixture of K Gaussians with unknown means $p = \sum_{i=1}^{K} \frac{1}{K} \mathcal{N}(\mu_i, I_d)$

Let's try to calculate the E and M steps.

**M-step**: given a quess $q_i^t(z|x_i)$ , we can rewrite the maximization for $\theta$ as:

$$\max_{\theta} \ \mathbb{E}_{q_i^t(z|x_i)}[\ \log p_\theta(x|z)] = \max_{\theta} \ -\sum_{i=1}^{n} \sum_{k=1}^{K} q_i^t(z=k|x_i)||x_i - \mu_k\,||^2$$

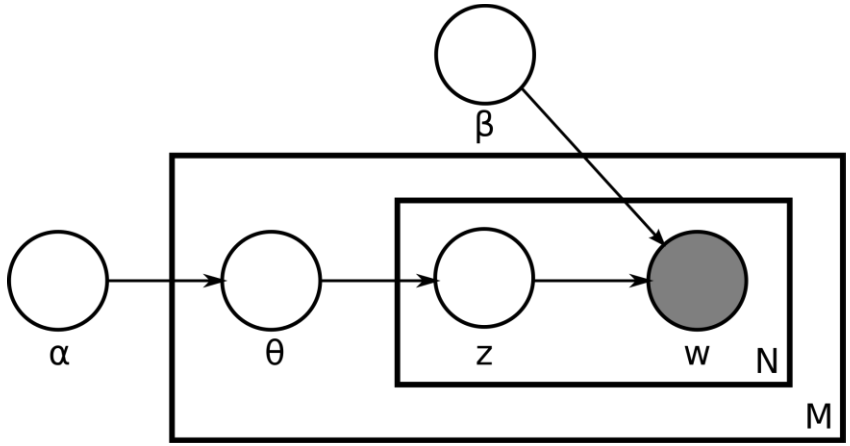
Setting the derivative wrt to $\mu_k$ to 0, we have:

$$\mu_k^{t+1} = \sum_{i=1}^{n} \frac{e^{-\left|\left|x_i - \mu_k^t\right|\right|^2}}{\sum_{k'} e^{-\left|\left|x_i - \mu_{k'}^t\right|\right|^2}} x_i$$

# Examples

2: Latent Dirichlet Allocation

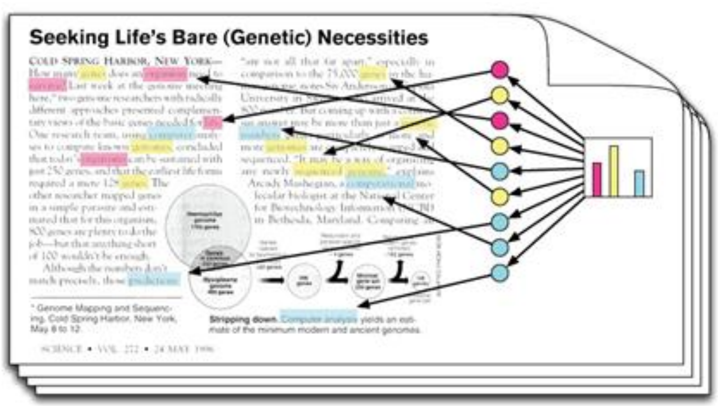
The **parameters** are: $\{\alpha_i\}_{i=1}^{K}$ (Dirichlet parameters) and matrix $\beta \in \mathbb{R}_+^{N\times K}$, where N is the size of the vocabulary.




The columns of $\beta$ satisfy $\sum_{j=1}^{N} \beta_{ij} = 1$ (the distribution of words in a topic i)

To produce document:

- First, sample $\theta \sim \mathrm{Dir}(\cdot\,|\alpha)$: this will be the topic proportion vector for the document.
- Each word in the document is generated in order, independently.
- To generate word i:
  - Sample topic $z_i$ with categorical distribution with parameters $\theta$
  - Sample word $w_i$ with categorical distribution with parameters $\beta_{z_i}$



# Examples

The E-step cannot be done in closed form:

$$p(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K} \mid w_{1:D,1:N}, \alpha, \eta) = \frac{p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}{\int_{\vec{\beta}_{1:K}} \int_{\vec{\theta}_{1:D}} \sum_{\vec{z}} p(\vec{\theta}_{1:D}, \vec{z}_{1:D}, \vec{\beta}_{1:K} \mid \vec{w}_{1:D}, \alpha, \eta)}$$

(In fact, can be shown to be #P-hard to perform in the worst case.)

The variational family to approximate the posterior is commonly chosen to be a mean-field family:

$$q(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K}) = \prod_{k=1}^{K} q(\vec{\beta}_k \mid \vec{\lambda}_k) \prod_{d=1}^{D} \left( q(\vec{\theta}_{d d} \mid \vec{\gamma}_d) \prod_{n=1}^{N} q(z_{d,n} \mid \vec{\phi}_{d,n}) \right)$$

- **Probability of topic** $z$ **given document** $d$: $q(\theta_d \mid \gamma_d)$
  Each document has its own Dirichlet prior $\gamma_d$
- **Probability of word** $w$ **given topic** $z$: $q(\beta_z \mid \lambda_z)$
  Each topic has its own Dirichlet prior $\lambda_z$
- **Probability of topic assignment to word** $w_{d,n}$: $q(z_{d,n} \mid \varphi_{d,n})$
  Each word position *word[d][n]* has its own prior $\varphi_{d,n}$

# Examples

$$q(\vec{\theta}_{1:D}, z_{1:D,1:N}, \vec{\beta}_{1:K}) = \prod_{k=1}^{K} q(\vec{\beta}_k \mid \vec{\lambda}_k) \prod_{d=1}^{D} \left( q(\vec{\theta}_{d\,d} \mid \vec{\gamma}_d) \prod_{n=1}^{N} q(z_{d,n} \mid \vec{\phi}_{d,n}) \right)$$

- **Probability of topic** $z$ **given document** $d$: $q(\theta_d \mid \gamma_d)$
  Each document has its own Dirichlet prior $\gamma_d$
- **Probability of word** $w$ **given topic** $z$: $q(\beta_z \mid \lambda_z)$
  Each topic has its own Dirichlet prior $\lambda_z$
- **Probability of topic assignment to word** $w_{d,n}$: $q(z_{d,n} \mid \varphi_{d,n})$
  Each word position *word[d][n]* has its own prior $\varphi_{d,n}$

**One iteration of mean field variational inference for LDA**

(1) For each topic $k$ and term $v$:

(8) $$\lambda_{k,v}^{(t+1)} = \eta + \sum_{d=1}^{D} \sum_{n=1}^{N} 1(w_{d,n} = v)\phi_{n,k}^{(t)}.$$

(2) For each document $d$:

(a) Update $\gamma_d$:

(9) $$\gamma_{d,k}^{(t+1)} = \alpha_k + \sum_{n=1}^{N} \phi_{d,n,k}^{(t)}.$$

(b) For each word $n$, update $\vec{\phi}_{d,n}$:

(10) $$\phi_{d,n,k}^{(t+1)} \propto \exp\left\{ \Psi(\gamma_{d,k}^{(t+1)}) + \Psi(\lambda_{k,w_n}^{(t+1)}) - \Psi(\textstyle\sum_{v=1}^{V} \lambda_{k,v}^{(t+1)}) \right\},$$

where $\Psi$ is the digamma function, the first derivative of the $\log \Gamma$ function.

Parameter updates:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}.$$