

Variational Inference

Michael Thomas Wojnowicz

November 11, 2020

Data Intensive Studies Center, Tufts University

Table of contents

1. Variational Inference: Overview
2. Variational Inference: Basics
3. Variational Inference and Expectation Maximization
4. Coordinate Ascent Variational Inference (CAVI)
5. Example: Bayesian Gaussian Mixture Model
6. Example: Bayesian Hidden Markov Model
7. Example: Latent Dirichlet Allocation (LDA)

Variational Inference: Overview

Some questions

- What is variational inference?
- When is it useful?
- Is it the same as variational bayes?
- Why is it called *variational* inference?
- What is Variational Expectation Maximization (VEM)? Variational Bayes Expectation Maximization (VBEM)?
- How can we apply VI to inference problems?

Approximate Bayesian Inference

- The two most prominent strategies for approximating intractable posteriors are VI and Markov Chain Monte Carlo (MCMC).
- MCMC uses **sampling**. We construct a Markov chain over model parameters. The stationary distribution is the posterior. We approximate the posterior with samples.
- VI uses **approximation**. A tractable approximating family is chosen, and parameters are optimized to be close to the posterior.

Variational Inference vs MCMC

Variational Inference scales better to large datasets.

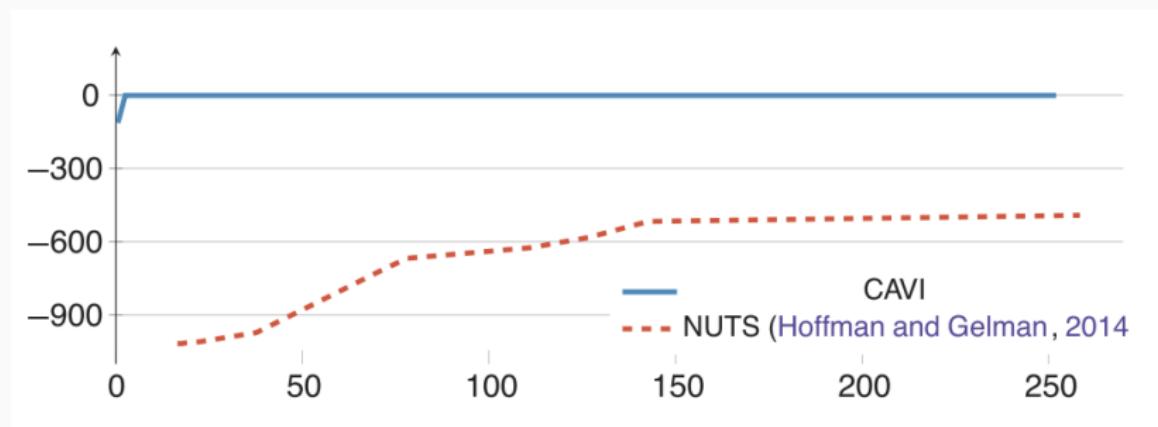
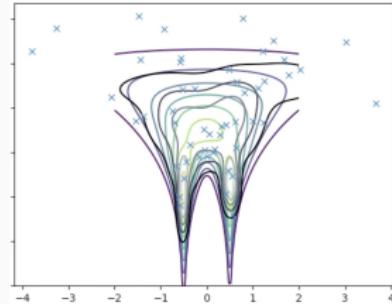
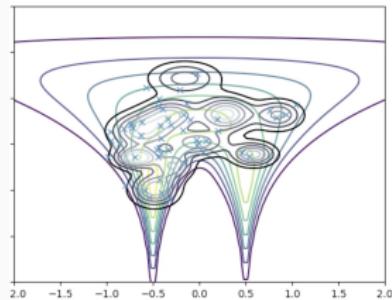
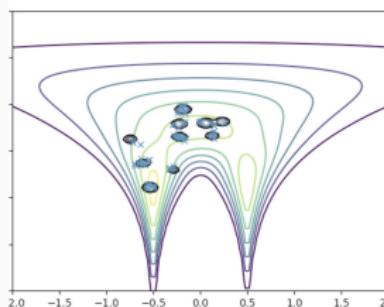
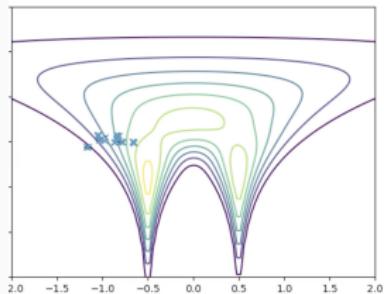


Figure 1: Comparison of CAVI to a Hamiltonian Monte Carlo-based sampling technique. The plot shows log predictive test set accuracy by training time (minutes). CAVI fits a Gaussian mixture model to 10,000 images in less than a minute.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

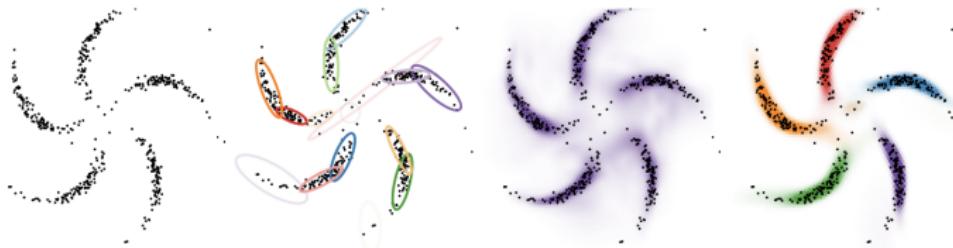
Illustration

Here we approximate an probability distribution by finding the best approximation from tractable family $\mathcal{Q} = \{10\text{-component Gaussian mixture models}\}$



Modern application

We can compose probabilistic graphical models with neural networks to exploit their complementary strengths.



The resulting model is expressive, but also interpretable/decomposable.

Variational Inference: Basics

Parametric statistical models

Parametric statistical models

A *parametric statistical model* posits

- x : observed data
- θ : parameters
- z (possibly): latent random variables

Parameters vs. latent variables

Both z and θ are unobserved, but only the dimensionality of z increases with the number of samples in x .

Frequentist vs. Bayesian variants

Frequentists take parameters θ to be fixed (but unknown) constants, whereas the Bayesians take θ to be random variables.

Three statistical modeling paradigms of interest

Variational inference can be useful for:

1. *Bayesian latent variable models*

- e.g. Bayesian Mixture Model, Bayesian Hidden Markov Model, Latent Dirichlet Allocation, Bayesian Nonparametric versions of the preceding

2. *Frequentist latent variable models*

- Variational Autoencoders (the classical kind), Bayesian Generalized Linear Mixed Effects Models,

3. *Bayesian (non-latent variable) models*

- Many hierarchical Bayesian models, Any model with non-conjugate prior,

Statistical inference

Bayesian non-latent variable models

We want the posterior:

$$p(\theta | \mathbf{x}) = \frac{p(\mathbf{x} | \theta)p(\theta)}{p(\mathbf{x})}$$

Note that we need to compute the *evidence*, i.e. the marginal likelihood of the data:

$$p(\mathbf{x}) = \int p(\theta, \mathbf{x}) d\theta \tag{2.1}$$

Statistical inference

Bayesian latent-variable models

We want the posterior:

$$p(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}, \boldsymbol{\theta})}{p(\mathbf{x})}$$

Note that we need to compute the *evidence*, i.e. the marginal likelihood of the data:

$$p(\mathbf{x}) = \int p(\boldsymbol{\theta}, \mathbf{x}, \mathbf{z}) \, d\boldsymbol{\theta} \, d\mathbf{z} \quad (2.2)$$

Statistical inference

Frequentist latent variable models

We want the maximum likelihood value:

$$\boldsymbol{\theta}_{\text{ML}} := \operatorname{argmax}_{\boldsymbol{\theta}} p(\mathbf{x} \mid \boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \int p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) d \mathbf{z} \quad (2.3)$$

In particular, one requires access to the *marginal* likelihood

$$p(\mathbf{x} \mid \boldsymbol{\theta})_{\text{marginal likelihood}} = \int_{\text{joint (or "complete") likelihood}} p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) d \mathbf{z} \quad (2.4)$$

Statistical inference

In general

We must compute the marginal

$$p(\mathbf{x} \mid \mathbf{c}) = \int p(\mathbf{x}, \mathbf{u} \mid \mathbf{c}) \, d\mathbf{u} \quad (2.5)$$

where

- \mathbf{x} : observed data
- \mathbf{u} : unobserved random variables
- \mathbf{c} : constant values

The need for marginalization in statistical inference

Model	Inferential goal	Target marginal
		$p(\mathbf{x} \mathbf{c})$
Bayesian (non-latent)	$p(\boldsymbol{\theta} \mathbf{x})$	$p(\mathbf{x}) = \int p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta}$
Bayesian latent	$p(\mathbf{z}, \boldsymbol{\theta} \mathbf{x})$	$p(\mathbf{x}) = \int p(\boldsymbol{\theta}, \mathbf{x}, \mathbf{z}) d\boldsymbol{\theta} dz$
Frequentist latent	$\text{argmax}_{\boldsymbol{\theta}} p(\mathbf{x} \boldsymbol{\theta})$	$p(\mathbf{x} \boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z} \boldsymbol{\theta}) dz$

Problem: These marginalizations may be intractable

Example: Hidden Markov Model

Define T : the state transition matrix

ϵ_j : the j th emission distribution, $j = 1, \dots, k$

π : the initial latent state distribution

$$\begin{aligned} p(\mathbf{x} | \theta) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z} | \theta) \\ &= \sum_{\mathbf{z}=(z_1, \dots, z_n)} p(\mathbf{x}, \mathbf{z} | \theta) \\ &= \sum_{\mathbf{z}=(z_1, \dots, z_n)} \pi_{z_1} \epsilon_{z_1}(x_1) T_{z_1, z_2} \epsilon_{z_2}(x_2) T_{z_2, z_3}, \dots, T_{z_{n-1}, z_n} \epsilon_{z_n}(x_n) \end{aligned}$$

has $\mathcal{O}(n k^n)$ complexity. 

Consider e.g. that $(k, n) = (5, 100) \rightarrow 10^{72}$ calculations.



Towards variational inference

We construct a lower bound on the target marginal.

Variational Lower Bound (VLBO)

Let q be any probability density over \mathbf{u} . Then:

$$\begin{aligned}\ln p(\mathbf{x} \mid \mathbf{c}) &= \ln \int p(\mathbf{u}, \mathbf{x} \mid \mathbf{c}) d\mathbf{u} \\ &= \ln \int q(\mathbf{u}) \frac{p(\mathbf{u}, \mathbf{x} \mid \mathbf{c})}{q(\mathbf{u})} d\mathbf{u} \\ &\stackrel{\text{Jensen's}}{\geq} \int q(\mathbf{u}) \ln \left(\frac{p(\mathbf{u}, \mathbf{x} \mid \mathbf{c})}{q(\mathbf{u})} \right) d\mathbf{u} \\ &:= \text{VLBO}(q)\end{aligned}$$

Variational Inference: Maximizing the VLBO

Variational Inference

Variational inference (VI) proceeds by finding q^* , the variational density in tractable family \mathcal{Q} which maximizes the VLBO:

$$q^*_{\text{solution}} = \underset{\substack{q \in \mathcal{Q} \\ \text{approximating family}}}{\operatorname{argmax}} \text{VLBO}(q)$$

Decompositions of the VLBO

Energy/Entropy Decomposition of the VLBO

By simply appealing to properties of the logarithm and the definition of expectation, we obtain

$$\begin{aligned}\text{VLBO}(q) &= \int q(\mathbf{u}) \ln p(\mathbf{x}, \mathbf{u} \mid \mathbf{c}) d\mathbf{u} - \int q(\mathbf{u}) \ln q(\mathbf{u}) d\mathbf{u} \\ &= \underset{\text{energy}}{\mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} \mid \mathbf{c})]} + \underset{\text{entropy}}{\mathbb{H}[q(\mathbf{u})]}\end{aligned}$$

Decompositions of the VLBO

Likelihood/Prior Decomposition of the VLBO

By applying the chain rule to the preceding, and then reapplying the definition of KL divergence, we obtain another nice form

$$\begin{aligned}\text{VLBO}(q) &= \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} | \mathbf{c})] + \mathbb{H}[q(\mathbf{u})] \\ &= \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} | \mathbf{c})] - \mathbb{E}_q[\log q(\mathbf{u})] \\ &= \mathbb{E}_q[\log p(\mathbf{x} | \mathbf{u}, \mathbf{c})] + \mathbb{E}_q[\log p(\mathbf{u} | \mathbf{c})] - \mathbb{E}_q[\log q(\mathbf{u})] \\ &= \mathbb{E}_q[\log p(\mathbf{x} | \mathbf{u}, \mathbf{c})] - \text{KL}(q(\mathbf{u}) || p(\mathbf{u} | \mathbf{c}))\end{aligned}$$

expected log likelihood divergence from prior

Note that the first term grows in magnitude as the number of samples increases; thus, the prior's influence diminishes asymptotically.

Maximizing the VLBO minimizes the KL divergence

By definition, the KL divergence from the target posterior to the variational density is given by

$$\text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) = \mathbb{E}_q \left[\log \frac{q(\mathbf{u})}{p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})} \right]$$

By the chain rule, we get

$$\begin{aligned} \text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) &= \mathbb{E}_q[\log q(\mathbf{u})] - \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} \mid \mathbf{c})] && + \log p(\mathbf{x} \mid \mathbf{c}) \\ &= -\text{VLBO}(q) && + \text{constant} \end{aligned}$$

Maximizing the VLBO minimizes the KL divergence

By definition, the KL divergence from the target posterior to the variational density is given by

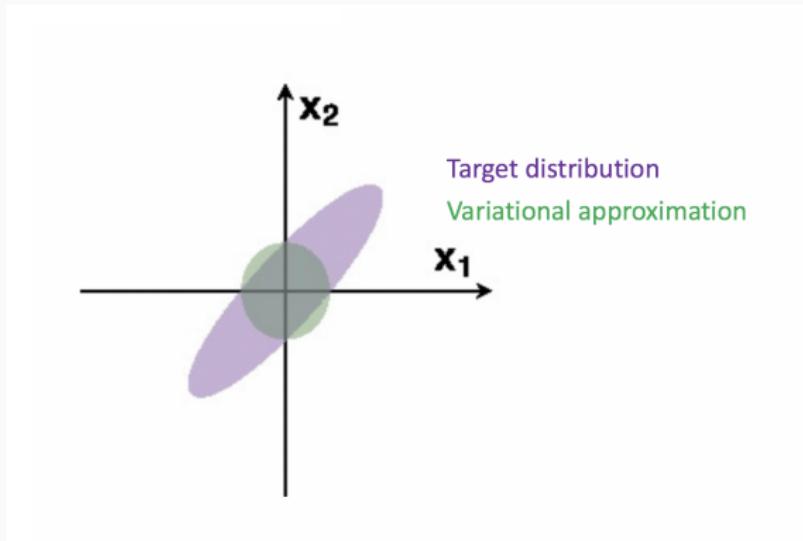
$$\text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) = \mathbb{E}_q \left[\log \frac{q(\mathbf{u})}{p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})} \right]$$

By the chain rule, we get

$$\begin{aligned} \text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) &= \mathbb{E}_q[\log q(\mathbf{u})] - \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{u} \mid \mathbf{c})] + \log p(\mathbf{x} \mid \mathbf{c}) \\ &= -\text{VLBO}(q) + \text{constant} \end{aligned}$$

Note: The optimal variational density, $q^*(\mathbf{u})$ is the target posterior density $p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})$ when the underlying variational family \mathcal{Q} is unrestricted

Problem: VI underestimates variance of the true posterior



$$\text{KL}(q(\mathbf{u}) \parallel p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})) = \mathbb{E}_q \left[\log \frac{q(\mathbf{u})}{p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})} \right]$$

Intuition

- If $q(\mathbf{u})$ is low, then we don't care (because of the expectation).
- If $q(\mathbf{u})$ is high and $p(\mathbf{x}, \mathbf{u} \mid \mathbf{c})$ is low, then we pay a price

Summary

- VI is a general tool. It is useful whenever you face intractable marginals.

Model	Inferential goal	Target marginal	Variational density	Target posterior
General case	infer about θ	$p(\mathbf{x} \mid \mathbf{c})$	$q(\mathbf{u})$	$p(\mathbf{u} \mid \mathbf{x}, \mathbf{c})$
Bayesian (non-latent)	$p(\theta \mid \mathbf{x})$	$p(\mathbf{x}) = \int p(\theta, \mathbf{x}) d\theta$	$q(\theta)$	$p(\theta \mid \mathbf{x})$
Bayesian latent	$p(\mathbf{z}, \theta \mid \mathbf{x})$	$p(\mathbf{x}) = \int p(\theta, \mathbf{x}, \mathbf{z}) d\theta \, d\mathbf{z}$	$q(\mathbf{z}, \theta)$	$p(\mathbf{z}, \theta \mid \mathbf{x})$
Frequentist latent	$\text{argmax}_{\theta} p(\mathbf{x} \mid \theta)$	$p(\mathbf{x} \mid \theta) = \int p(\mathbf{x}, \mathbf{z} \mid \theta) d\mathbf{z}$	$q(\mathbf{z})$	$p(\mathbf{z} \mid \mathbf{x}, \theta)$

Variational Inference and Expectation Maximization

Expectation Maximization (EM)

The EM algorithm refines an initial guess $\theta^{(0)}$ via the recursion

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \mathbb{E}_{p(z|x, \theta^{(t)})} \left[\ln p(x, z | \theta) \right]$$

until convergence to a local optimum.

Example: Exponential Hidden Markov Model

E-step: Compute $p_i := p(z_i | x_i, \theta^{(t)})$ via the forward-backward algorithm.

M-step: Just a computation of **weighted** empirical reciprocal means:

$$\hat{\theta}_k^{(t)} = \frac{\sum_i (p_i = k)}{\sum_i (p_i = k) x_i}$$

Example in cybersecurity: Kantchelian, A., et al. (2015). *Better malware ground truth: Techniques for weighting anti-virus vendor labels*. In Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security (pp. 45-56). ACM.

Expectation Maximization (EM)

The EM algorithm refines an initial guess $\theta^{(0)}$ via the recursion

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \mathbb{E}_{p(z|x, \theta^{(t)})} \left[\ln p(x, z | \theta) \right]$$

until convergence to a local optimum.

Example: Gaussian Hidden Markov Model

E-step: Compute $p_i := p(z_i | x_i, \theta^{(t)})$ via the forward-backward algorithm.

M-step: Just a computation of **weighted** empirical means and variances:

$$\hat{\mu}_k^{(t)} = \frac{\sum_i (p_i = k) x_i}{\sum_i (p_i = k)}, \quad \hat{\Sigma}_k^{(t)} = \frac{\sum_i (p_i = k) (x_i - \hat{\mu}^{(t)}) (x_i - \hat{\mu}^{(t)})^T}{\sum_i (p_i = k)}$$

Example in cybersecurity: Kantchelian, A., et al. (2015). *Better malware ground truth: Techniques for weighting anti-virus vendor labels*. In Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security (pp. 45-56). ACM.

EM from the perspective of VI

For a frequentist latent variable model, the VLBO is

$$\text{VLBO}(q_z, \theta) = \mathbb{E}_q [\log p(x, z | \theta)] + \mathbb{H}[q(z)]$$

Applying coordinate ascent (in the sense of variational calculus), we get the following update equations:

$$\mathbf{q \; update :} \quad q_z^{(t+1)} = \operatorname{argmax}_{q_z} \text{VLBO}(q_z; \theta^{(t)}) \quad (3.1)$$

$$\mathbf{\theta \; update :} \quad \theta^{(t+1)} = \operatorname{argmax}_{\theta} \text{VLBO}(q_z^{(t+1)}; \theta) \quad (3.2)$$

As argued earlier, we can solve the *q update* exactly by setting

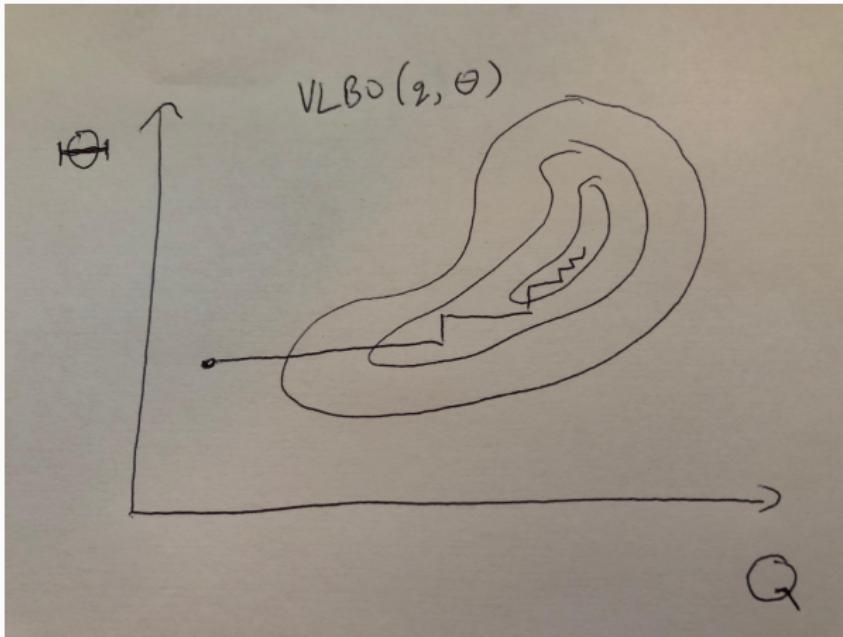
$$q_z^{(t+1)} = p(z | x; \theta^{(t)})$$

in which case the *θ update* becomes

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \mathbb{E}_{p(z | x, \theta^{(t)})} \left[\ln p(x, z | \theta) \right] \quad (3.3)$$

which is precisely the EM algorithm.

EM as coordinate ascent on the VLBO



- If \mathcal{Q} unrestricted, we have EM
- What if we restrict \mathcal{Q} ?

Variational Expectation Maximization (VEM)

Consider a frequentist latent variable model. Since we don't always have access to $p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta})$, we may restrict our variational family \mathcal{Q} to some convenient form. In this case, coordinate ascent on the VLBO is given by:

$$\begin{aligned} q_{\mathbf{z}}^{(t+1)} &= \operatorname{argmax}_{q_{\mathbf{z}} \in \mathcal{Q}} \text{VLBO}(q_{\mathbf{z}}; \boldsymbol{\theta}^{(t)}) \\ \boldsymbol{\theta}^{(t+1)} &= \operatorname{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{q_{\mathbf{z}}^{(t+1)}} \left[\ln p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\theta}) \right] \end{aligned}$$

which generalizes the EM algorithm.

Extension: Variational MAP

Consider a **Bayesian latent variable model**. Recall that a(n unscrupulous) Bayesian will sometimes compute a maximum a posteriori (MAP) estimate of parameter θ .

$$\theta_{\text{MAP}} := \operatorname{argmax}_{\theta} p(\mathbf{x}, \theta) = \operatorname{argmax}_{\theta} \int p(\mathbf{x}, \mathbf{z}, \theta) d \mathbf{z} \quad (3.4)$$

To obtain θ_{MAP} here, we swap $p(\mathbf{x}, \mathbf{z}, \theta)$ for $p(\mathbf{x}, \mathbf{z} | \theta)$ in the VLBO and then immediately obtain the same algorithm but with altered M-step

$$q_z^{(t+1)} = \operatorname{argmax}_{q_z \in \mathcal{Q}} \text{VLBO}(q_z; \theta^{(t)})$$

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} \mathbb{E}_{q_z^{(t+1)}} \left[\ln p(\mathbf{x}, \mathbf{z} | \theta) \right] + \text{ln } p(\theta)$$

Variational Bayes Expectation Maximization (VBEM)

Consider a **Bayesian latent variable model**. So again we marginalize over $p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta})$ instead of $p(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta})$. But this time we want the full posterior.

Suppose that we construct variational density with the factorization

$$q(\mathbf{z}, \boldsymbol{\theta}) = q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$$

In this case, the VLBO becomes

$$\text{VLBO}(q_{\mathbf{z}}(\mathbf{z}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) := \int \int q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \left(\frac{p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{x})}{q_{\mathbf{z}}(\mathbf{z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} d\mathbf{z} \quad (3.5)$$

We can perform coordinate ascent on the VLBO with respect to the densities $q_{\mathbf{z}}$ and $q_{\boldsymbol{\theta}}$:

$$\text{VB-E step : } q_{\mathbf{z}}^{(t+1)} = \operatorname{argmax}_{q_{\mathbf{z}}} \text{VLBO}(q_{\mathbf{z}}; q_{\boldsymbol{\theta}}^{(t)})$$

$$\text{VB-M step : } q_{\boldsymbol{\theta}}^{(t+1)} = \operatorname{argmax}_{q_{\boldsymbol{\theta}}} \text{VLBO}(q_{\mathbf{z}}^{(t+1)}; q_{\boldsymbol{\theta}})$$

VBEM: Derivation

See notes.

VBEM: Update Equations

The coordinate ascent equations have the form

$$\textbf{VB-E step : } q_z^{(t+1)} \propto \exp \left(\mathbb{E}_{q_\theta^{(t)}} [\ln p(x, z | \theta)] \right) \quad (3.6)$$

$$\textbf{VB-M step : } q_\theta^{(t+1)} \propto p(\theta) \exp \left(\mathbb{E}_{q_z^{(t)}} [\ln p(x, z | \theta)] \right) \quad (3.7)$$

Prior-likelihood decomposition

Bayes' rule

$$p(\theta | x) \propto \underset{\textit{posterior}}{p(\theta)} \underset{\textit{prior}}{p(x | \theta)} \underset{\textit{likelihood}}{}$$

VB-M update

$$\underset{\textit{variational posterior}}{q_\theta^{(t+1)}} \propto \underset{\textit{prior}}{p(\theta)} \underset{\textit{expected likelihood under variational distribution}}{\exp \left(\mathbb{E}_{q_z^{(t)}} [\ln p(x, z | \theta)] \right)}$$

VI and EM: Summary

Variational inference can be considered as a generalization of the expectation maximization algorithm (which is generally used by frequentists). It

- relaxes the need for tractable computation of the posterior distribution $p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta})$.
- relaxes the assumption that $\boldsymbol{\theta}$ is a deterministic variable; variational calculus lets us do coordinate ascent on the *distribution* governing $\boldsymbol{\theta}$.

Coordinate Ascent Variational Inference (CAVI)

Towards generalizing the VBEM updates

The VB-E and VB-M steps in (3.6) and (3.7) have different forms, suggesting a dependence on the “semantic” distinction between z and θ (as scaling vs. not scaling with the data x).

But the VLBO in (3.5) is symmetric in its arguments, and the derivation of the VBEM update equations does not rely upon this semantic difference.

Towards generalizing the VBEM updates

For example, we could have equivalently presented the VBEM updates as

$$q_{\theta}^{(t+1)} \propto \exp \left(\mathbb{E}_{q_z^{(t)}} [\ln p(x, \theta | z)] \right) \quad (4.1)$$

$$q_z^{(t+1)} \propto p(z) \exp \left(\mathbb{E}_{q_{\theta}^{(t)}} [\ln p(x, \theta | z)] \right) \quad (4.2)$$

by simply interchanging the syntactic roles of z and θ , although the updates are not typically presented this way because joint likelihood is a conventional object.

Coordinate Ascent Variational Inference (CAVI)

Coordinate ascent variational inference (CAVI) is a general approach to fitting models using VI.

This approach generalizes VBEM.

Mean Field Coordinate Ascent Variational Inference (MF-CAVI)

Mean field variational families

A variational family \mathcal{Q} is mean field if it factorizes

$$q(u_1, \dots, u_K) = \prod_{k=1}^K q_k(u_k) \quad (4.3)$$

Mean field coordinate ascent variational inference (MF-CAVI) is CAVI performed under the mean field assumption (4.3).

Update equations for MF-CAVI

To perform coordinate ascent on the VLBO under the mean field assumption (4.3), we iteratively update our variational factors $\{q_k\}_k$ via

$$q_k(u_k) \propto \exp \left\{ \mathbb{E}_{q_{-k}} \left[\log p(u_k \mid \mathbf{u}_{-k}, \mathbf{x}, \mathbf{c}) \right] \right\} \quad (4.4)$$

The derivation uses variational calculus, and is nearly syntactically identical to the derivation of the VBEM updates.

Coordinate Ascent Variational Inference (CAVI)

**Exponential Family Complete
Conditionals**

Motivation

The form of (4.4) suggests that MF-CAVI can be simplified when the complete conditional has known form. We consider the case where complete conditionals are in the exponential family

This situation describes a lot of models:

Models with exponential family complete conditionals

- Bayesian mixture models (where the mixture components are exponential families and conjugate priors are used)
- Bayesian Hidden Markov Models (HMMs)
- Hierarchical HMMs
- Switching Kalman Filters
- Certain hierarchical regression models (Linear regression, Poisson regression, probit regression)
- Matrix factorization models
- [...]

The Exponential Family

We define an *exponential family* of probability distributions as those distributions whose density has the following form

$$p(x | \eta) = h(x) \exp\{\eta^T s(x) - a(\eta)\} \quad (4.5)$$

where we refer to h as the base measure, η as the natural parameter, s as the sufficient statistics, and a as the log normalizer.

Note: $x \perp\!\!\!\perp \theta | t(x)$

MF-CAVI updates on random variables with exponential family complete conditionals

Claim

Consider a model with unobserved random variables (u_1, \dots, u_k) . Let

1. Variational density q have mean field factorization $q = q_k(u_k)q_{-k}(u_{-k})$.
2. $p(u_k | u_{-k}, x)$ be in exponential family \mathcal{E} with natural parameter $\eta_k(u_{-k}, x)$.

Then optimal mean field CAVI update (4.4) puts

$$q_k \in \mathcal{E} \tag{4.6}$$

with natural parameter

$$\nu_k = \mathbb{E}_{q_{-k}}[\eta_k(u_{-k}, x)] \tag{4.7}$$

Take Home

- Variational factor is in same family as complete conditional.
- Its natural parameter is the expectation (with respect to the other variational factors) of the natural parameter of the complete conditional.

Proof

If the k th complete conditional is in the exponential family, then we have

$$p(u_k \mid u_{-k}, x) = h(u_k) \exp\{\eta_k^T s(u_k) - a(\eta_k)\} \quad (4.8)$$

Note that our notation suppresses that the natural parameter η_k depends on the conditioning variables (u_{-k}, x) .

By substituting (4.8) into (4.4) and discarding factors that do not depend on u_k , we obtain

$$\begin{aligned} q_k(u_k) &\propto \exp \left\{ \mathbb{E}_{q_{-k}} \left[\eta_k^T s(u_k) + \log h(u_k) \right] \right\} \\ &\propto h(u_k) \exp \left\{ \mathbb{E}_{q_{-k}} [\eta_k]^T s(u_k) \right\} \end{aligned}$$

Since h and s are identical to those of (4.8), the claim holds.

Coordinate Ascent Variational Inference (CAVI)

Evaluation

Evaluation

CAVI uses deterministic optimization methods, and thereby requires analytical expansions of the expectations.

- ✗ This can require expert analysis, in terms of setting up the model, carefully choosing the variational family, and carrying out the integrals. Moreover, expert analysis is required each time the model changes.
- ✓ However, CAVI is still used in state-of-the-art models for efficient subroutines in black box models.

Example: Bayesian Gaussian Mixture Model

Example: Bayesian Gaussian Mixture Model

To see the mean field CAVI algorithm (4.4) in a concrete context, consider a version of the Bayesian Gaussian Mixture Model.

$$\mu_k \sim \text{Normal}(M_k = 0, V_k = \sigma^2) \quad k = 1, \dots, K$$

$$c_i \sim \text{Categorical}(\pi_1, \dots, \pi_K) \quad i = 1, \dots, n$$

$$x_i \mid c_i, \mu \sim \text{Normal}(\mu_{c_i}, 1) \quad i = 1, \dots, n$$

(The model is simple in that it assumes univariate observations and that each mixture component has unit variance.)

The joint density, by chain rule, is

$$p(x, c, \mu) = p(\mu) \prod_{i=1}^n p(c_i) p(x_i \mid c_i, \mu)$$

And a mean-field variational family is given by

$$q(c, \mu) = \prod_{k=1}^K q(\mu_k) \prod_{i=1}^n q(c_i)$$

Example: Bayesian Gaussian Mixture Model

We apply (4.4) to determine the coordinate updates for q_{c_i} , the variational factors governing cluster assignments.

$$\begin{aligned} q(c_{ik}) &\propto \exp \left\{ \mathbb{E}_{q_{\mu_k}} \left[\log p(c_i = k) + \log p(x_i | c_i = k, \mu) \right] \right\} \\ &\propto \exp \left\{ \mathbb{E}_{q_{\mu_k}} \left[\log \pi_k + x_i \mu_k - \frac{1}{2} \mu_k^2 \right] \right\} \\ &\propto \pi_k \exp \left\{ x_i \mathbb{E}_{q_{\mu_k}} [\mu_k] - \frac{1}{2} \mathbb{E}_{q_{\mu_k}} [\mu_k^2] \right\} \end{aligned}$$

The coordinate updates for q_{μ_k} are derived similarly. They reveal that q_{μ_k} are Gaussian, and hence the above expectations are easy to compute.

Note: We abuse notation, and write $q(c_{ik})$ as shorthand for $q(c_i = k)$

Bayesian Gaussian Mixture Model: Updates to mixture component means

Using the same strategy as when updating cluster assignments c_i , we obtain

$$\begin{aligned} q(\mu_k) &\propto \exp \left\{ \mathbb{E}_{-q_{\mu_k}} \left[\log p(\mu_k) + \sum_{i=1}^n \log p(x_i \mid c_i = k, \mu) \right] \right\} \\ &\propto \exp \left\{ -\frac{1}{2\sigma^2} \mu_k^2 + \sum_{i=1}^n \mathbb{E}_{q_c} \left[\mathbf{1}_{c_i=k} \left(x_i \mu_k - \frac{1}{2} \mu_k^2 \right) \right] \right\} \\ &\propto \exp \left\{ \left(\sum_{i=1}^n q(c_{ik}) x_i \right) \mu_k + -\frac{1}{2} \left(\frac{1}{\sigma^2} + \sum_{i=1}^n q(c_{ik}) \right) \mu_k^2 \right\} \end{aligned}$$

which is an exponential family distribution with sufficient statistics (μ_k, μ_k^2) and base measure $\propto 1$; hence it is Gaussian.

Bayesian Gaussian Mixture Model: Updates to mixture component means

It is easy to show that for a Gaussian with mean M and variance V , the natural parameters are given by

$$\eta_1 = \frac{M}{V}, \quad \eta_2 = -\frac{1}{2V}$$

From the last slide, the variational density $q(\mu_k)$ has natural parameters

$$\eta_1 = \left(\sum_{i=1}^n q(c_{ik})x_i \right), \quad \eta_2 = -\frac{1}{2} \left(\frac{1}{\sigma^2} + \sum_{i=1}^n q(c_{ik}) \right)$$

Using this, we can backsolve to determine the updates to the mean and variance of the Gaussian variational density governing the k th cluster mean:

$$M_k = \frac{\sum_{i=1}^n q(c_{ik})x_i}{1/\sigma^2 + \sum_{i=1}^n q(c_{ik})}, \quad V_k = \frac{1}{1/\sigma^2 + \sum_{i=1}^n q(c_{ik})}$$

VB Predictive vs. ML Solution

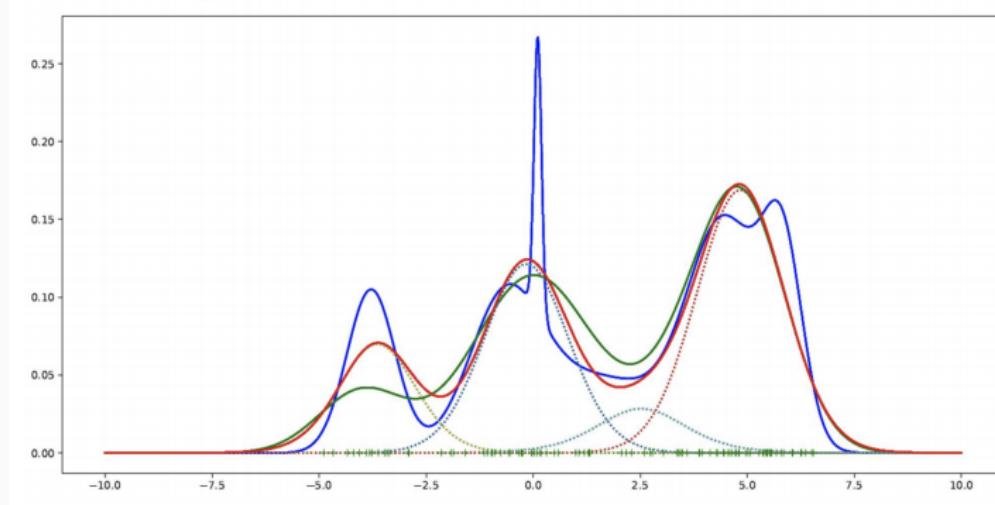


Image Credit: Lukas Burget

- VB was initialized from the ML solution
- VB recovers from ML overfitting and is closer to the true distribution for generating the training data

Example: Bayesian Hidden Markov Model

Hidden Markov Model

A hidden Markov model (HMM) is a tool for representing probability distributions over sequences of observations.

The HMM assumes that

- The observation at time t , y_t , was generated by some process whose state x_t is hidden from the observer.
- The sequence of states satisfies the *Markov property*: conditional on the current state x_t , past and future hidden states are independent.
- There is an additional Markov property on outputs: conditional on the current state x_t , the output y_t is independent of all other hidden states and outputs.

Notation

- $y_{1:T} = (y_1, \dots, y_T)$ observed sequence
- $x_{1:T} = (x_1, \dots, x_T)$: hidden state sequence ($x_t \in \{1, \dots, K\}$)
- $\pi = \{\pi_k\}, \pi_k = P(x_1 = k)$: initial state distribution
- $A = \{A_{kk'}\}, A_{kk'} = P(x_t = k' \mid x_{t-1} = k)$: state transition probability matrix
- $\phi = (\phi_k)_{k=1}^K$ a set of parameters, each governing an output distribution (also called emissions distribution) associated to each hidden state; that is, $P(y_t \mid x_t = k) = P(y_t \mid \phi_k)$.
- $\theta = (\pi, A, \phi)$: model parameters

HMM: Complete Data Likelihood (CDL)

The complete data likelihood for the HMM is given by

$$\begin{aligned} p(x_{1:T}, y_{1:T} \mid \theta) &= p(x_1 \mid \theta)p(y_1 \mid x_1, \theta) \prod_{t=2}^T p(x_t \mid x_{t-1}, \theta)p(y_t \mid x_t, \theta) \\ &= p(x_1 \mid \pi)p(y_1 \mid x_1, \phi) \prod_{t=2}^T p(x_t \mid x_{t-1}, A)p(y_t \mid x_t, \phi) \\ &= \pi_{x_1} \prod_{t=2}^T A_{x_{t-1}, x_t} \prod_{t=1}^T p(y_t \mid \phi_{x_t}) \end{aligned} \tag{6.1}$$

HMM: Complete Data Likelihood is an Exponential Family

We continue

$$p(x_{1:T}, y_{1:T} \mid \theta) = \exp \left\{ \log p(x_1 \mid \pi) + \sum_{t=2}^T \log p(x_t \mid x_{t-1}, A) + \sum_{t=1}^T \log p(y_t \mid x_t, \phi) \right\}$$

$$= \exp \left\{ \log \pi_{x_1} + \sum_{t=2}^T \log A_{x_{t-1}, x_t} + \sum_{t=1}^T \log p(y_t \mid \phi_{x_t}) \right\} \quad (6.2)$$

$$\begin{aligned} &= \exp \left\{ \sum_{k=1}^K x_1^k \log \pi_k + \sum_{t=2}^T \sum_{k,k'=1}^K x_{t-1}^k x_t^{k'} \log A_{kk'} \right. \\ &\quad \left. + \sum_{t=1}^T \sum_{k=1}^K x_t^k \log p(y_t \mid x_t, \phi_k) \right\} \end{aligned} \quad (6.3)$$

where we have defined

$$x_t^k = \begin{cases} 1, & \text{if the latent state at time } t \text{ is } k \\ 0, & \text{otherwise} \end{cases}$$

Thus, the complete data likelihood (although not the marginal likelihood $p(x_{1:T} \mid \theta)$) is in the exponential family, so long as the emissions distributions are. The sufficient statistics for $\log \pi_k$ are x_1^k , and the sufficient statistics for $\log A_{kk'}$ are $\sum_{t=2}^T x_{t-1}^k x_t^{k'}$.

Prior Distributions

Let us impose a prior whose density factorizes as follows

$$p(\theta) = p(A)p(\phi)p(\pi) \quad (6.4)$$

and let us further assume that the priors have the form

$$p(\pi) = \text{Dirichlet}(\pi | \alpha^\pi) \quad (6.5)$$

$$p(A) = \prod_{k=1}^K p(A_k | \alpha^{A_k}) = \prod_{k=1}^K \text{Dirichlet}(A_k | \alpha^{A_k}) \quad (6.6)$$

where A_k designates the k th row of A .

π has a Dirichlet complete conditional

By the structure of the complete data likelihood (6.3) and prior (6.4), (6.5) , we see that the complete conditional of π is a Dirichlet

$$\begin{aligned} p(\pi \mid \theta_{-\pi}, x_{1:T}, y_{1:T}) &\propto p(\pi, \theta_{-\pi}, x_{1:T}, y_{1:T}) \\ &\propto \exp \left\{ \sum_{k=1}^K (\alpha_k^\pi - 1) \log \pi_k \right\} \exp \left\{ \sum_{k=1}^K x_1^k \log \pi_k \right\} \\ &\propto \exp \left\{ \sum_{k=1}^K \left(\alpha_k^\pi - 1 + x_1^k \right) \log \pi_k \right\} \end{aligned}$$

since this is an exponential family with base measure $h(\pi) \propto 1$ and the sufficient statistics are $\log \pi = [\log \pi_1, \dots, \log \pi_K]^T$.

Variational update for π

From the last slide, the natural parameter for the complete conditional for π is given by

$$\eta^\pi = \begin{bmatrix} \alpha_1^\pi - 1 + x_1^1 \\ \dots \\ \alpha_K^\pi - 1 + x_1^K \end{bmatrix}$$

Let us assume the factorization $q(A, \phi, \pi, z) = q(\phi, A, z)q(\pi)$.

Then we can apply our Claim (Slide 44) to discover that the optimal MF-CAVI update takes $q(\pi)$ to be Dirichlet with variational natural parameter

$$\nu^\pi = \mathbb{E}_{q_{-\pi}}[\eta^\pi] = \begin{bmatrix} \alpha_1^\pi - 1 + \mathbb{E}_{q_x}[x_1^1] \\ \dots \\ \alpha_K^\pi - 1 + \mathbb{E}_{q_x}[x_1^K] \end{bmatrix} = \alpha^\pi + \mathbb{E}_{q_x}[x_1]$$

The rows of A have Dirichlet complete conditionals

The updates for the rows A_k can be determined using a similar argument as for π . By the structure of the complete data likelihood (6.3) and prior (6.4), (6.6), we see that the complete conditional of A_k is also Dirichlet

$$\begin{aligned} p(A_k \mid \theta_{-A_k}, x_{1:T}, y_{1:T}) &\propto p(A_k, \theta_{-A_k}, x_{1:T}, y_{1:T}) \\ &\propto \exp \left\{ \sum_{k'=1}^K (\alpha_{k'}^{A_k} - 1) \log A_{kk'} \right\} \exp \left\{ \sum_{k'=1}^K \sum_{t=2}^T x_{t-1}^k x_t^{k'} \log A_{kk'} \right\} \\ &\propto \exp \left\{ \sum_{k'=1}^K \left((\alpha_{k'}^{A_k} - 1) + \sum_{t=2}^T x_{t-1}^k x_t^{k'} \right) \log A_{kk'} \right\} \end{aligned}$$

since this is an exponential family with base measure $h(A_k) \propto 1$ and sufficient statistics $\log A_k = [\log A_{k1}, \dots, \log A_{kK}]^T$.

Variational update for A_k

From the last slide, the natural parameter for the complete conditional for A_k is given by

$$\eta^{A_k} = \begin{bmatrix} (\alpha_1^{A_k} - 1) + \sum_{t=2}^T x_{t-1}^k x_t^1 \\ \dots \\ (\alpha_K^{A_k} - 1) + \sum_{t=2}^T x_{t-1}^k x_t^K \end{bmatrix}$$

Let us assume factorization $q(A, \phi, \pi, z) = q(\phi, \pi, z) \prod_{k=1}^K q(A_k)$.

Then we can apply our Claim (Slide 44) to discover that the optimal MF-CAVI update takes $q(A_k)$ to be Dirichlet with variational natural parameter

$$\nu^{A_k} = \mathbb{E}_{q_{\phi}}[\eta^{A_k}] = \begin{bmatrix} (\alpha_1^{A_k} - 1) + \sum_{t=2}^T \mathbb{E}_{q_x}[x_{t-1}^k x_t^1] \\ \dots \\ (\alpha_K^{A_k} - 1) + \sum_{t=2}^T \mathbb{E}_{q_x}[x_{t-1}^k x_t^K] \end{bmatrix} = \alpha^{A_k} + \sum_{t=2}^T \mathbb{E}_{q_x}[x_{t-1}^k x_t]$$

Variational update for x

By applying (4.4) and substituting in the form of the complete data likelihood (6.1), we obtain

$$\begin{aligned} q(x) &\propto \exp \left\{ \mathbb{E}_{-q(z)} [\log p(x_{1:T}, y_{1:T}, \theta)] \right\} \\ &\propto \exp \left\{ \mathbb{E}_{q(\pi)} \log \pi_{x_1} \right\} \prod_{t=2}^T \exp \left\{ \mathbb{E}_{q(A)} \log A_{x_{t-1}, x_t} \right\} \prod_{t=1}^T \exp \left\{ \mathbb{E}_{q(\phi)} p(y_1 \mid \phi_{x_t}) \right\} \end{aligned}$$

Compare this to the original complete data likelihood (6.1). The structure is identical if we define the alterations $(A, \pi, p) \rightarrow (A^*, \pi^*, p^*)$.

In particular, we define

$$\pi^* = \{\pi_k^*\}_{k=1}^K : \pi_k^* = \exp \left\{ \mathbb{E}_{q(\pi)} \log \pi_{x_1} \right\} = \exp \left\{ \psi(\nu_k^\pi) - \psi \left(\sum_{j=1}^K \nu_j^\pi \right) \right\}$$

$$A^* = \{A_{kk'}^*\}_{k,k'=1}^K : A_{kk'}^* = \exp \left\{ \mathbb{E}_{q(A)} \log A_{kk'} \right\} = \exp \left\{ \psi(\nu_{k'}^{A_k}) - \psi \left(\sum_{j=1}^K \nu_j^{A_k} \right) \right\}$$

$$p^* = \{p^*(\cdot | \phi_k)\}_{k=1}^K : p_k^* \propto \exp \left\{ \mathbb{E}_{q(\phi_k)} \log p(\cdot | \phi_k) \right\}$$

$\psi(\cdot)$ is the digamma function; the expectation of the log of a component of a Dirichlet-distributed probability vector is well-known.

Then we can use the forward-backwards algorithm (FBA) to update $q(x)$.

- A standard HMM applies FBA to (A, π, p) in order determine the marginals $p(x_t | y, \theta)$ and pairwise marginals $p(x_{t-1}, x_t | y, \theta)$.
- Likewise, we can apply FBA to (A^*, π^*, p^*) to determine the marginals $q(x_t)$ and pairwise marginals $q(x_{t-1}, x_t)$ (which, as we have seen, are sufficient for performing the other updates).

Example: Latent Dirichlet Allocation (LDA)

Acknowledgements

This section, especially the intro, borrows heavily from David Blei's 2012 ICML tutorial.

Overview

LDA is a generative probabilistic model of a corpus of documents of text.

LDA assumes:

- There is a set of topics that describe the corpus
- Each document exhibits these topics to varying degrees (each word in a document was generated by one of these topics.).

So:

- The topics and how they relate to the documents are hidden structure
- The main computational problem is to infer this hidden structure

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* 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

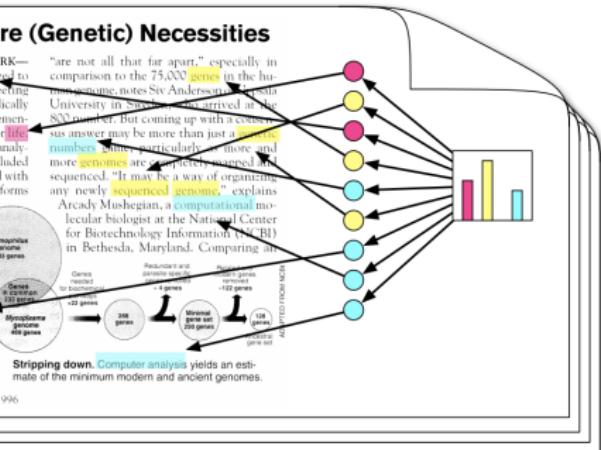
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Stéphane Andréasson, a geneticist at the University of Stockholm who arrived at the 800 number. But coming up with a count of this answer may be more than just a numbers game; particularly, more and more genomes are being collected, parsed, and sequenced. "It may be a way of organizing and easily sequencing genomes," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

the genomes, he says, "we can find the minimum set of genes that are common to all of them."

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

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



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

Source: David Blei, 2012 ICML Tutorial

Topics



Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,¹ 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

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

SCIENCE • VOL. 272 • 24 MAY 1996

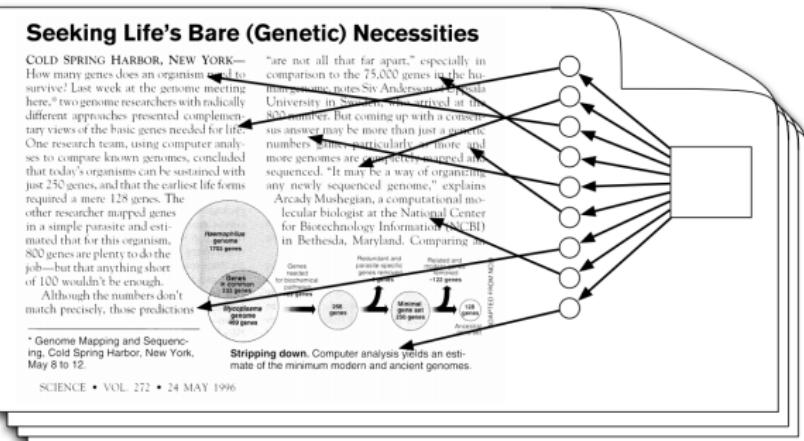
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andreassen of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a matter of numbers. Some predictability: more and more genomes are being pieced together 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

Hemophilus influenzae 1746 genes
Genes in common 128 genes
Hypothetical genes 60 genes

Genes needed to maintain minimum life 128 genes
Residual and overlapping genes 500 genes
Required and minimal genes 128 genes
Minimal gene set 128 genes
Actual 128 genes
Surprised me

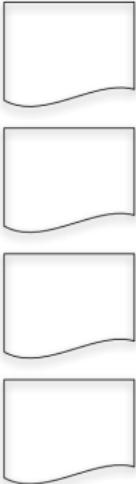
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Topic proportions and assignments

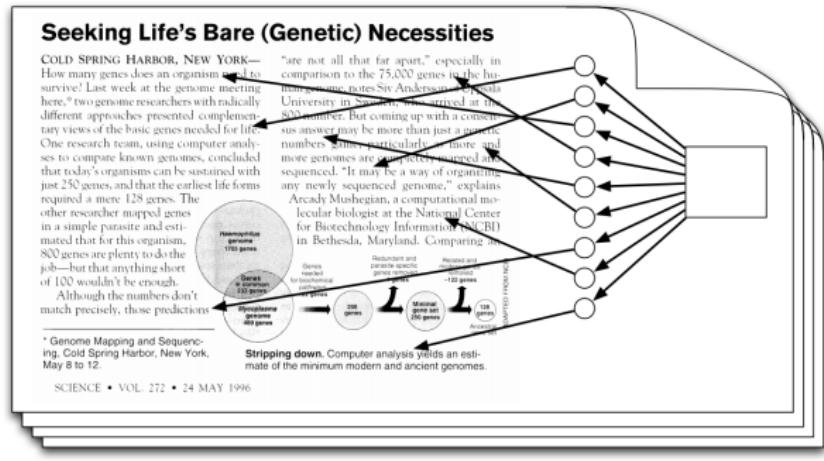


- In reality, we only observe the documents.
- The other structure is **hidden variables**.

Topics



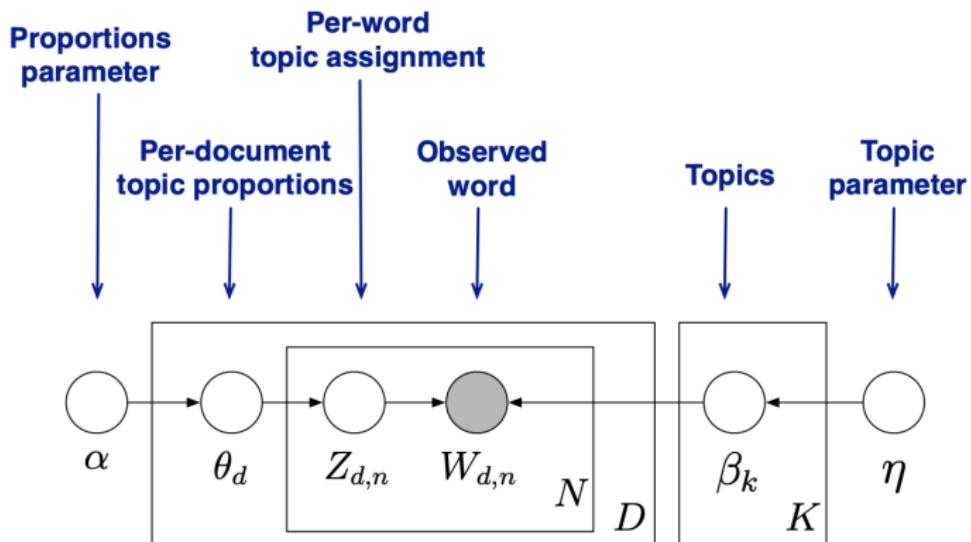
Documents



- Our goal is to **infer** the hidden variables.
- I.e., compute their distribution conditioned on the documents

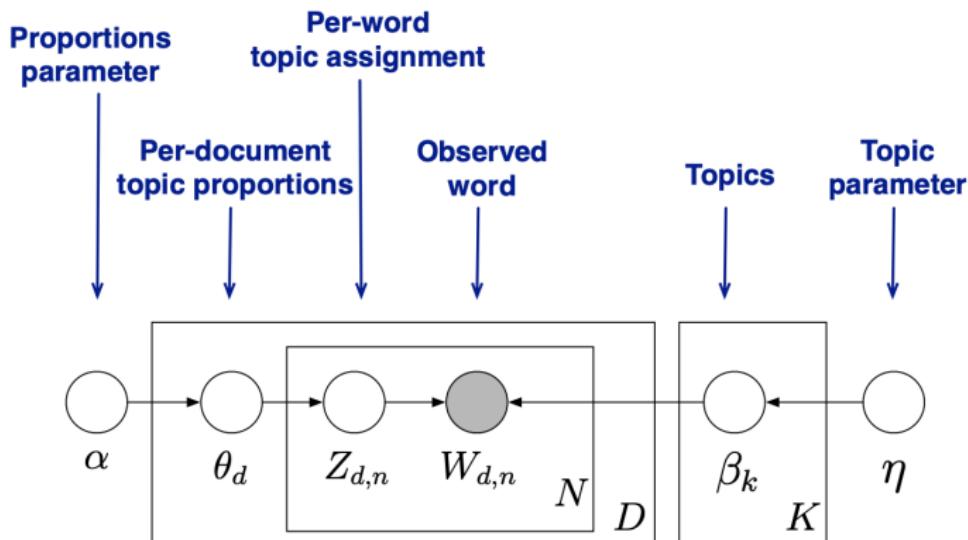
$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

LDA as a graphical model



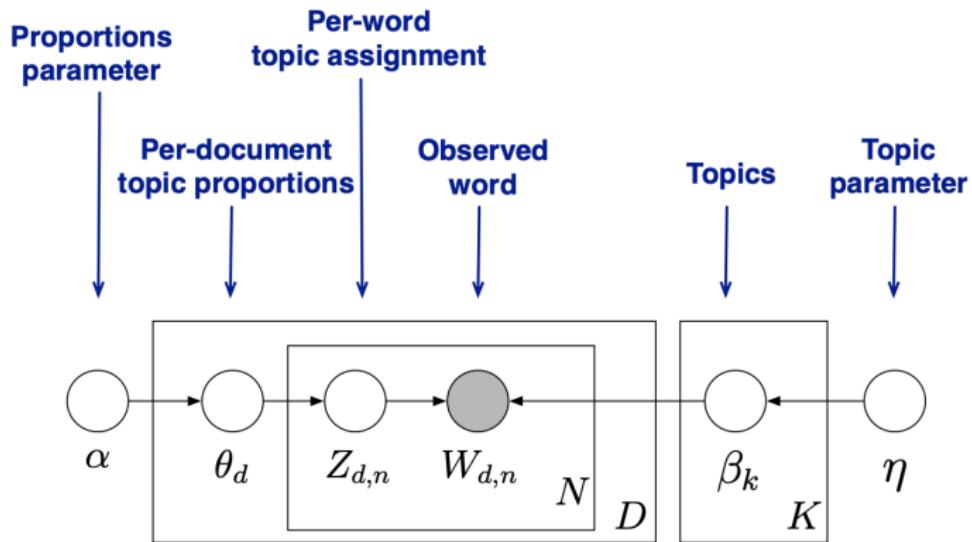
- Encodes assumptions
- Defines a factorization of the joint distribution

LDA as a graphical model



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

Joint distribution



$$p(z, \theta, w, \beta | \alpha, \eta) = \prod_{k=1}^K p(\beta_k | \eta) \underset{\text{Dirichlet}}{\prod_{d=1}^D p(\theta_d | \alpha)} \underset{\text{Categorical}}{\prod_{n=1}^N p(z_n | \theta)} \underset{\text{Categorical}}{p(w_n | z_n, \beta)} \quad (7.1)$$

LDA: Generative Process

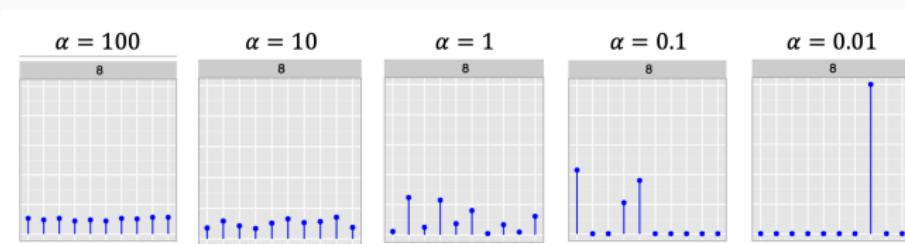
- Choose V vocabulary words, K topics.
- $\eta \in \mathbb{R}^V, \alpha \in \mathbb{R}^K$ are *hyperparameters*.
- For k in $(1, K)$:
 - Choose *per-corpus topic distribution* $\beta_k \in \mathbb{R}^V \sim \text{Dir}(\eta)$
- For d in $(1, D)$:
 - Choose *per-document topic proportions* $\theta_d \in \mathbb{R}^K \sim \text{Dir}(\alpha)$
 - For j in $(1, N_d)$:
 - Choose *topic* $z_{d,n} \sim \text{Categorical}_K(\theta_d)$
 - Choose *word* $w_{d,n} \sim \text{Categorical}_V(\beta_{z_{d,n}})$

Aside: Dirichlet Distribution

- Dirichlet distribution is *conjugate prior* of Multinomial

$$p(\theta | \alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \cdots \theta_k^{\alpha_k-1}$$

- The parameter α controls the shape and sparsity of θ
 - high α = uniform θ , small α = sparse θ

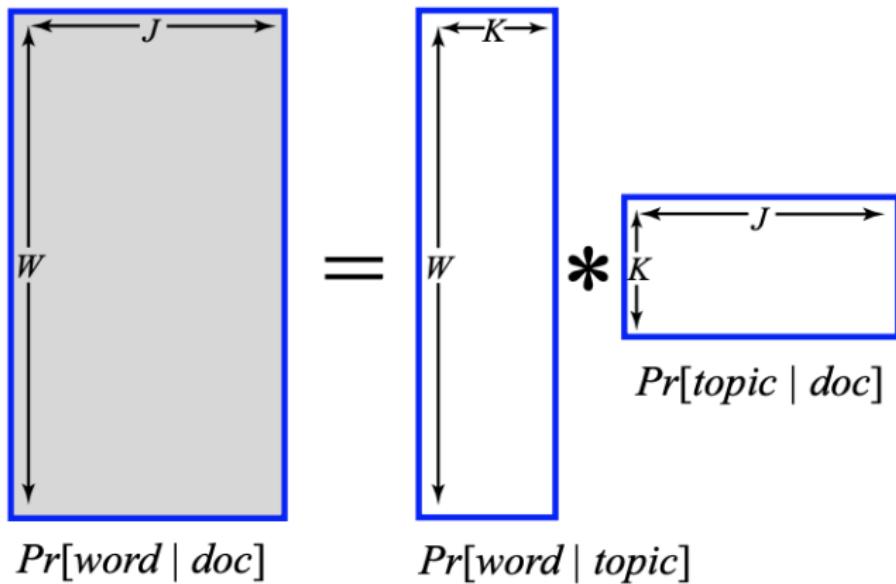


LDA Example Inference

- Data: 17K Science documents from 1990-2000 (11M words, 20K unique terms)
- Model: 100-topic LDA model, fit using variational inference

1 dna gene sequence genes sequences human genome genetic analysis two	2 protein cell cells proteins receptor fig binding activity activation kinase	3 water climate atmospheric temperature global surface ocean carbon atmosphere changes	4 says researchers new university just science like work first years	5 mantle high earth pressure seismic crust temperature earths lower earthquakes
6 end article start science readers service news card circle letters	7 time data two model to system number different with on	8 materials surface high structure temperature molecules chemical molecular to university	9 dna rna transcription protein site binding sequence proteins specific sequences	10 disease cancer patients human gene medical studies drug normal drugs
11 years million ago age university north early fig evidence record	12 species evolution population evolutionary university populations natural studies genetic today	13 protein structure proteins two amino binding acid residues molecular structural	14 cells cell virus hiv infection immune human antigen infected viral	15 space solar observations earth stars university mass sun astronomers telescope
16 fax manager science aaas advertising sales member recruitment associate washington	17 cells cell gene genes expression development mutant mice fig biology	18 energy electron state light quantum physics electrons high laser magnetic	19 research science national scientific scientists new states university united health	20 neurons brain cells activity fig channels university cortex neuronal visual

LDA as Probabilistic Matrix Factorization



LDA can be seen as a probabilistically constrained factorization of the matrix describing the bag of words composing each group, or document.

The number K of latent topics determines the factorization's rank.

Example: Latent Dirichlet Allocation (LDA)

Old LDA Stuff (Still needs integration)

Structure

LDA assumes that each latent topic is associated with a distribution over a vocabulary of V words. It also assumes that each word in a document was generated from one of K latent topics.

Definitions: Observations

- A *vocabulary* is a list of V possible words.
- A *word*, $\mathbf{w}_n \in \text{OneHot}(V)$ indicates which element of the vocabulary was observed as the n th word of the document.¹
- A *document*, $\mathbf{w} \in \text{OneHot}(V)^N$, is a list of N words. It is represented as a $N \times V$ matrix.

¹We define $\text{OneHot}(K)$ as a K dimensional vector having one entry equal to 1 and all other entries equal to 0. Note that this is the support of the $\text{Multinoulli}(K)$ distribution.

Definitions: Hidden Variables

- A *topic indicator* is an integer in $\{1, \dots, K\}$.
- A (*per-word*) *topic assignment*, $z_n \in \text{OneHot}(K)$ indicates which topic generated the n th word in a document
- The (*per-word*) *topic assignments*, $z \in \text{OneHot}(K)^N$, is a list of N topic indicators, one for each word. It is represented as a $N \times K$ matrix.
- The (*per-document*) *topic proportions*, θ , is a (document-specific) probability distribution over the topics.

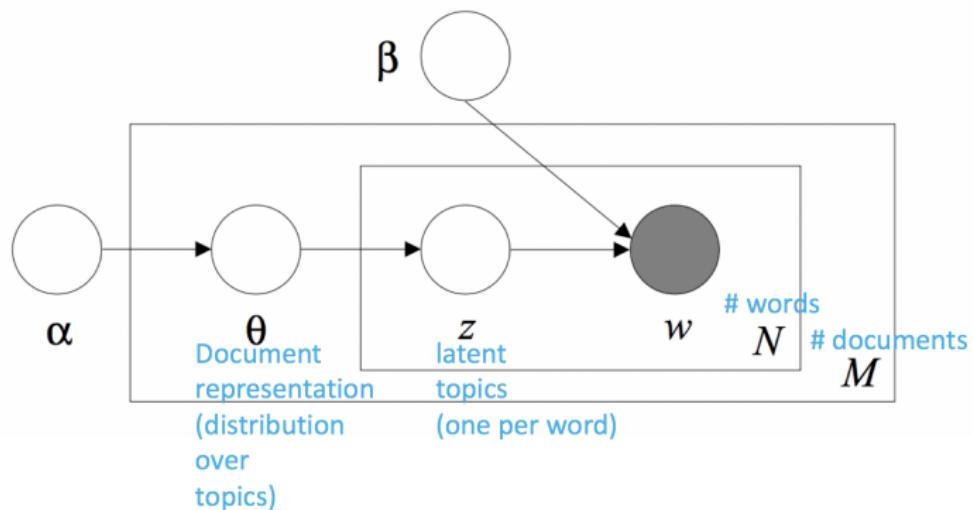
Hyperparameters

- $\alpha \in (\mathbb{R}^+)^K$ is interpreted as the *prior counts of topic indicators*.
- $\beta \in (\Delta^{V-1})^K$ is interpreted as the *topic-conditional word probabilities*, In particular, note that, by construction β is defined by

$$\beta_{kv} = P(w_{n,v} = 1 \mid z_{n,k} = 1),$$

i.e. it is the right-stochastic matrix, where each row is a categorical distribution over vocabulary words given a (latent) topic.

Graphical Model



Generative process

For each document $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_n)$ in a corpus, the generative model is

- Choose a topic distribution for the document $\theta \sim \text{Dirichlet}_K(\alpha)$
- For each word \mathbf{w}_n :
 - Choose a topic $\mathbf{z}_n \sim \text{Multinoulli}_K(\theta)$
 - Choose a word $\mathbf{w}_n \sim \text{Multinoulli}_V(\beta_{\mathbf{z}_n, \cdot})$

Remark. Note that LDA is a “bag-of-words” model; i.e. the probability of a word (or document) is invariant to word order.

Joint distribution

The joint distribution for the generative process is given by

$$p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \underset{\text{Dirichlet}}{\prod_{n=1}^N} p(\mathbf{z}_n \mid \boldsymbol{\theta}) \underset{\text{Multinoulli}}{p(\mathbf{w}_n \mid \mathbf{z}_n, \boldsymbol{\beta})} \quad (7.2)$$

Definition

([·]) We define the operator $[\cdot] : \text{OneHot}(K) \rightarrow \{1, \dots, K\}$ as that which takes a one-hot encoded vector and returns the (unique) index which is non-zero.

Using Definition 1, we can specify the functional forms for the multinoulli log likelihoods as

$$\log p(\mathbf{z}_n \mid \boldsymbol{\theta}) = \log \boldsymbol{\theta}_{[\mathbf{z}_n]} \quad (7.3)$$

$$\log p(\mathbf{w}_n \mid \mathbf{z}_n, \boldsymbol{\beta}) = \log \boldsymbol{\beta}_{[\mathbf{z}_n], [\mathbf{w}_n]} \quad (7.4)$$

Variational distribution

We approximate the posterior $p(\theta | z, w)$ using mean field variational inference (4.3). In particular, we assume that the variational family \mathcal{Q} factorizes as

$$\begin{aligned} q &= q_\delta(\theta) q_\tau(z) \\ &= \underbrace{q_\delta(\theta)}_{\text{Dirichlet}} \prod_{n=1}^N \underbrace{q_{\tau_n}(z_n)}_{\text{Multinoulli}} \end{aligned} \tag{7.5}$$

Update equations

LDA coordinate ascent update equations

$$\tau_{n,k} \underset{\text{var. multinoulli (topics-for-word)}}{\propto} \left(\Psi(\delta_k) - \Psi\left(\sum_j \delta_j\right) \right) \beta_{k,[w_n]} \quad (7.6)$$

$$= \exp \left\{ \mathbb{E}_{q_\delta(\theta)} \left[\log \theta_i \right] \right\} \underset{\text{var. "prior" over topics}}{\beta_{k,[w_n]}} \underset{\text{likelihood}}{\beta_{k,[w_n]}}$$

$$\delta_k \underset{\text{var. dirichlet (documents)}}{=} \underset{\text{prior counts}}{\alpha_k} + \sum_{n=1}^N \underset{\text{var. prob of topics}}{\tau_{n,k}} \quad (7.7)$$

where $\Psi(\cdot)$ is the first derivative of the log Γ function.

- Derivable via VBEM (see notes).
- Could also fit α, β to data via VEM; i.e. VEM does "empirical Bayes" for you.
- Characteristic form: latent variable update depends on the data, global parameter update depends on the latent variable

The role of analytical computations

The variational multinomial update crucially hinges on facts about the exponential family.

In particular, the meat of the proof of the variational multinomial update depends crucially on the fact that the Dirichlet of a single probability component is given by

$$\mathbb{E}_{q_\delta(\theta)} \left[\log \theta_i \right] = \Psi(\delta_i) - \Psi\left(\sum_k \delta_k\right) \quad (7.8)$$

where $\Psi(\cdot)$ is the first derivative of the $\log \Gamma$ function.

This fact is justified via facts about the exponential family (such as that the derivative of the log normalization factor with respect to the natural parameter is equal to the sufficient statistic).

Summary

1. We can derive (see notes) the LDA update equations from the VBEM algorithm, so that we have

$$\text{MF-CAVI} \rightarrow \text{VBEM} \rightarrow \text{LDA}$$

and LDA instantiates the big picture.

2. We have highlighted the potential obstacles for deriving VI updates via classical approaches. This will foreshadow and motivate the development of black-box VI.

Notes!

Questions?

How does VI accommodate the goal of statistical inference?

The target marginal perspective

Given selection of variational family \mathcal{Q} ,

- *Bayesian models*: The optimal variational density $q^*(\mathbf{u})$ raises the (approximate) evidence term $p(\mathbf{x})$ (the term used for Bayesian model comparison) as high as possible.
- *Frequentist models*: The VLBO approximates the true marginal likelihood, $p(\mathbf{x} | \theta)$, which we wanted to maximize.

How would VI accommodate the goal of statistical inference?

The target posterior perspective

Given selection of variational family \mathcal{Q} ,

- *Bayesian models*: The fitted $q^*(\mathbf{u})$ is as similar as possible to the target posterior $p(\mathbf{u} \mid \mathbf{x})$
- *Frequentist models*: The value θ^* is as close as possible to the solution provided by the classical EM algorithm. (See next section; we substitute $q^*(\mathbf{z}) \approx p(\mathbf{z} \mid \mathbf{x}, \theta^{\text{curr}})$ in the E-step.)

VBEM: Derivation

Because the VLBO is a functional, the update equations can be derived via variational calculus.

Theorem (Gelfand & Foman) Let $V[q]$ be a functional of the form

$$\int_a^b F(u, q, q') dx$$

defined on the set of functions $q(u)$ which have continuous first derivatives in $[a, b]$ and satisfy the boundary conditions

$q(a) = A, q(b) = B$. Then a necessary condition for $V[q]$ to have an extremum for a given function $q(u)$ is that $q(u)$ satisfy Euler's equation

$$F_q - \frac{d}{du} F_{q'} = 0$$

VBEM: Derivation

We sketch the derivation for the VB-E step. Take the VLBO and use the chain rule of probability to **isolate** terms that do not involve q_z

$$\text{VLBO}(q_z(z), q_{\theta}(\theta)) := \int d\theta q_{\theta}^{(t)}(\theta) (\ln p(\theta) - \ln q_{\theta}^{(t)}(\theta)) + \int d\theta q_{\theta}^{(t)}(\theta) \int dz q_z(z) (\ln p(z, x | \theta) - \ln q_z(z))$$

The first summand is constant in q_z ; therefore it does not play a role in the argmax over q_z . Thus, our coordinate ascent update must satisfy this criterion for a critical point

$$0 = \frac{\partial}{\partial q_z} \left(\int d\theta q_{\theta}^{(t)}(\theta) \int dz q_z(z) (\ln p(z, x | \theta) - \ln q_z(z)) \right)$$

Applying the Theorem as well as the product rule yields an extremum at

$$0 = \left(\int d\theta q_{\theta}^{(t)}(\theta) (\ln p(z, x | \theta) - \ln q_z(z)) \right) - \frac{q_z}{q_z}$$

Pulling out the constant term, $\ln q_z$, rearranging, and taking the exponential yields the VB-E update.