Variational Inference

Michael Thomas Wojnowicz

November 11, 2020

Data Intensive Studies Center, Tufts University

Table of contents

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?

Overview

Overview

The problem: marginalization

Parametric statistical models

Parametric statistical models

A *parametric statistical model* posits

- $\boldsymbol{x}$: observed data
- $\boldsymbol{\theta}$: parameters
- $\boldsymbol{z}$ (possibly): latent random variables

Parameters vs. latent variables

Both $\boldsymbol{z}$ and $\boldsymbol{\theta}$ are unobserved, but only the dimensionality of $\boldsymbol{z}$ increases with the number of samples in $\boldsymbol{x}$.

Frequentist vs. Bayesian variants

Frequentists take parameters $\boldsymbol{\theta}$ to be fixed (but unknown) constants, whereas the Bayesians take $\boldsymbol{\theta}$ to be random variables.

Three statistical modeling paradigms of interest

Let us consider models that present an **intractable marginal**.

Bayesian latent variable models

Examples: Bayesian Mixture Model, Bayesian Hidden Markov Model, Latent Dirichlet Allocation, Bayesian nonparametric versions of the preceding

Bayesian (non-latent variable) models

Examples: Non-conjugate models, Many hierarchical Bayesian models

Frequentist latent variable models

Examples: Hidden Markov Models (although we have handled this case) , Variational Autoencoders (the classical kind) , Bayesian Generalized Linear Mixed Effects Models

Statistical inference and marginalization

Bayesian latent-variable models

We want the posterior:

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

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

$$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{z}) \, d\,\boldsymbol{\theta} \;\; d\,\boldsymbol{z} \tag{1.1}$$

Statistical inference and marginalization

Bayesian non-latent variable models

We want the posterior:

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

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

$$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{x}) \, d\boldsymbol{\theta} \tag{1.2}$$

Statistical inference and marginalization

Frequentist latent variable models

We want the maximum likelihood value:

$$\boldsymbol{\theta}_{\mathrm{ML}} := \mathrm{argmax}_{\boldsymbol{\theta}}\ p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{\theta}} \int p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\, d\,\boldsymbol{z} \tag{1.3}$$

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

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{\theta})}_{\textit{marginal likelihood}} = \int \underbrace{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}_{\textit{joint (or "complete") likelihood}} d\,\boldsymbol{z} \tag{1.4}$$

Statistical inference

In general

We must compute the marginal

$$p(\boldsymbol{x} \mid \boldsymbol{c}) = \int p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c}) \, d\, \boldsymbol{u} \tag{1.5}$$

where

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

The need for marginalization in statistical inference

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

Problem: These marginalizations may be intractable

Example: Hidden Markov Model

Define T : the state transition matrix

ϵ_j : the jth emission distribution, $j = 1, ..., k$

π : the initial latent state distribution

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

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

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

Overview

The technique: functional optimization

Towards variational inference

We construct a lower bound on the target marginal.

Variational Lower Bound (VLBO)

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

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

Variational Inference: Maximizing the VLBO

Variational Inference

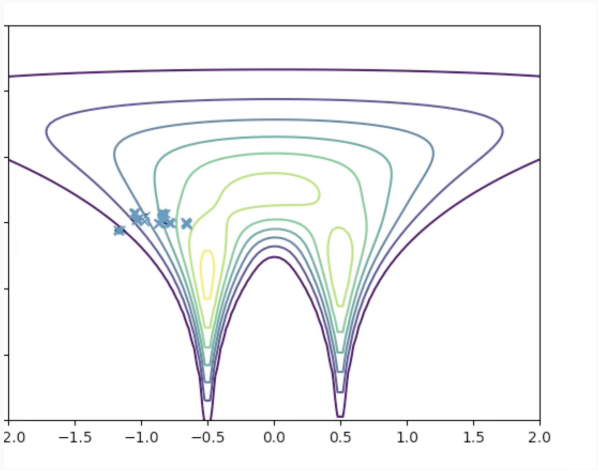
Variational inference (VI) proceeds by finding q^*, the variational density in tractable family $\mathcal{Q}$ which maximizes the VLBO:

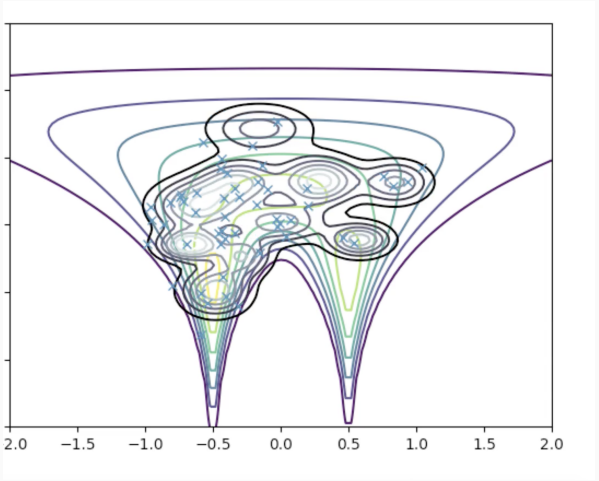
$$\underset{\textit{solution}}{q^*} = \underset{\substack{q \in \mathcal{Q} \\ \textit{approximating family}}}{\operatorname{argmax}} \texttt{VLBO}(q)$$

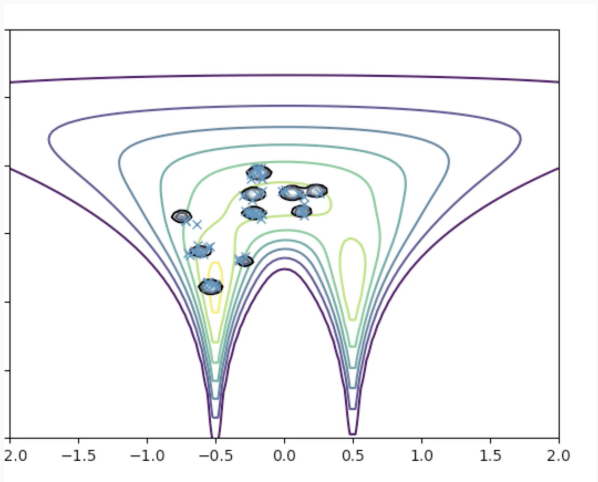
Rk: Note that we are trying to optimize over a function space (of a particular kind).

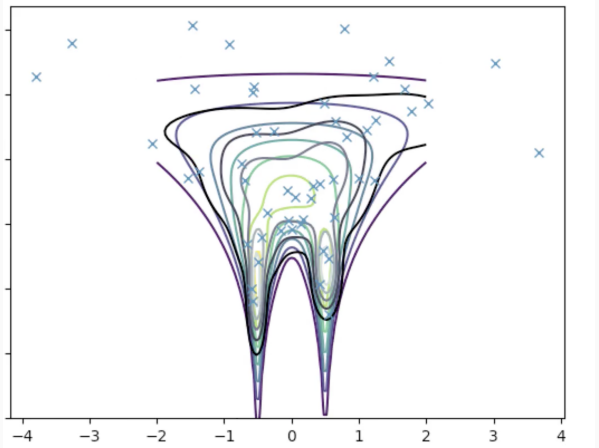
Illustration

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









Ranganath, R., Gerrish, S., & Blei, D. (2014, April). Black box variational inference. In Artificial Intelligence and Statistics (pp. 814-822).

Overview

Decompositions: Intuition on the cost function

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

$$\begin{aligned}\texttt{VLBO}\,(q) &= \int q(\boldsymbol{u})\ \ln p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c})\ d\,\boldsymbol{u} \ - \int q(\boldsymbol{u})\ \ln q(\boldsymbol{u})\ d\,\boldsymbol{u} \\ &= \underbrace{\mathbb{E}_q\big[\log p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c})\big]}_{\textit{energy}} + \underbrace{\mathbb{H}\big[q(\boldsymbol{u})\big]}_{\textit{entropy}}\end{aligned}$$

Q What is the effect of the entropy term?

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

$$\begin{aligned}
\texttt{VLBO}\,(q) &= \mathbb{E}_q\big[\log p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c})\big] + \mathbb{H}\big[q(\boldsymbol{u})\big] \\
&= \mathbb{E}_q\big[\log p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c})\big] - \mathbb{E}_q\big[\log q(\boldsymbol{u})\big] \\
&= \mathbb{E}_q\big[\log p(\boldsymbol{x} \mid \boldsymbol{u}, \boldsymbol{c})\big] + \mathbb{E}_q\big[\log p(\boldsymbol{u} \mid \boldsymbol{c})\big] - \mathbb{E}_q\big[\log q(\boldsymbol{u})\big] \\
&= \underbrace{\mathbb{E}_q\big[\log p(\boldsymbol{x} \mid \boldsymbol{u}, \boldsymbol{c})\big]}_{\textit{expected log likelihood}} - \underbrace{\texttt{KL}\big(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u} \mid \boldsymbol{c})\big)}_{\textit{divergence from prior}}
\end{aligned}$$

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

Overview

The posterior perspective

Maximizing the VLBO minimizes the KL divergence (to the posterior)

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

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

By the chain rule, we get

$$\begin{aligned} \mathrm{KL}(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})) &= \underbrace{\mathbb{E}_q[\log q(\boldsymbol{u})] - \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c})]}_{\text{energy/entropy decomposition}} + \log p(\boldsymbol{x} \mid \boldsymbol{c}) \\ &= -\ \mathrm{VLBO}\,(q) + \text{constant} \end{aligned}$$

Discuss: What is the optimal variational density?

Maximizing the VLBO minimizes the KL divergence (to the posterior)

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

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

By the chain rule, we get

$$\begin{aligned}\mathtt{KL}(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})) &= \underbrace{\mathbb{E}_q[\log q(\boldsymbol{u})] - \mathbb{E}_q[\log p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c})]}_{\text{energy/entropy decomposition}} + \log p(\boldsymbol{x} \mid \boldsymbol{c}) \\ &= -\ \mathtt{VLBO}\,(q) + \text{constant}\end{aligned}$$

The optimal variational density

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

Summary

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

Model	Inferential goal	Intractable marginal	Variational density	Posterior
General case	infer about $\boldsymbol{\theta}$	$p(\boldsymbol{x} \mid \boldsymbol{c})$	$q(\boldsymbol{u})$	$p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})$
Frequentist latent	$\operatorname{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{x} \mid \boldsymbol{\theta})$	$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, d\boldsymbol{z}$	$q(\boldsymbol{z})$	$p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta})$
Bayesian (non-latent)	$p(\boldsymbol{\theta} \mid \boldsymbol{x})$	$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{x}) \, d\boldsymbol{\theta}$	$q(\boldsymbol{\theta})$	$p(\boldsymbol{\theta} \mid \boldsymbol{x})$
Bayesian latent	$p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{x})$	$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{z}) \, d\boldsymbol{\theta} \; d\boldsymbol{z}$	$q(\boldsymbol{z}, \boldsymbol{\theta})$	$p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{x})$

How does VI accommodate the goal of statistical inference?

Given selection of variational family $\mathcal{Q}$, the optimal variational density q^* ...

The marginal perspective

- *For frequentist models*: ... makes the VLBO best approximate the target marginal likelihood, $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, which is what we wanted to maximize.
- *For Bayesian models*: ... raises the (approximate) evidence term $p(\boldsymbol{x})$ (the term used for Bayesian model comparison) as high as possible.

The posterior perspective

- *For Bayesian models*: ... is the family member which is closest to the target posterior $p(\boldsymbol{u} \mid \boldsymbol{x})$.
- *For frequentist models*: ... provides the best substitution $q^*(\boldsymbol{z}) \approx p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{\text{curr}})$ into the E-step of the EM algorithm[1]

[1] See next section for more information.

Overview

Evaluation (in context)

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.

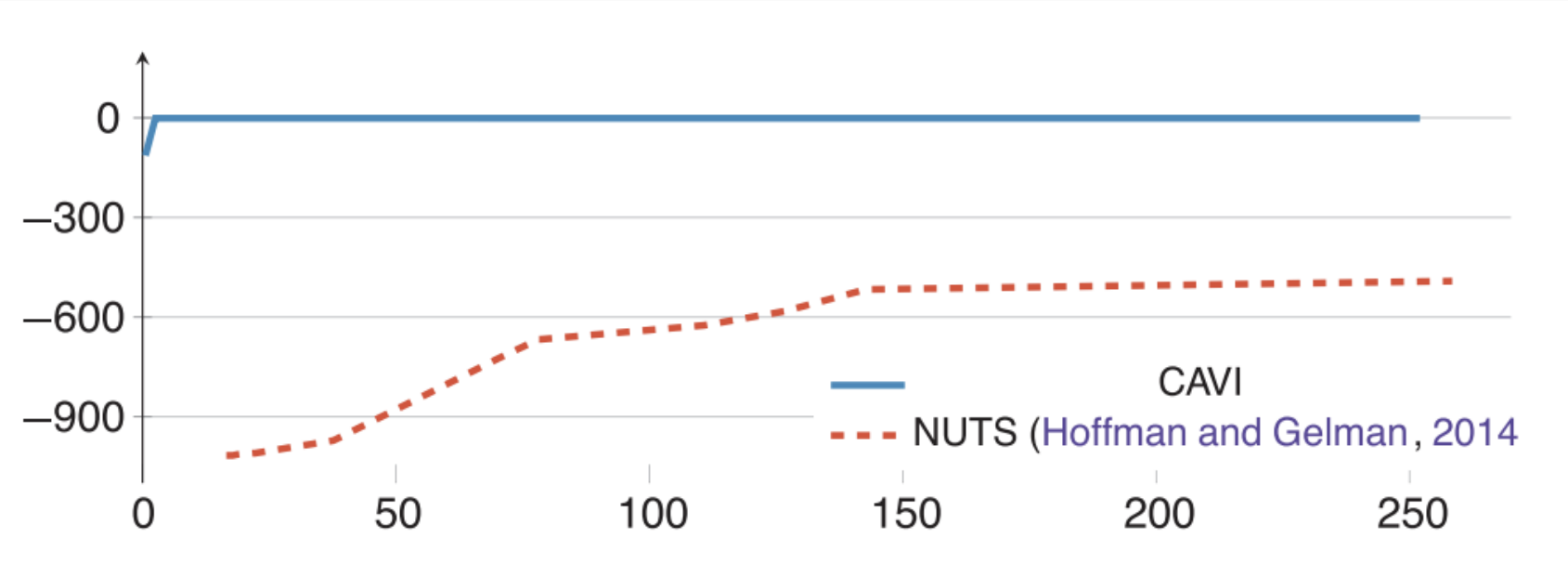
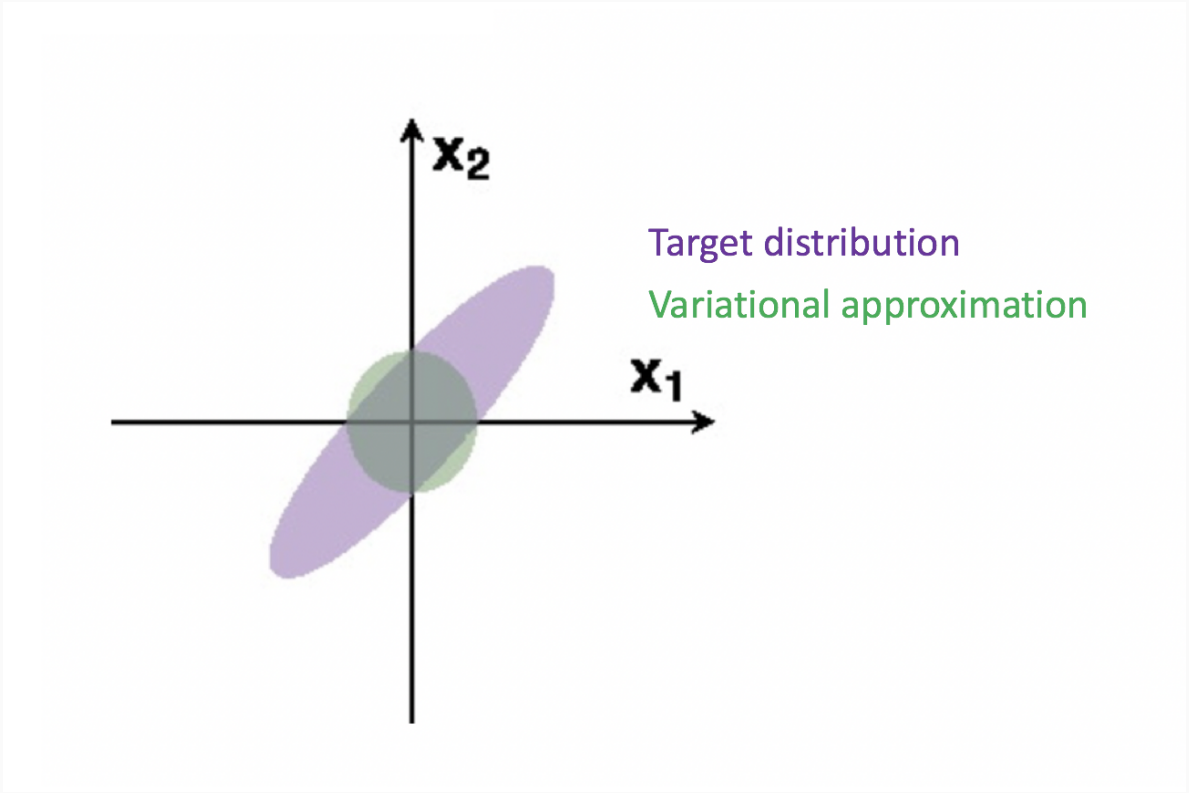


Figure 1: Comparison of CAVI to a Hamiltonian Monte Carlo-based sampling technique. The plot shows log predictive test set accuracy by training time (minutes). CAVI fits a Gaussian mixture model to 10,000 images in less than a minute.

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.

Shortcoming: VI underestimates variance of the true posterior



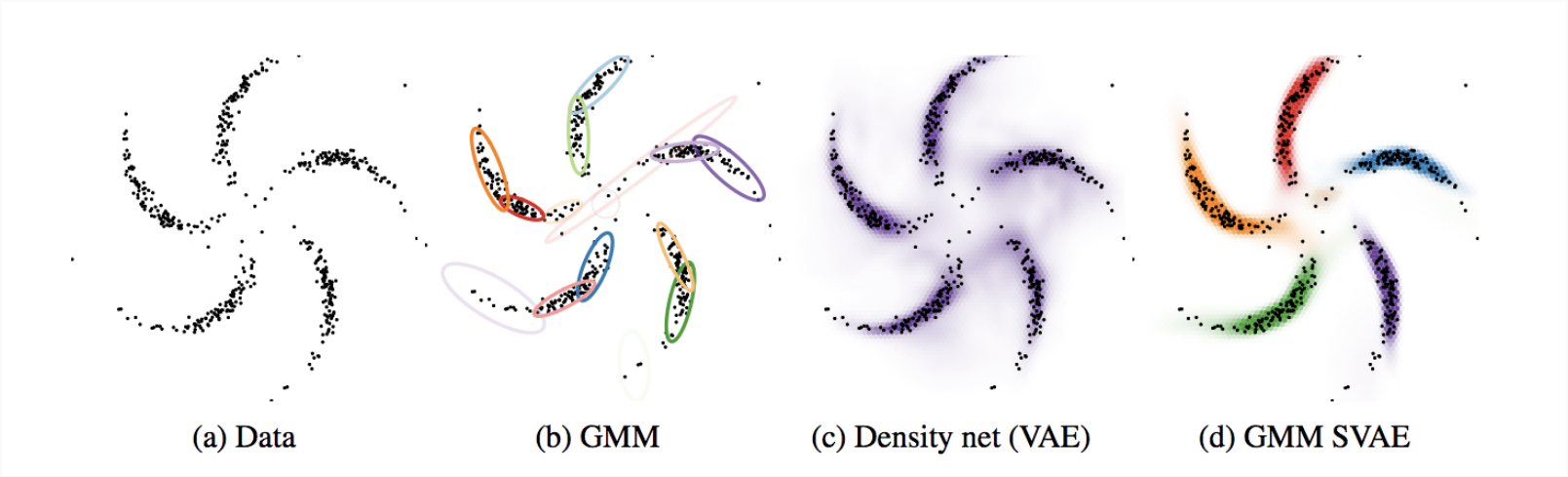
$$\mathrm{KL}(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})) = \mathbb{E}_q \left[\log \frac{q(\boldsymbol{u})}{p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})} \right]$$

Intuition

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

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.

Johnson, M., Duvenaud, D. K., Wiltschko, A., Adams, R. P., & Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. In Advances in neural information processing systems (pp. 2946-2954).

Variational Inference and Expectation Maximization

Expectation Maximization (EM)

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

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

until convergence to a local optimum.

Example: Exponential Hidden Markov Model

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

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

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

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 $\boldsymbol{\theta}^{(0)}$ via the recursion

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

until convergence to a local optimum.

Example: Gaussian Hidden Markov Model

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

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

$$\widehat{\boldsymbol{\mu}}_k^{(t)} = \frac{\sum_i (p_i = k)\, \boldsymbol{x}_i}{\sum_i (p_i = k)}, \qquad \widehat{\boldsymbol{\Sigma}}_k^{(t)} = \frac{\sum_i (p_i = k)\, (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}^{(t)})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}^{(t)})^T}{\sum_i (p_i = k)}$$

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

$$\texttt{VLBO}\,(q_{\boldsymbol{z}}, \theta) = \mathbb{E}_q\big[\log p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\big] + \mathbb{H}\big[q(\boldsymbol{z})\big]$$

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

$$\textbf{q update}: \quad q_{\boldsymbol{z}}^{(t+1)} = \text{argmax}_{q_z}\ \texttt{VLBO}\,(q_{\boldsymbol{z}}; \boldsymbol{\theta}^{(t)}) \tag{2.1}$$

$$\boldsymbol{\theta}\ \textbf{update}: \quad \boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}}\ \texttt{VLBO}\,(q_{\boldsymbol{z}}^{(t+1)}; \boldsymbol{\theta}) \tag{2.2}$$

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

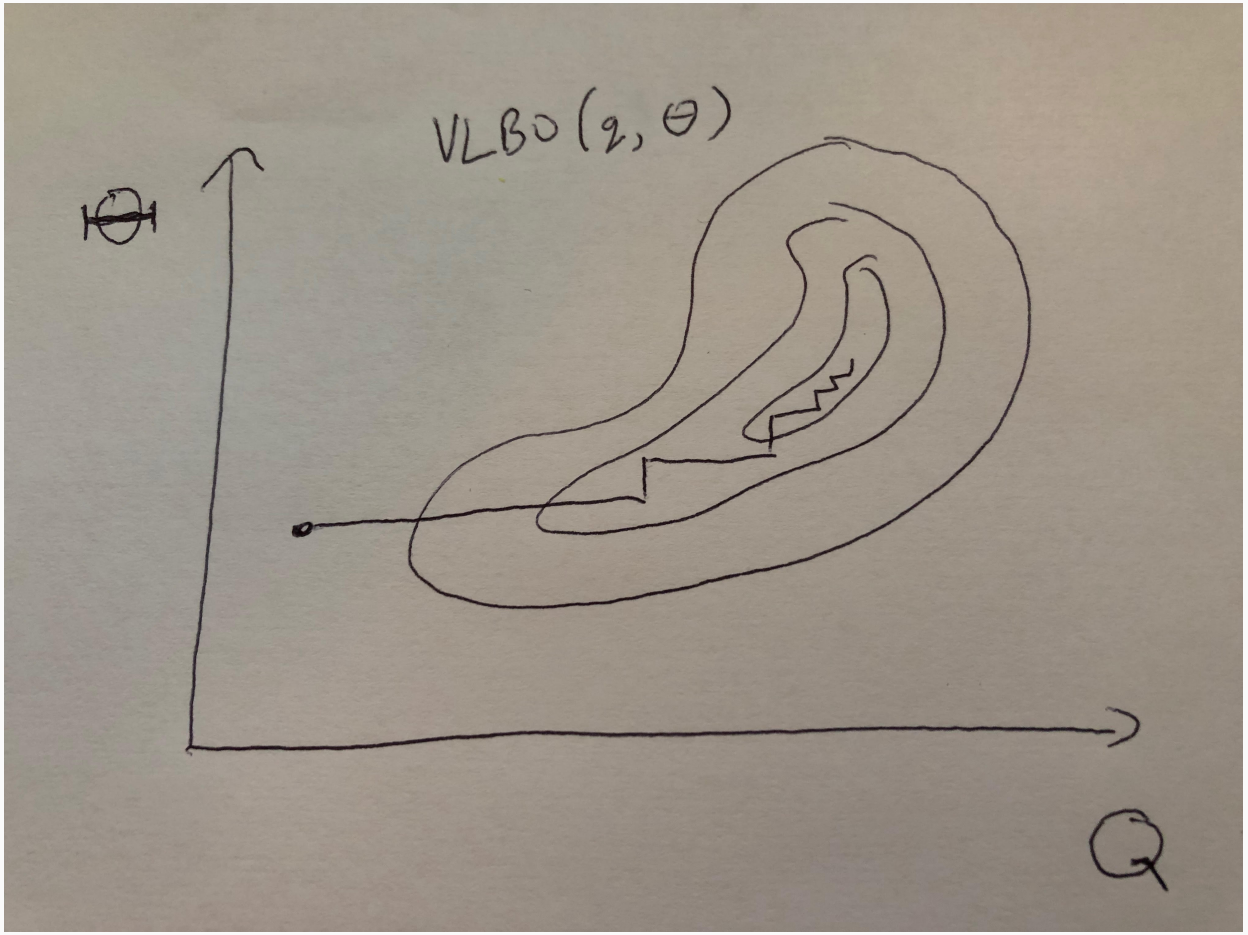
$$q_{\boldsymbol{z}}^{(t+1)} = p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

in which case the $\boldsymbol{\theta}$ *update* becomes

$$\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}}\ \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}\Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big] \tag{2.3}$$

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 $p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta})$, we may restrict our variational family $\mathcal{Q}$ to some convenient form. In this case, coordinate ascent on the $\texttt{VLBO}$ is given by:

$$q_{\boldsymbol{z}}^{(t+1)} = \text{argmax}_{q_{\boldsymbol{z}} \in \mathcal{Q}} \ \texttt{VLBO}\left(q_{\boldsymbol{z}}; \boldsymbol{\theta}^{(t)}\right)$$

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

which generalizes the EM algorithm.

Variational Bayes Expectation Maximization (VBEM)

Consider a Bayesian latent variable model. So we need to swap $p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta})$ for $p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})$ in the VLBO.

If we construct the variational density with the factorization

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

then the VLBO becomes

$$\texttt{VLBO}\left(q_{\boldsymbol{z}}(\boldsymbol{z}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) := \int \int q_{\boldsymbol{z}}(\boldsymbol{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ \ln \left(\frac{p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{x})}{q_{\boldsymbol{z}}(\boldsymbol{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} \ \ d\boldsymbol{z} \tag{2.4}$$

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

$$\begin{aligned} \textbf{VB-E step}: \quad & q_{\boldsymbol{z}}^{(t+1)} = \operatorname{argmax}_{q_{\boldsymbol{z}}} \ \texttt{VLBO}\left(q_{\boldsymbol{z}} \,;\, q_{\boldsymbol{\theta}}^{(t)}\right) \\ \textbf{VB-M step}: \quad & q_{\boldsymbol{\theta}}^{(t+1)} = \operatorname{argmax}_{q_{\boldsymbol{\theta}}} \ \texttt{VLBO}\left(q_{\boldsymbol{z}}^{(t+1)} \,;\, q_{\boldsymbol{\theta}}\right) \end{aligned}$$

VBEM: Derivation

See notes.

VBEM: Update Equations

The coordinate ascent equations have the form

$$\textbf{VB-E step}: \quad q_{\boldsymbol{z}}^{(t+1)} \propto \exp\left(\mathbb{E}_{q_{\boldsymbol{\theta}}^{(t)}}\left[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\right]\right) \tag{2.5}$$

$$\textbf{VB-M step}: \quad q_{\boldsymbol{\theta}}^{(t+1)} \propto p(\boldsymbol{\theta}) \exp\left(\mathbb{E}_{q_{\boldsymbol{z}}^{(t)}}\left[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\right]\right) \tag{2.6}$$

Prior-likelihood decomposition

Bayes' rule

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

VB-M update

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

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 $p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta})$.
- 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)

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

$$q(u_1, ..., u_K) = \prod_{k=1}^{K} q_k(u_k) \tag{3.1}$$

Mean field coordinate ascent variational inference (MF-CAVI) is CAVI performed under the mean field assumption (3.1).

Update equations for MF-CAVI

To perform coordinate ascent on the `VLBO` under the mean field assumption (3.1), we iteratively update our variational factors $\{q_k\}_k$ via

$$q_k(u_k) \propto \exp\left\{\mathbb{E}_{q_{-k}}\left[\log p(u_k \mid \boldsymbol{u}_{-k}, \boldsymbol{x}, \boldsymbol{c})\right]\right\} \tag{3.2}$$

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

Example: Bayesian Gaussian Mixture Model

Example: Bayesian Gaussian Mixture Model

To see the mean field CAVI algorithm (3.2) in a concrete context, consider a version of the Bayesian Gaussian Mixture Model.

$$\begin{aligned} \mu_k &\sim \text{Normal}(M_k = 0, V_k = \sigma^2) && k = 1, ..., K \\ c_i &\sim \text{Categorical}(\pi_1, ..., \pi_K) && i = 1, ..., n \\ x_i \mid c_i, \mu &\sim \text{Normal}(\mu_{c_i}, 1) && i = 1, ..., n \end{aligned}$$

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

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

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 kth 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

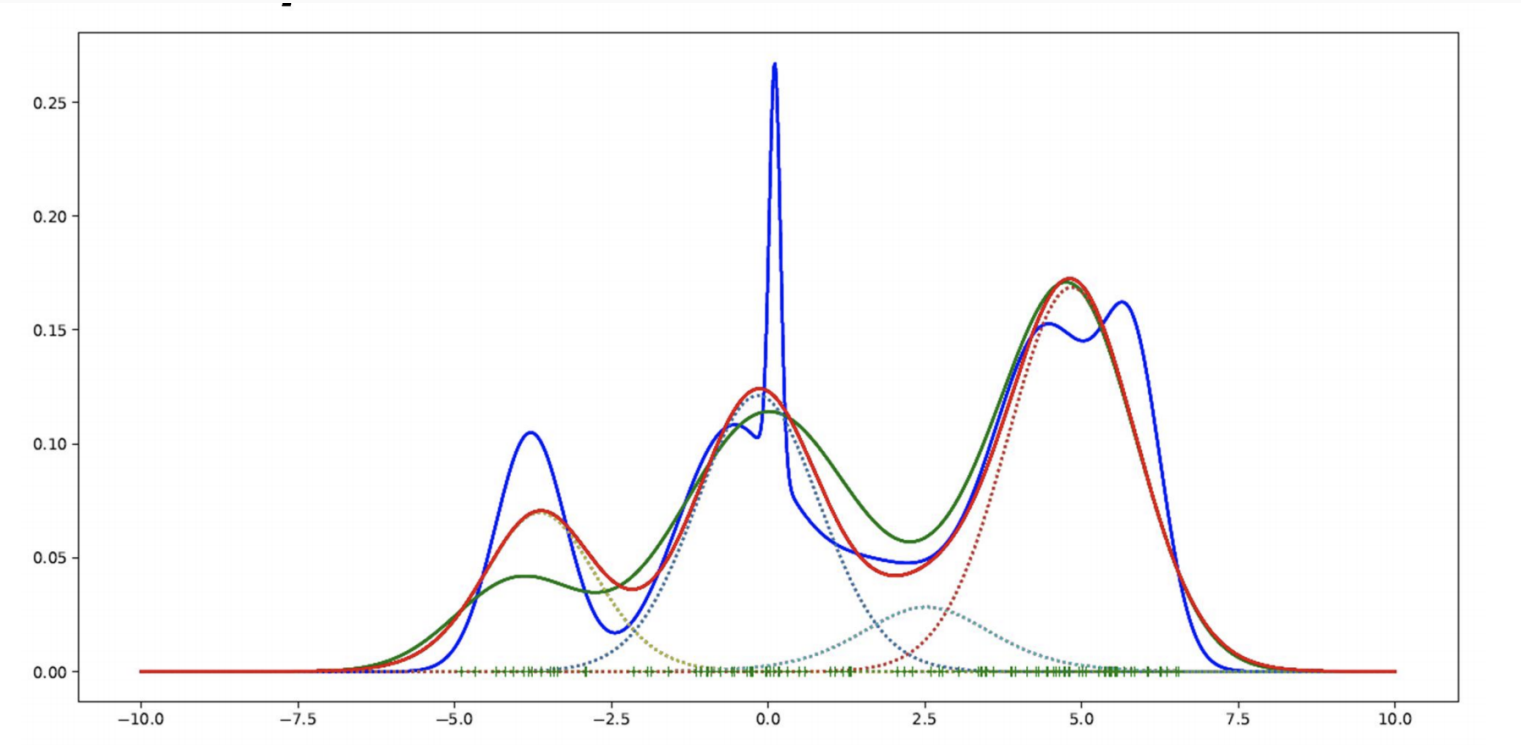


Image Credit: Lukas Burget

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

Tool: CAVI with Exponential Family Complete Conditionals

Motivation

The form of (3.2) 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

$$p(x \mid \eta) = h(x) \exp\{\eta^T s(x) - a(\eta)\} \tag{5.1}$$

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: $x \perp\!\!\!\perp \theta \mid t(x)$

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

Claim

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

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

Then optimal mean field CAVI update (3.2) puts

$$q_k \in \mathcal{E} \tag{5.2}$$

with natural parameter

$$\boxed{\nu_k = \mathbb{E}_{q_{-k}}[\eta_k(u_{-k}, x)]} \tag{5.3}$$

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 kth complete conditional is in the exponential family, then we have

$$p(u_k \mid u_{-k}, x) = h(u_k) \exp\{\eta_k^T s(u_k) - a(\eta_k)\} \tag{5.4}$$

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

By substituting (5.4) into (3.2) 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 (5.4), the claim holds.

Example: Latent Dirichlet Allocation (LDA)

Acknowledgements

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

Overview

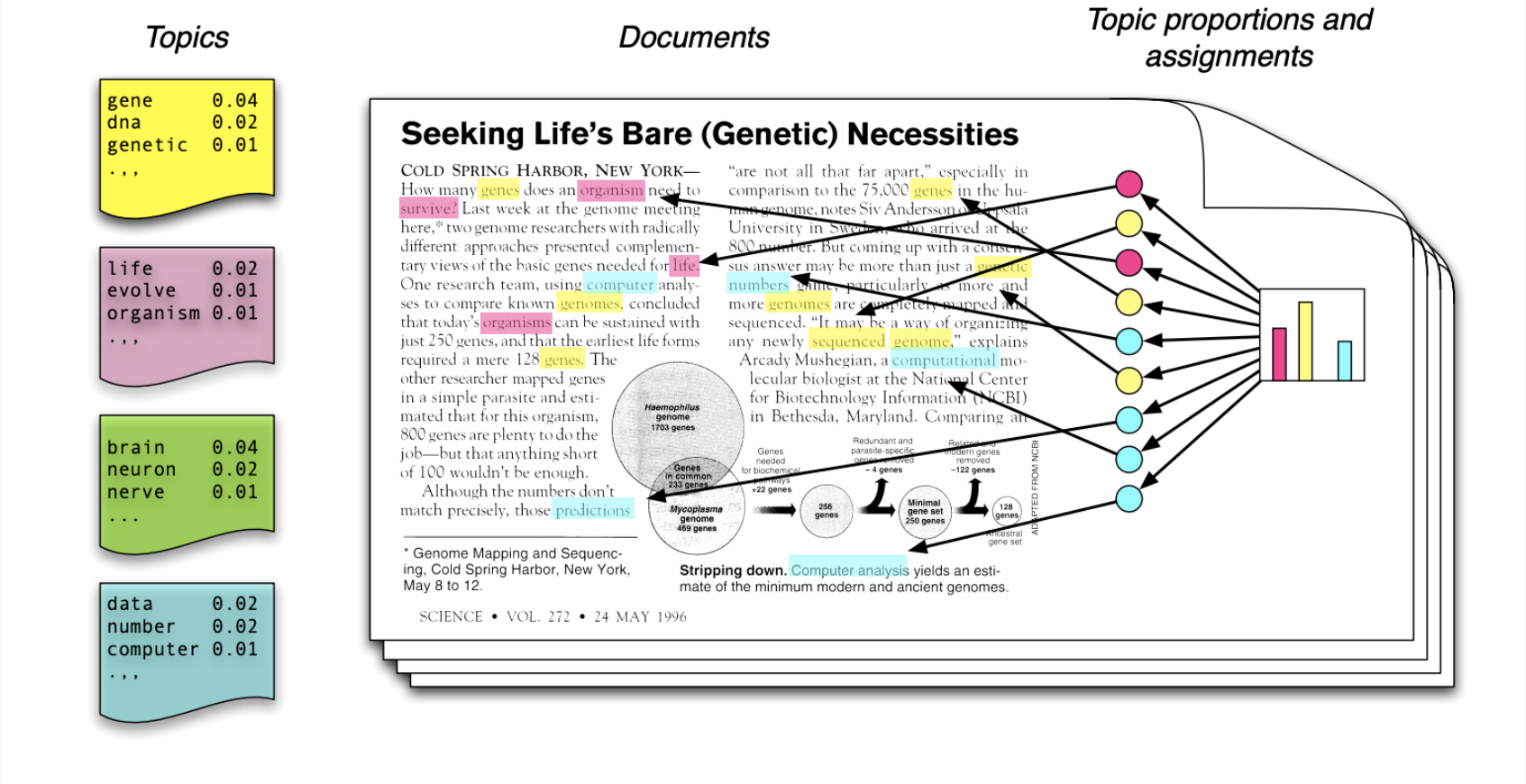
LDA is a generative probabilistic model of a corpus of documents of text.

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.).

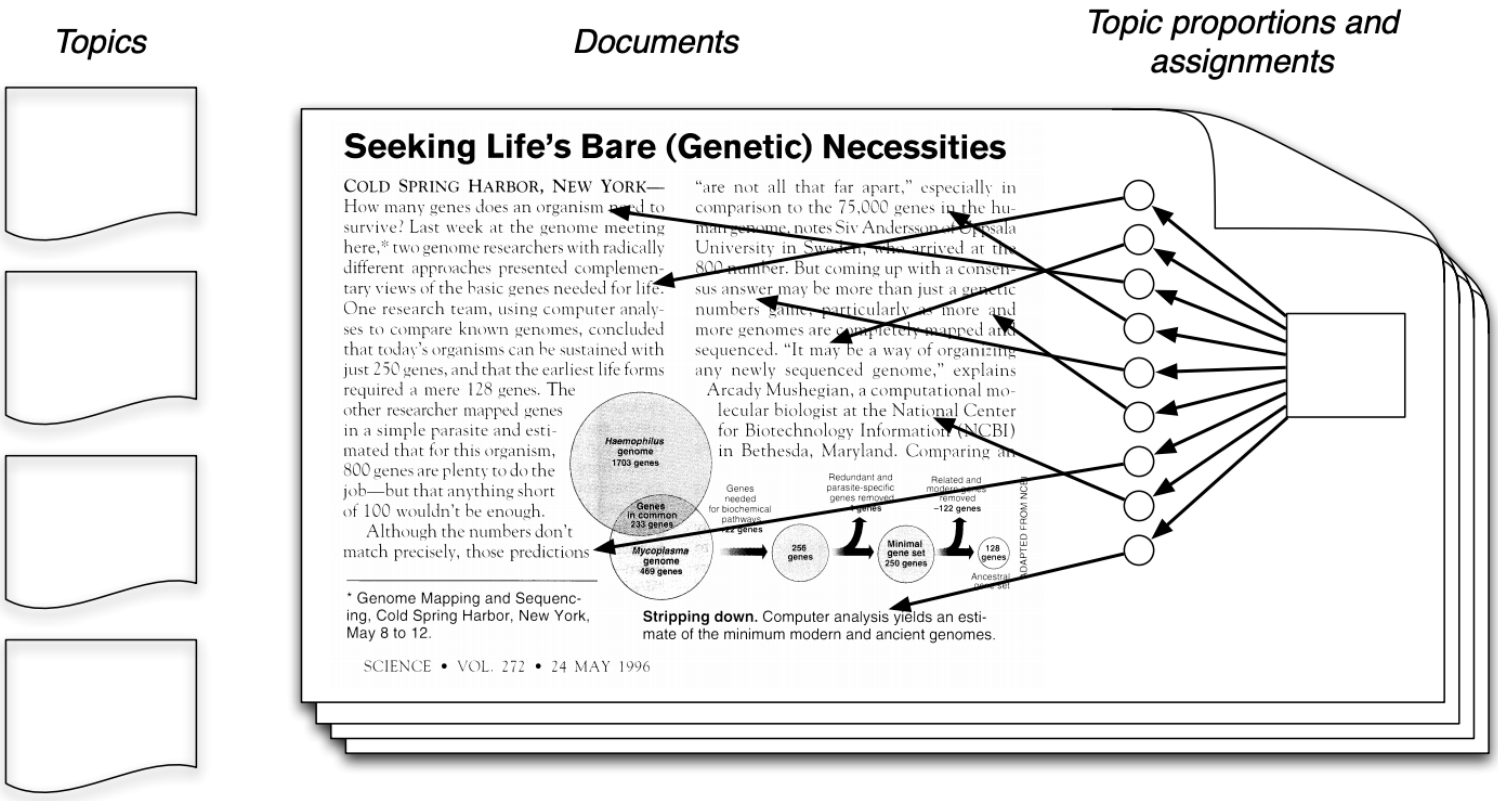
So:

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

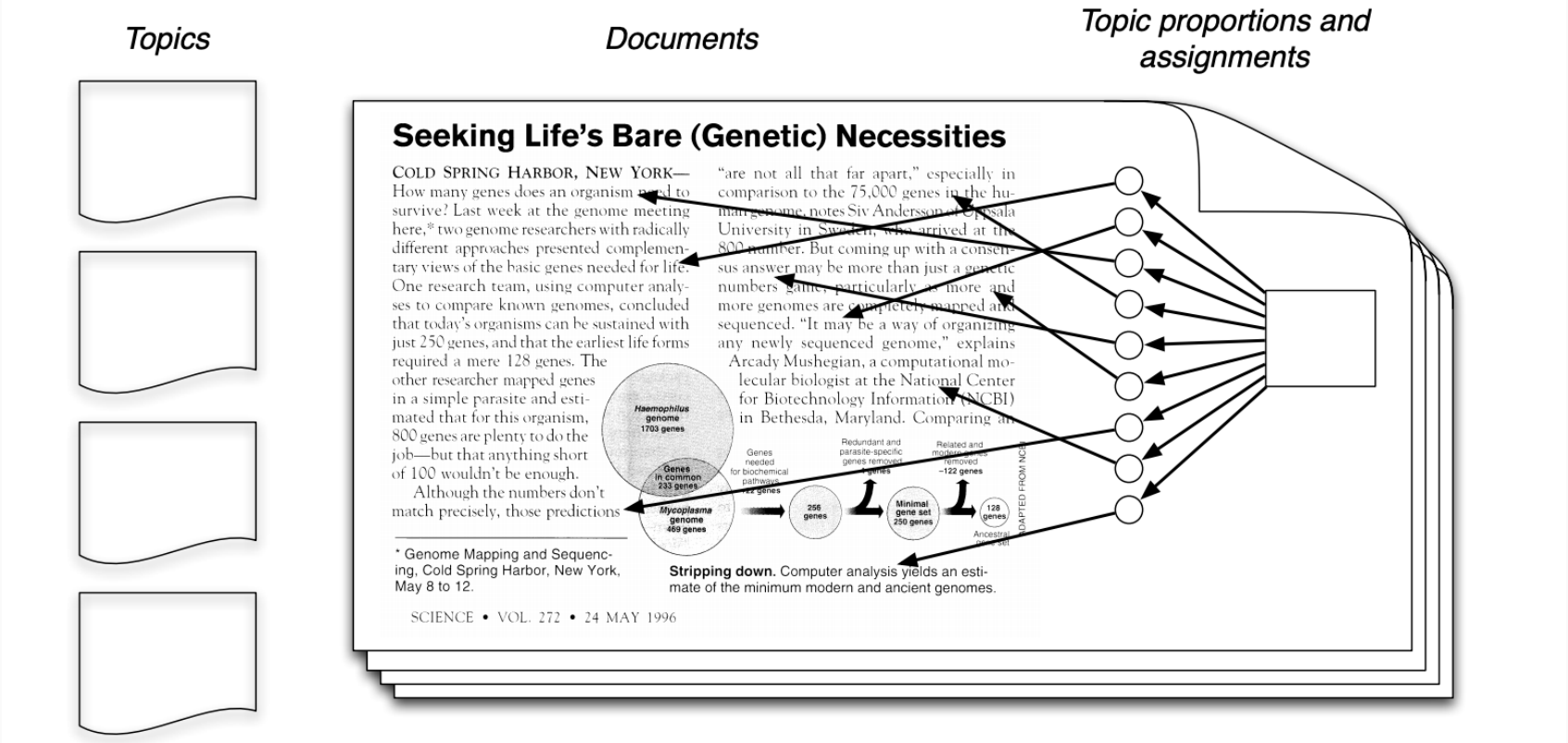


- Each **topic** is a distribution over words.
- Each **document** is a mixture of corpus-wide topics.
- Each **word** is drawn from one of those topics.

Source: David Blei, 2012 ICML Tutorial



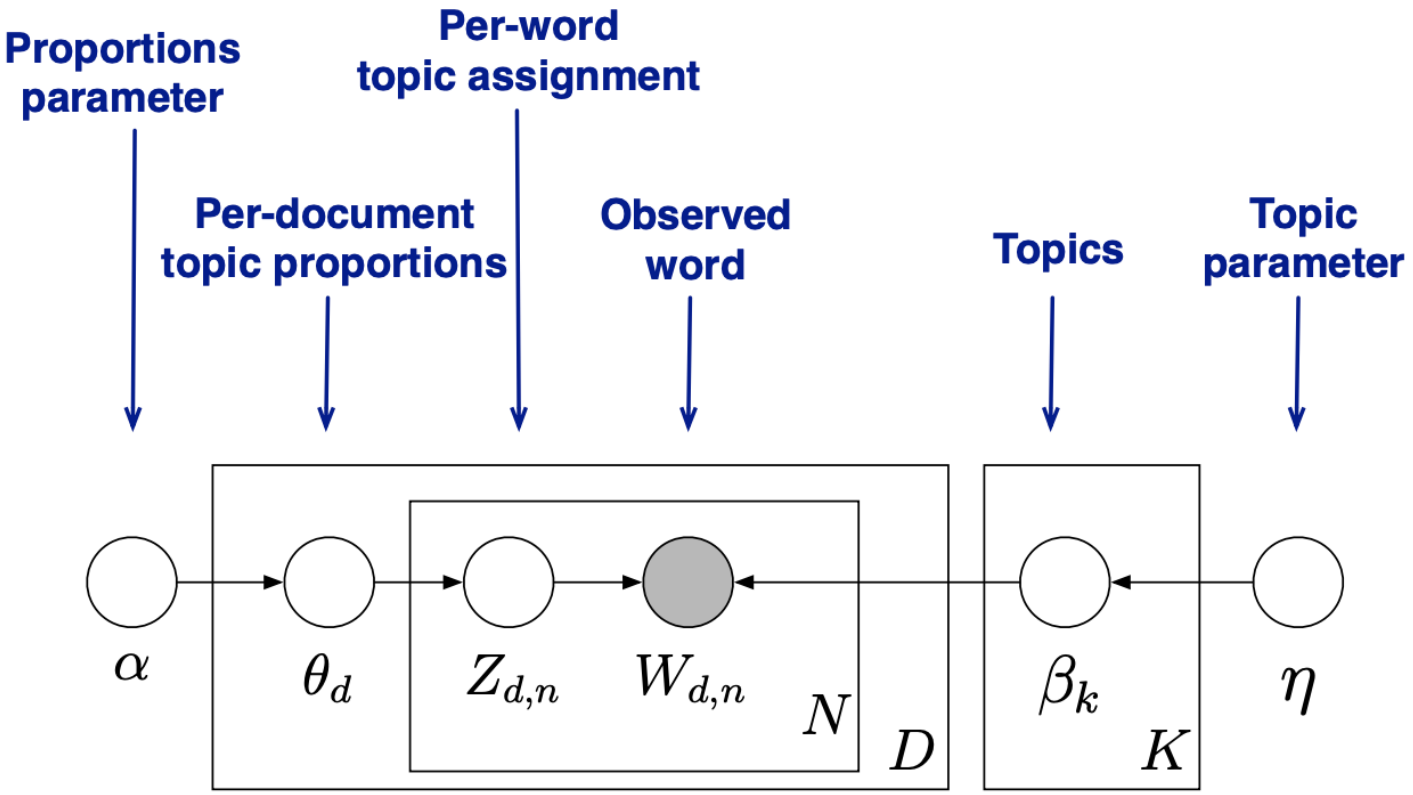
- In reality, we only observe the documents.
- The other structure is **hidden variables**.



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

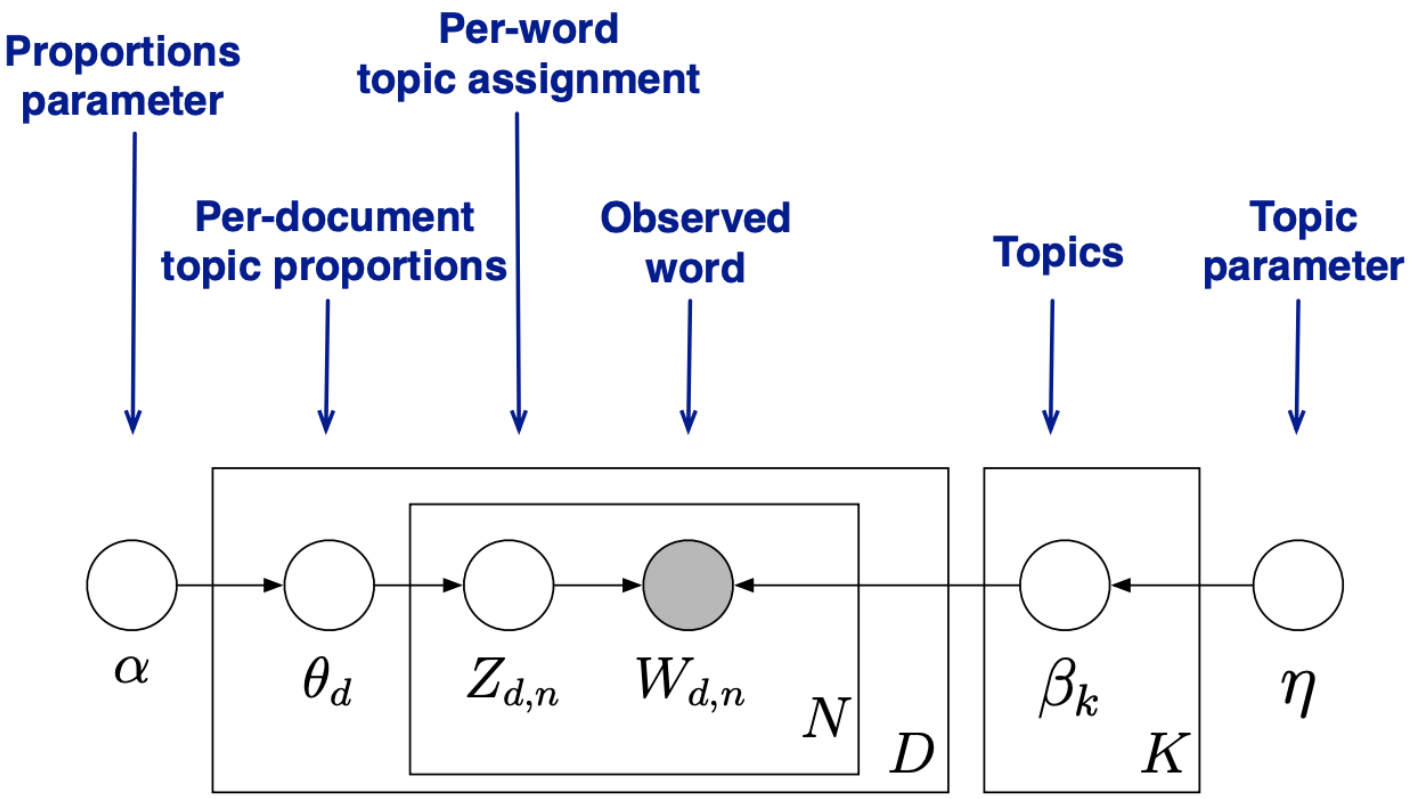
$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

LDA as a graphical model



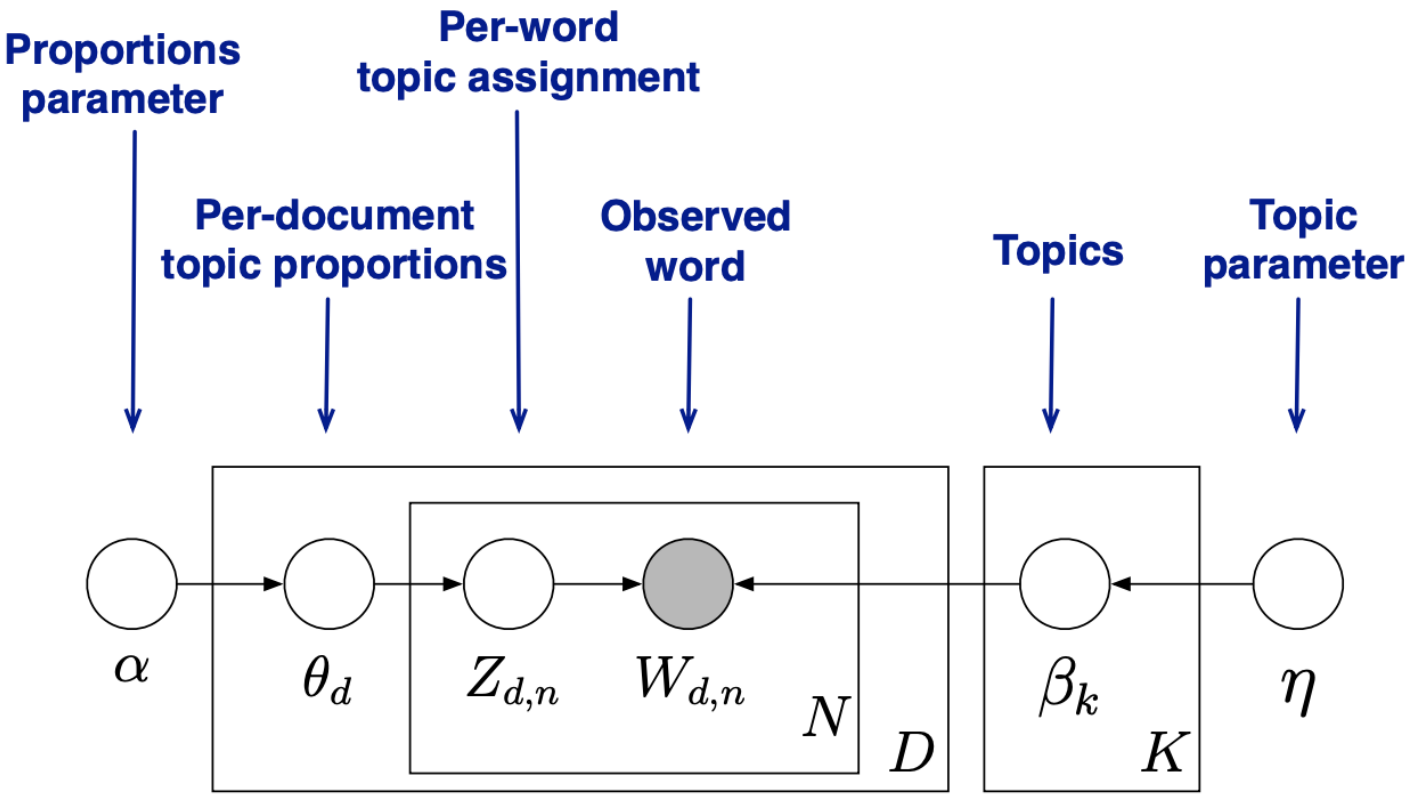
- Encodes assumptions
- Defines a factorization of the joint distribution

LDA as a graphical model



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

Joint distribution



$$p(\boldsymbol{z}, \theta, \boldsymbol{w}, \beta \mid \alpha, \eta) = \prod_{k=1}^{K} \underbrace{p(\beta_k \mid \eta)}_{\textit{Dirichlet}} \prod_{d=1}^{D} \underbrace{p(\theta_d \mid \alpha)}_{\textit{Dirichlet}} \prod_{n=1}^{N} \underbrace{p(\boldsymbol{z}_n \mid \theta)}_{\textit{Categorical}} \underbrace{p(\boldsymbol{w}_n \mid \boldsymbol{z}_n, \beta)}_{\textit{Categorical}} \quad (6.1)$$

LDA: Generative Process

- Choose the vocabulary of V words, and set the number of topics, K.
- Set hyperparameters $\eta \in \mathbb{R}^V, \alpha \in \mathbb{R}^K$.
- For k in $(1, K)$:
  - Choose *per-topic word distribution* $\beta_k \in \mathbb{R}^V \sim \text{Dir}(\eta)$.[2]
- For d in $(1, D)$:
  - Choose *per-document topic distribution* $\theta_d \in \mathbb{R}^K \sim \text{Dir}(\alpha)$.[3]
  - For j in $(1, N_d)$:
    - Choose *topic* $z_{d,n} \sim \text{Categorical}_K(\theta_d)$
    - Choose *word* $w_{d,n} \sim \text{Categorical}_V(\beta_{z_{d,n}})$

[2] Interpretation of η: psuedocount of vocabulary words observed across topics.

[3] Interpretation of α: psuedocount of topics observed across documents.

Aside: Dirichlet Distribution

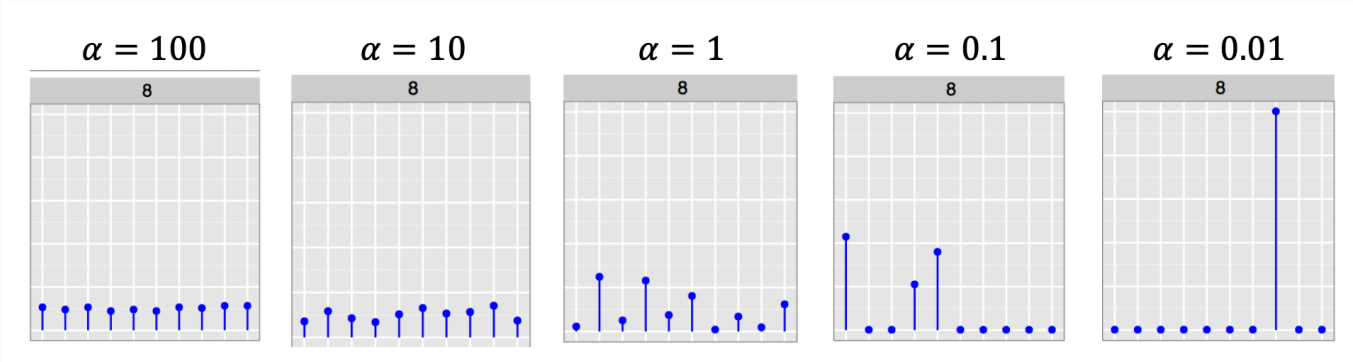
- Dirichlet distribution is *conjugate prior* of Multinomial

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

- The parameter α controls the shape and sparsity of the θ_d's.

(per-document topic distribution)

  - high α: typical θ_d will be uniform
  - small α: a typical θ_d will be sparse

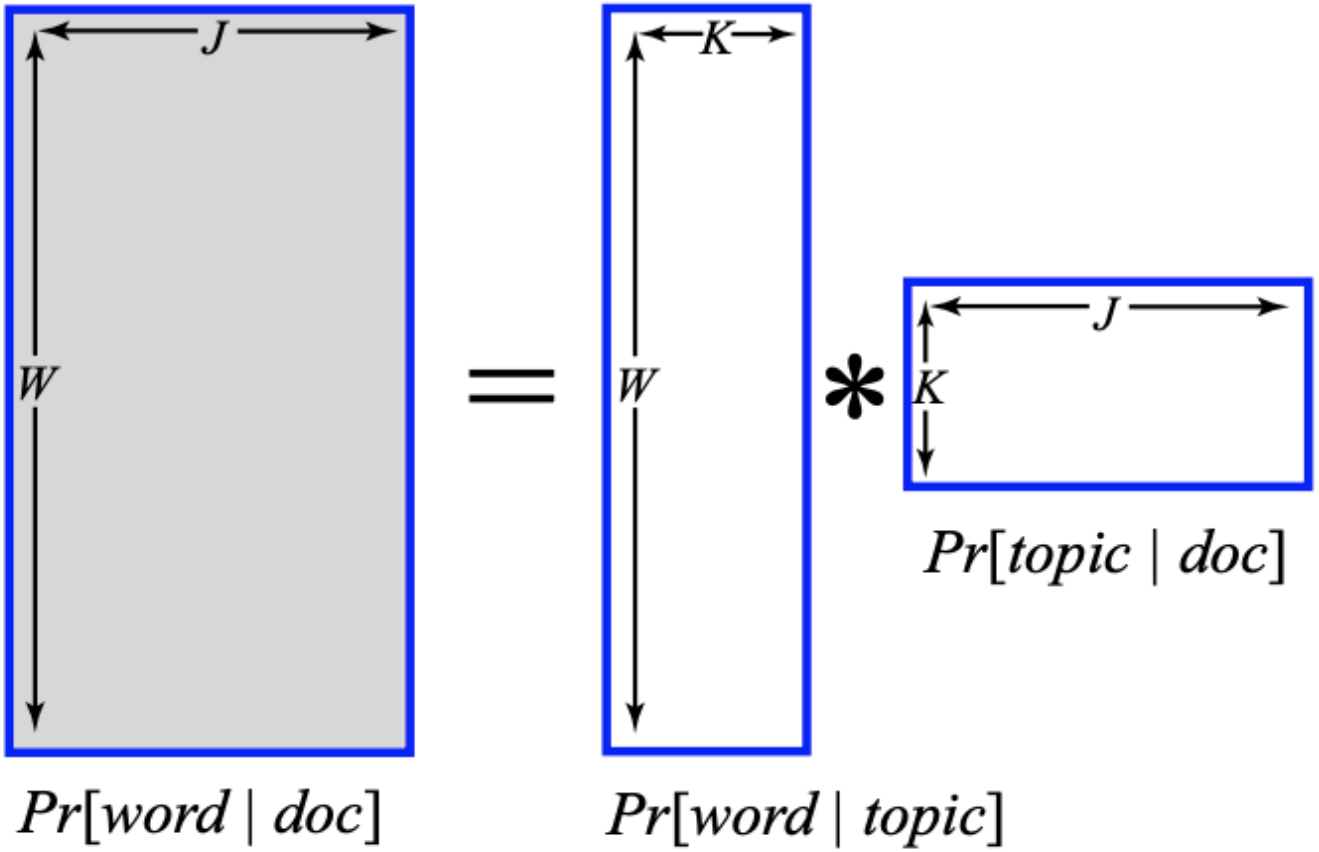


- Likewise, η controls the shape and sparsity of the β_k's (per-topic word distribution)

LDA Example Inference

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

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.

Image Credit : Erik Sudderth

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*, $\boldsymbol{w}_n \in \text{OneHot}(V)$ indicates which element of the vocabulary was observed as the nth word of the document.[4]
- A *document*, $\boldsymbol{w} \in \text{OneHot}(V)^N$, is a list of N words. It is represented as a $N \times V$ matrix.

[4] We define 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 Multinoulli(K) distribution.

Definitions: Hidden Variables

- A *topic indicator* is an integer in $\{1, ...K\}$.
- A *(per-word) topic assignment*, $\boldsymbol{z}_n \in \text{OneHot}(K)$ indicates which topic generated the nth word in a document
- The *(per-word) topic assignments*, $\boldsymbol{z} \in \text{OneHot}(K)^N$, 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*, $\boldsymbol{\theta}$, is a (document-specific) probability distribution over the topics.

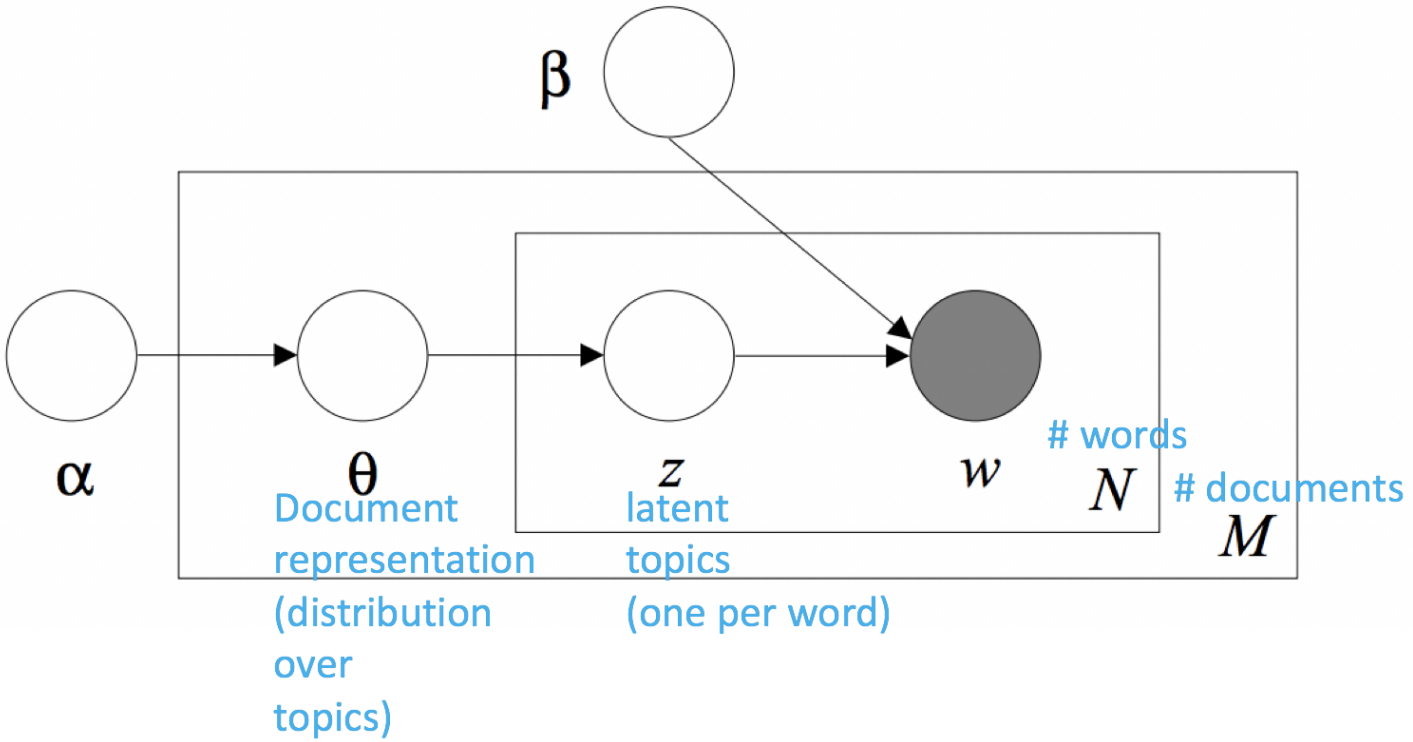
Hyperparameters

- $\boldsymbol{\alpha} \in (\mathbb{R}^+)^K$ is interpreted as the *prior counts of topic indicators*.
- $\boldsymbol{\beta} \in (\Delta^{V-1})^K$ is interpreted as the *topic-conditional word probabilities*, In particular, note that, by construction $\boldsymbol{\beta}$ 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 $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ in a corpus, the generative model is

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

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

$$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})}_{\textit{Dirichlet}} \prod_{n=1}^{N} \underbrace{p(\boldsymbol{z}_n \mid \boldsymbol{\theta})}_{\textit{Multinoulli}} \underbrace{p(\boldsymbol{w}_n \mid \boldsymbol{z}_n, \boldsymbol{\beta})}_{\textit{Multinoulli}} \tag{6.2}$$

Definition

([·]) We define the operator $[\cdot] : \textit{OneHot}(K) \rightarrow \{1, ..., K\}$ 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

$$\log p(\boldsymbol{z}_n \mid \boldsymbol{\theta}) = \log \boldsymbol{\theta}_{[\boldsymbol{z}_n]} \tag{6.3}$$

$$\log p(\boldsymbol{w}_n \mid \boldsymbol{z}_n, \boldsymbol{\beta}) = \log \boldsymbol{\beta}_{[\boldsymbol{z}_n],[\boldsymbol{w}_n]} \tag{6.4}$$

Variational distribution

We approximate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{w})$ using mean field variational inference (3.1). In particular, we assume that the variational family $\mathcal{Q}$ factorizes as

$$\begin{aligned} q &= q_{\boldsymbol{\delta}}(\boldsymbol{\theta})\, q_{\boldsymbol{\tau}}(\boldsymbol{z}) \\ &= \underbrace{q_{\boldsymbol{\delta}}(\boldsymbol{\theta})}_{\text{Dirichlet}} \prod_{n=1}^{N} \underbrace{q_{\boldsymbol{\tau}_n}(\boldsymbol{z}_n)}_{\text{Multinoulli}} \end{aligned} \tag{6.5}$$

Update equations

LDA coordinate ascent update equations

$$\underset{\substack{\text{var. multinoulli (topics-for-word)}}}{\tau_{n,k}} \propto \left(\Psi(\delta_k) - \Psi(\sum_j \delta_j)\right) \beta_{k,[w_n]} \tag{6.6}$$

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

$$\underset{\substack{\text{var. dirichlet (documents)}}}{\delta_k} = \underset{\text{prior counts}}{\alpha_k} + \sum_{n=1}^{N} \underset{\text{var. prob of topics}}{\tau_{n,k}} \tag{6.7}$$

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

$$\mathbb{E}_{q_\delta(\theta)}\left[\log \boldsymbol{\theta}_i\right] = \Psi(\delta_i) - \Psi(\sum_k \delta_k) \tag{6.8}$$

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.

Summary

Evaluation of CAVI

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.

Notes!

Questions?