Variational Inference

Michael Thomas Wojnowicz

November 11, 2020

Data Intensive Studies Center, Tufts University

Table of contents

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.

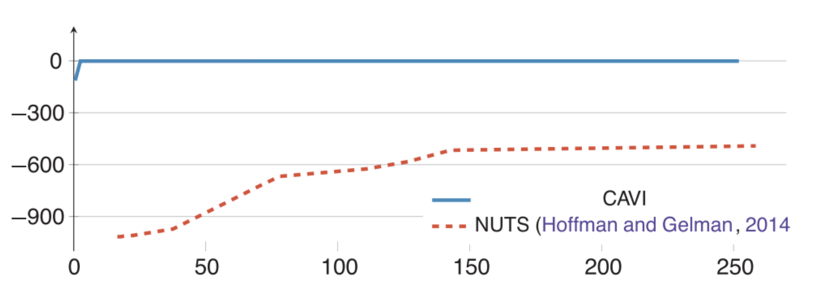
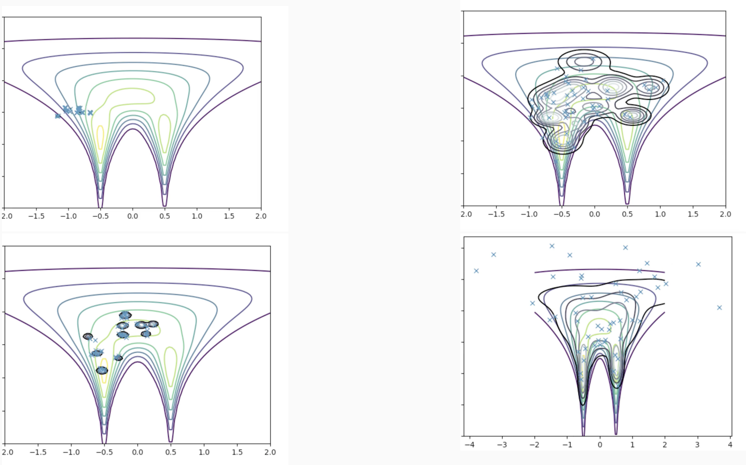


Figure 1: Comparison of CAVI to a Hamiltonian Monte Carlo-based sampling technique. The plot shows log predictive test set accuracy by training time (minutes). CAVI fits a Gaussian mixture model to 10,000 images in less than a minute.

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

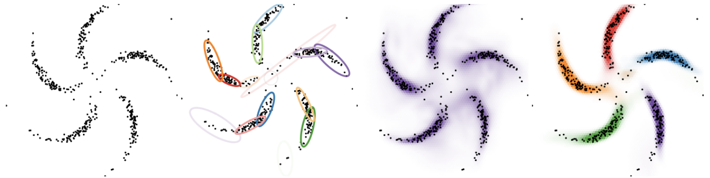
Here we approximate an probability distribution by finding the best approximation from tractable family $\mathcal{Q} = \{$10-component Gaussian mixture models$\}$



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

Modern application

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



(a) Data (b) GMM (c) Density net (VAE) (d) GMM SVAE

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: Basics

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

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:

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

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

$$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{x}) \, d\,\boldsymbol{\theta} \tag{1}$$

Statistical inference

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:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{z}) \, d\,\boldsymbol{\theta} \;\; d\,\boldsymbol{z} \tag{2}$$

Statistical inference

Frequentist latent variable models

We want the maximum likelihood value:

$$\boldsymbol{\theta}_{\mathrm{ML}} := \mathrm{argmax}_{\boldsymbol{\theta}}\; p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{\theta}} \int p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\; d\,\boldsymbol{z} \tag{3}$$

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

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{\theta})}_{\textit{marginal likelihood}} = \int \underbrace{p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})}_{\textit{joint (or "complete") likelihood}} \; d\,\boldsymbol{z} \tag{4}$$

Statistical inference

In general

We must compute the marginal

$$p(\boldsymbol{x} \mid \boldsymbol{c}) = \int p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c}) \, d\, \boldsymbol{u} \tag{5}$$

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	$\text{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{x} \mid \boldsymbol{\theta})$	$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\, d\,\boldsymbol{z}$

Problem: These marginalizations may be intractable

Example: Hidden Markov Model

$$
\begin{aligned}
\text{Define } T &: \text{ the state transition matrix} \\
\epsilon_j &: \text{ the } j\text{th emission distribution, } j = 1, ..., k \\
\pi &: \text{ the initial latent state distribution}
\end{aligned}
$$

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

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

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

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:

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

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 $\texttt{VLBO}$:

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

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

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

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.

Maximizing the $\texttt{VLBO}$ minimizes the KL divergence

By definition, the $\texttt{KL}$ divergence from the target posterior to the variational density is given by

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

By the chain rule, we get

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

Maximizing the VLBO minimizes the KL divergence

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

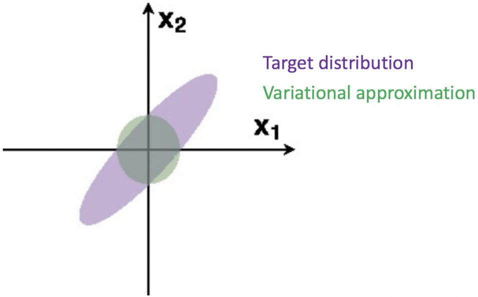
$$\mathtt{KL}(q(\boldsymbol{u}) \mid\mid p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})) = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{u})}{p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})}\right]$$

By the chain rule, we get

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

Note: 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

Problem: 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

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 $\boldsymbol{\theta}$	$p(\boldsymbol{x} \mid \boldsymbol{c})$	$q(\boldsymbol{u})$	$p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})$
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	$p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{x})$	$p(\boldsymbol{x}) = \int p(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{z}) \, d\,\boldsymbol{\theta} \;\; d\,\boldsymbol{z}$	$q(\boldsymbol{z}, \boldsymbol{\theta})$	$p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{x})$
Frequentist latent	$\operatorname{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{x} \mid \boldsymbol{\theta})$	$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \int p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}) \, d\,\boldsymbol{z}$	$q(\boldsymbol{z})$	$p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta})$

Variational Inference and Expectation Maximization

Expectation Maximization (EM)

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

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

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

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

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:

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

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 $\texttt{VLBO}$ 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:

$$\begin{aligned} \textbf{q update}: &\quad q_{\boldsymbol{z}}^{(t+1)} = \mathrm{argmax}_{q_{\boldsymbol{z}}}\ \texttt{VLBO}\,(q_{\boldsymbol{z}}; \boldsymbol{\theta}^{(t)}) & (6) \\ \boldsymbol{\theta}\ \textbf{update}: &\quad \boldsymbol{\theta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\theta}}\ \texttt{VLBO}\,(q_{\boldsymbol{z}}^{(t+1)}; \boldsymbol{\theta}) & (7) \end{aligned}$$

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

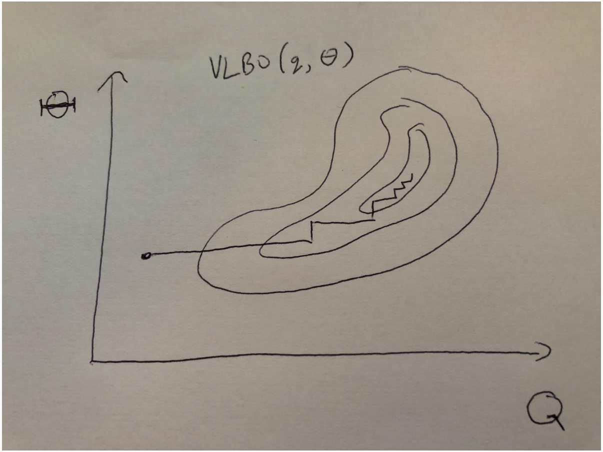
$$q_{\boldsymbol{z}}^{(t+1)} = p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

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

$$\boldsymbol{\theta}^{(t+1)} = \mathrm{argmax}_{\boldsymbol{\theta}}\ \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}\Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big] \qquad (8)$$

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 `VLBO` is given by:

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

which generalizes the EM algorithm.

Extension: Variational MAP

Consider a Bayesian latent variable model. Recall that a(n unscrupulous) Bayesian will sometimes compute a maximimum a posteriori (MAP) estimate of parameter $\boldsymbol{\theta}$.

$$\boldsymbol{\theta}_{\mathrm{MAP}} := \operatorname{argmax}_{\boldsymbol{\theta}} \, p(\boldsymbol{x}, \boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \int p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}) \; d\,\boldsymbol{z} \tag{9}$$

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

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

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

Variational Bayes Expectation Maximization (VBEM)

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

Suppose that we construct variational density with the factorization

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

In this case, the VLBO becomes

$$\texttt{VLBO}\left(q_{\boldsymbol{z}}(\boldsymbol{z}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) := \int \int q_{\boldsymbol{z}}(\boldsymbol{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ \ln\left(\frac{p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{x})}{q_{\boldsymbol{z}}(\boldsymbol{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\right) d\,\boldsymbol{\theta} \ \ d\,\boldsymbol{z} \tag{10}$$

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

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

VBEM: Derivation

See notes.

VBEM: Update Equations

The coordinate ascent equations have the form

$$\textbf{VB-E step}: \quad q_{\boldsymbol{z}}^{(t+1)} \propto \exp\left(\mathbb{E}_{q_{\boldsymbol{\theta}}^{(t)}}\left[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\right]\right) \tag{11}$$

$$\textbf{VB-M step}: \quad q_{\boldsymbol{\theta}}^{(t+1)} \propto p(\boldsymbol{\theta}) \exp\left(\mathbb{E}_{q_{\boldsymbol{z}}^{(t)}}\left[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\right]\right) \tag{12}$$

Prior-likelihood decomposition

Bayes' rule

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

VB-M update

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

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)

Towards generalizing the VBEM updates

The VB-E and VB-M steps in (11) and (12) have different forms, suggesting a dependence on the "semantic" distinction between $\boldsymbol{z}$ and $\boldsymbol{\theta}$ (as scaling vs. not scaling with the data $\boldsymbol{x}$).

But the VLBO in (10) 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

$$q_{\boldsymbol{\theta}}^{(t+1)} \propto \exp\left(\mathbb{E}_{q_{\boldsymbol{z}}^{(t)}}\left[\ln p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{z})\right]\right) \tag{13}$$

$$q_{\boldsymbol{z}}^{(t+1)} \propto p(\boldsymbol{z}) \exp\left(\mathbb{E}_{q_{\boldsymbol{\theta}}^{(t)}}\left[\ln p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{z})\right]\right) \tag{14}$$

by simply interchanging the syntactic roles of $\boldsymbol{z}$ and $\boldsymbol{\theta}$, 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

$$q(u_1, ..., u_K) = \prod_{k=1}^{K} q_k(u_k) \tag{15}$$

Mean field coordinate ascent variational inference (MF-CAVI) is CAVI performed under the mean field assumption (15).

Update equations for MF-CAVI

To perform coordinate ascent on the VLBO under the mean field assumption (15), we iteratively update our variational factors $\{q_k\}_k$ via

$$q_k(u_k) \propto \exp\left\{\mathbb{E}_{q_{-k}}\left[\log p(u_k \mid \boldsymbol{u}_{-k}, \boldsymbol{x}, \boldsymbol{c})\right]\right\} \tag{16}$$

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 (16) 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

$$p(x \mid \eta) = h(x) \exp\{\eta^T s(x) - a(\eta)\} \tag{17}$$

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 (16) puts

$$q_k \in \mathcal{E} \tag{18}$$

with natural parameter

$$\boxed{\nu_k = \mathbb{E}_{q_{-k}}[\eta_k(u_{-k}, x)]} \tag{19}$$

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

$$p(u_k \mid u_{-k}, x) = h(u_k) \exp\{\eta_k^T s(u_k) - a(\eta_k)\} \tag{20}$$

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

By substituting (20) into (16) 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 (20), 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 (16) in a concrete context, consider a version of the Bayesian Gaussian Mixture Model.

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

(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 (16) 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

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

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

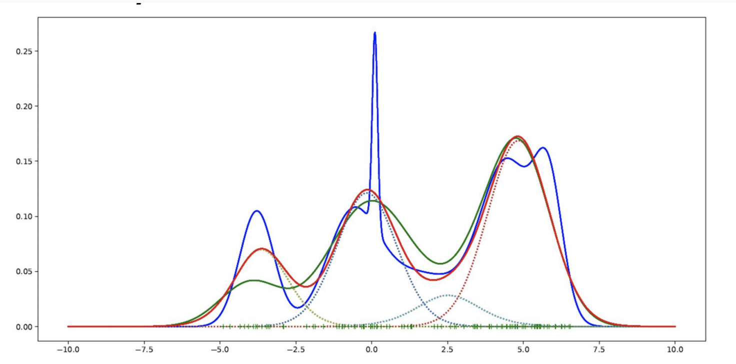


Image Credit: Lukas Burget

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

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

- $y_{1:T} = (y_1, ..., y_T)$ observed sequence
- $x_{1:T} = (x_1, ..., x_T)$: hidden state sequence ($x_t \in \{1, ..., K\}$)
- $\pi = \{\pi_k\}, \pi_k = P(x_1 = k)$: 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

$$
\begin{aligned}
p(x_{1:T}, y_{1:T} \mid \theta) &= p(x_1 \mid \theta) p(y_1 \mid x_1, \theta) \prod_{t=2}^{T} p(x_t \mid x_{t-1}, \theta) p(y_t \mid x_t, \theta) \\
&= p(x_1 \mid \pi) p(y_1 \mid x_1, \phi) \prod_{t=2}^{T} p(x_t \mid x_{t-1}, A) p(y_t \mid x_t, \phi) \\
&= \pi_{x_1} \prod_{t=2}^{T} A_{x_{t-1}, x_t} \prod_{t=1}^{T} p(y_1 \mid \phi_{x_t}) \qquad (21)
\end{aligned}
$$

HMM: Complete Data Likelihood is an Exponential Family

We continue

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

$$= \exp\left\{ \log \pi_{x_1} + \sum_{t=2}^{T} \log A_{x_{t-1}, x_t} + \sum_{t=1}^{T} \log p(y_1 \mid \phi_{x_t}) \right\} \qquad (22)$$

$$= \exp\left\{ \sum_{k=1}^{K} x_1^k \log \pi_k + \sum_{t=2}^{T} \sum_{k,k'=1}^{K} x_{t-1}^k x_t^{k'} \log A_{kk'} \right.$$

$$\left. + \sum_{t=1}^{T} \sum_{k=1}^{K} x_t^k \log p(y_t \mid x_t, \phi_k) \right\} \qquad (23)$$

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

$$p(\theta) = p(A)p(\phi)p(\pi) \tag{24}$$

and let us further assume that the priors have the form

$$p(\pi) = \mathrm{Dirichlet}(\pi \mid \alpha^{\pi}) \tag{25}$$

$$p(A) = \prod_{k=1}^{K} p(A_k \mid \alpha^{A_k}) = \prod_{k=1}^{K} \mathrm{Dirichlet}(A_k \mid \alpha^{A_k}) \tag{26}$$

where A_k designates the kth row of A.

π has a Dirichlet complete conditional

By the structure of the complete data likelihood (23) and prior (24), (25) , 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 $\log \pi = [\log \pi_1, ..., \log \pi_K]^T$.

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 (23) and prior (24), (26), 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 $\log A_k = [\log A_{k1}, ..., \log A_{kK}]^T$.

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

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

Variational update for x

By applying (16) and substituting in the form of the complete data likelihood (21), we obtain

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

Compare this to the original complete data likelihood (21). The structure is identical if we define the alterations $(A, \pi, p) \to (A^*, \pi^*, p^*)$.

In particular, we define

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

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

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

$\psi(\cdot)$ 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 $p(x_t \mid y, \theta)$ and pairwise marginals $p(x_{t-1}, x_t \mid y, \theta)$.
- 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)

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

Illustration

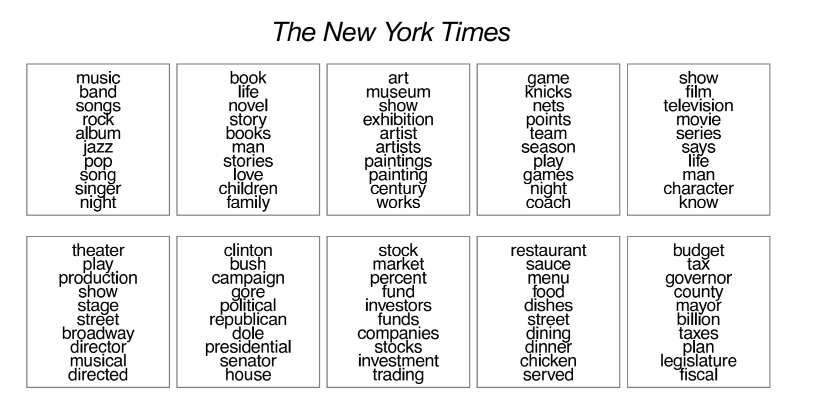


Figure 2: Posterior topics from a variant of LDA trained on 1.8M articles from the New York Times. Each topic is a weighted distribution over the vocabulary and each topic's plot illustrates its most frequent words.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. The Journal of Machine Learning Research, 14(1), 1303-1347.

Illustration

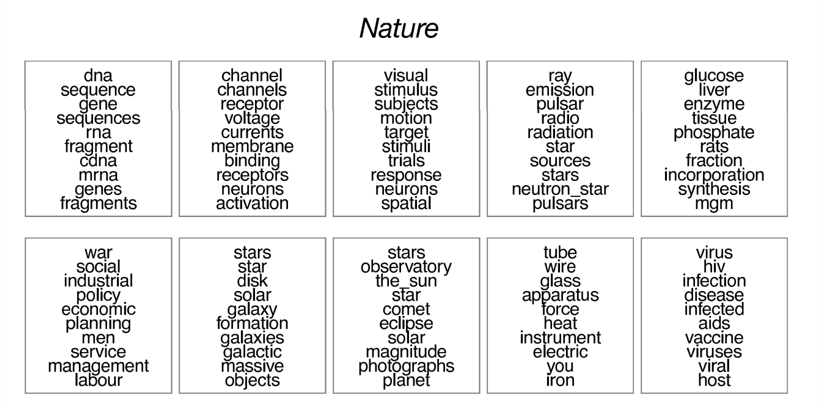


Figure 3: Posterior topics from a variant of LDA trained on 300K articles from Nature. Each topic is a weighted distribution over the vocabulary and each topic's plot illustrates its most frequent words.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. The Journal of Machine Learning Research, 14(1), 1303-1347.

Illustration

LDA assumes that each word in a document was generated from one of K latent topics.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

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*, $\boldsymbol{w}_n \in \mathrm{OneHot}(V)$ indicates which element of the vocabulary was observed as the nth word of the document.[1]
- A *document*, $\boldsymbol{w} \in \mathrm{OneHot}(V)^N$, is a list of N words. It is represented as a $N \times V$ matrix.

[1] We define $\mathrm{OneHot}(K)$ 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 $\mathrm{Multinoulli}(K)$ distribution.

Definitions: Hidden Variables

- A *topic indicator* is an integer in $\{1, \ldots K\}$.
- 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.

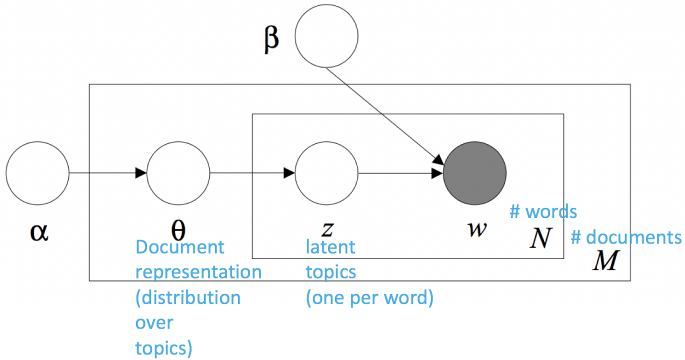
Hyperparameters

- $\boldsymbol{\alpha} \in (\mathbb{R}^+)^K$ 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

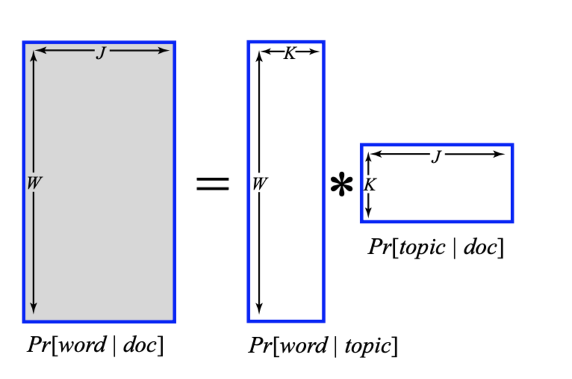
$$\beta_{kv} = P(w_{n,v} = 1 \mid z_{n,k} = 1),$$

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

Graphical Model



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.

The hyperparameters α and β define Dirichlet priors for the columns of the topic and word distribution matrices, respectively

Image Credit : Erik Sudderth

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 $\boldsymbol{w}_n \sim \text{Multinoulli}_V(\boldsymbol{\beta}_{\boldsymbol{z}_n,\cdot})$

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

$$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \underbrace{p(\boldsymbol{\theta} \mid \boldsymbol{\alpha})}_{\text{Dirichlet}} \prod_{n=1}^{N} \underbrace{p(\boldsymbol{z}_n \mid \boldsymbol{\theta})}_{\text{Multinoulli}} \underbrace{p(\boldsymbol{w}_n \mid \boldsymbol{z}_n, \boldsymbol{\beta})}_{\text{Multinoulli}} \tag{27}$$

Definition
($[\cdot]$) We define the operator $[\cdot] : \mathit{OneHot}(K) \to \{1, ..., K\}$ 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

$$\log p(\boldsymbol{z}_n \mid \boldsymbol{\theta}) = \log \boldsymbol{\theta}_{[z_n]} \tag{28}$$

$$\log p(\boldsymbol{w}_n \mid \boldsymbol{z}_n, \boldsymbol{\beta}) = \log \boldsymbol{\beta}_{[\boldsymbol{z}_n],[\boldsymbol{w}_n]} \tag{29}$$

Variational distribution

We approximate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{w})$ using mean field variational inference (15). In particular, we assume that the variational family $\mathcal{Q}$ factorizes as

$$\begin{aligned} q &= q_{\boldsymbol{\delta}}(\boldsymbol{\theta})\, q_{\boldsymbol{\tau}}(\boldsymbol{z}) \\ &= \underbrace{q_{\boldsymbol{\delta}}(\boldsymbol{\theta})}_{\text{Dirichlet}} \prod_{n=1}^{N} \underbrace{q_{\boldsymbol{\tau}_n}(\boldsymbol{z}_n)}_{\text{Multinoulli}} \end{aligned} \tag{30}$$

Update equations

LDA coordinate ascent update equations

$$\underset{\text{var. multinoulli (topics-for-word)}}{\tau_{n,k}} \propto \left(\Psi(\delta_k) - \Psi(\sum_j \delta_j)\right) \; \beta_{k,[w_n]} \tag{31}$$

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

$$\underset{\text{var. dirichlet (documents)}}{\delta_k} = \underset{\text{prior counts}}{\alpha_k} + \sum_{n=1}^{N} \underset{\text{var. prob of topics}}{\tau_{n,k}} \tag{32}$$

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

$$\mathbb{E}_{q_\delta(\theta)}\left[\log \boldsymbol{\theta}_i\right] = \Psi(\delta_i) - \Psi(\sum_k \delta_k) \tag{33}$$

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 $q^*(\boldsymbol{u})$ raises the (approximate) evidence term $p(\boldsymbol{x})$ (the term used for Bayesian model comparison) as high as possible.
- *Frequentist models*: The VLBO approximates the true marginal likelihood, $p(\boldsymbol{x} \mid \boldsymbol{\theta})$, 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 $q^*(\boldsymbol{u})$ is as similar as possible to the target posterior $p(\boldsymbol{u} \mid \boldsymbol{x})$
- *Frequentist models*: The value $\boldsymbol{\theta}^*$ is as close as possible to the solution provided by the classical EM algorithm. (See next section; we substitute $q^*(\boldsymbol{z}) \approx p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{\text{curr}})$ 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 $q_{\boldsymbol{z}}$

$$\texttt{VLBO}(q_{\boldsymbol{z}}(\boldsymbol{z}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})) := \int d\,\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta})\big(\ln p(\boldsymbol{\theta}) - \ln q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta})\big) + \int d\,\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \int d\,\boldsymbol{z}\; q_{\boldsymbol{z}}(\boldsymbol{z})\big(\ln p(\boldsymbol{z}, \boldsymbol{x} \mid \boldsymbol{\theta}) - \ln q_{\boldsymbol{z}}(\boldsymbol{z})\big)$$

The first summand is constant in $q_{\boldsymbol{z}}$; therefore it does not play a role in the argmax over $q_{\boldsymbol{z}}$. Thus, our coordinate ascent update must satisfy this criterion for a critical point

$$0 = \frac{\partial}{\partial q_{\boldsymbol{z}}}\left(\int d\,\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \int d\,\boldsymbol{z}\; q_{\boldsymbol{z}}(\boldsymbol{z})\big(\ln p(\boldsymbol{z}, \boldsymbol{x} \mid \boldsymbol{\theta})\big) - \ln q_{\boldsymbol{z}}(\boldsymbol{z})\big)\right)$$

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

$$0 = \left(\int d\,\boldsymbol{\theta}\; q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta})\big(\ln p(\boldsymbol{z}, \boldsymbol{x} \mid \boldsymbol{\theta})\big) - \ln q_{\boldsymbol{z}}(\boldsymbol{z})\big)\right) - \frac{q_{\boldsymbol{z}}}{q_{\boldsymbol{z}}}$$

Pulling out the constant term, $\ln q_{\boldsymbol{z}}$, rearranging, and taking the exponential yields the VB-E update.