Variational Inference

Michael Thomas Wojnowicz

June 14, 2021

Data Intensive Studies Center, Tufts University

Table of contents

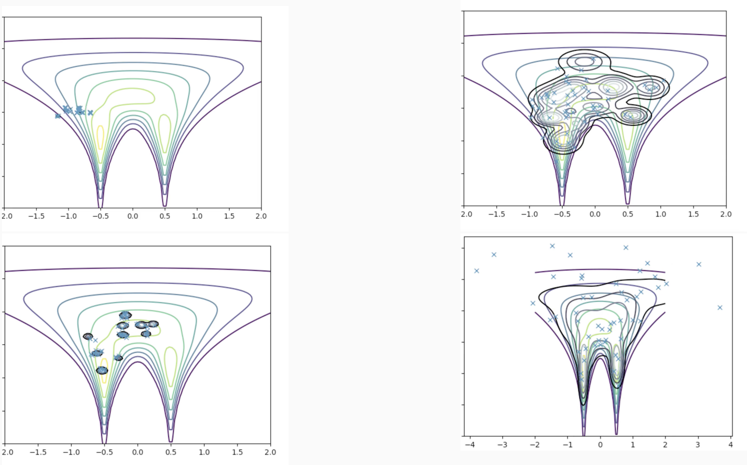
Some questions

- What is variational inference?
- When is it useful? Can it even be useful to frequentists?
- How can we apply VI to inference problems?
- What is *automatic* differentiation variational inference (ADVI)?

Overview

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).

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 (non-latent variable) models
Example: Any model with a non-conjugate prior

Three statistical modeling paradigms of interest

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

Bayesian (non-latent variable) models

Example: Any model with a non-conjugate prior

Bayesian latent variable models

Examples: Bayesian Mixture Model, Bayesian Hidden Markov Model, Latent Dirichlet Allocation,

Three statistical modeling paradigms of interest

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

Bayesian (non-latent variable) models

Example: Any model with a non-conjugate prior

Bayesian latent variable models

Examples: Bayesian Mixture Model, Bayesian Hidden Markov Model, Latent Dirichlet Allocation,

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

In general

We must compute the marginal

$$p(\boldsymbol{x} \mid \boldsymbol{c}) = \int p(\boldsymbol{x}, \boldsymbol{u} \mid \boldsymbol{c}) \; d\,\boldsymbol{u} \tag{1}$$

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

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:

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

Variational Inference: Maximizing the VLBO

Variational Inference

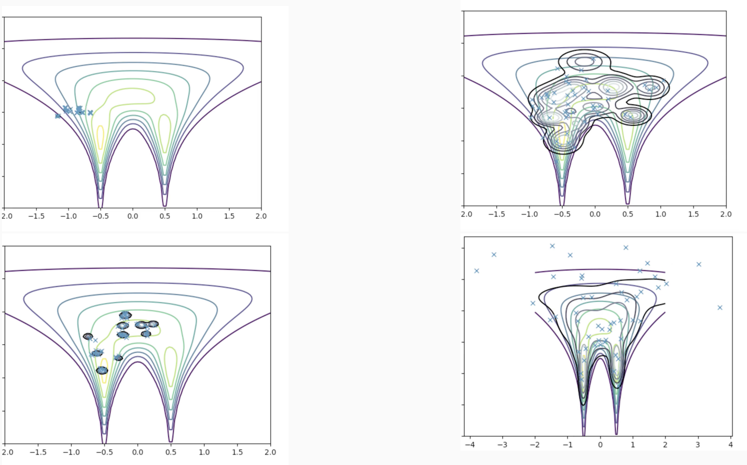
Variational inference (VI) proceeds by finding q^*, the variational density in tractable family $\mathcal{Q}$ which maximizes the $\texttt{VLBO}$:

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

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 $\mathcal{Q} = \{10\text{-component Gaussian mixture models}\}$



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

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

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

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

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 $\texttt{VLBO}$ minimizes the KL divergence (to the posterior)

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

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

Discuss: What is the optimal variational density?

Maximizing the $\texttt{VLBO}$ minimizes the KL divergence (to the posterior)

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

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

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

Overview

Summary

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

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

Note: Orange denotes the primary target, Blue denotes a helper.

How does VI accommodate the goal of statistical inference?

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

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.

How does VI accommodate the goal of statistical inference?

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

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]

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.

[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.

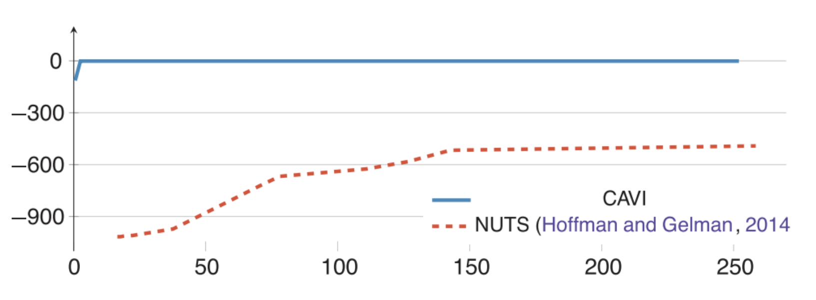


Figure 1: Comparison of CAVI to a Hamiltonian Monte Carlo-based sampling technique. The plot shows log predictive test set accuracy by training time (minutes). CAVI fits a Gaussian mixture model to 10,000 images in less than a minute.

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.

Variational Inference vs. Expectation Propagation

Let us fix a distribution P and consider two optimization strategies

Variational Inference

Minimizing
$\mathrm{KL}(Q||P)$
$= \mathbb{E}_Q\left[\log \frac{q(x)}{p(x)}\right]$

Expectation Propagation

Minimizing
$\mathrm{KL}(P||Q)$
$= \mathbb{E}_P\left[\log \frac{p(x)}{q(x)}\right]$

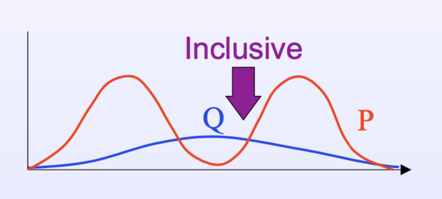
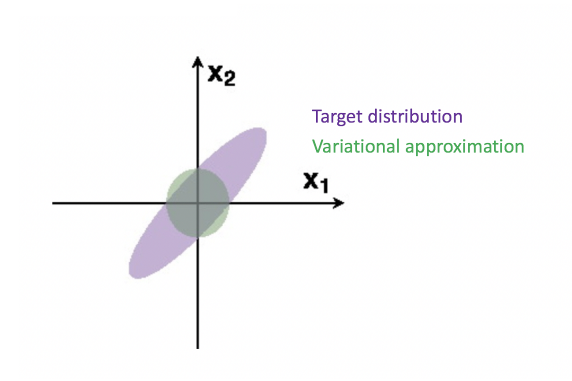


Image Credit: Tushar Tank

Q: What does this say about VI? Which one would you prefer to use?

Shortcoming: VI underestimates variance of the true posterior



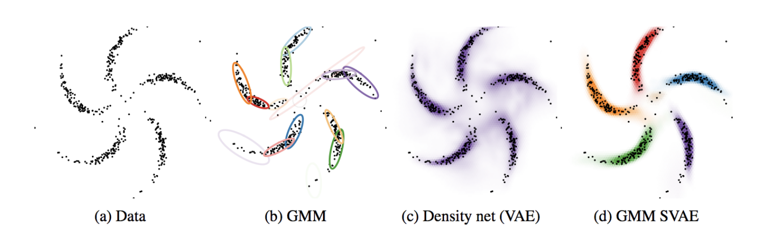
$$\mathrm{KL}(q(\boldsymbol{u}) \,||\, p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})) = \mathbb{E}_q\left[\log \frac{q(\boldsymbol{u})}{p(\boldsymbol{u} \mid \boldsymbol{x}, \boldsymbol{c})}\right]$$

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

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

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:

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

EM from the perspective of VI

For a frequentist latent variable model, the $\texttt{VLBO}$ is

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

EM from the perspective of VI

For a frequentist latent variable model, the VLBO is

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

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

$$\textbf{q update}: \quad q_{\boldsymbol{z}}^{(t+1)} = \text{argmax}_{q_{\boldsymbol{z}}}\ \texttt{VLBO}\,(q_{\boldsymbol{z}}; \boldsymbol{\theta}^{(t)}) \tag{2}$$

$$\boldsymbol{\theta}\ \textbf{update}: \quad \boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}}\ \texttt{VLBO}\,(q_{\boldsymbol{z}}^{(t+1)}; \boldsymbol{\theta}) \tag{3}$$

EM from the perspective of VI

For a frequentist latent variable model, the VLBO is

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

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

$$\textbf{q update}: \quad q_{\boldsymbol{z}}^{(t+1)} = \text{argmax}_{q_{\boldsymbol{z}}} \; \texttt{VLBO}\,(q_{\boldsymbol{z}}; \boldsymbol{\theta}^{(t)}) \tag{2}$$

$$\boldsymbol{\theta}\ \textbf{update}: \quad \boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} \; \texttt{VLBO}\,(q_{\boldsymbol{z}}^{(t+1)}; \boldsymbol{\theta}) \tag{3}$$

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 (what?)

EM from the perspective of VI

For a frequentist latent variable model, the `VLBO` is

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

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

$$\textbf{q update}: \quad q_{\boldsymbol{z}}^{(t+1)} = \text{argmax}_{q_{\boldsymbol{z}}} \ \texttt{VLBO}\,(q_{\boldsymbol{z}}; \boldsymbol{\theta}^{(t)}) \tag{2}$$

$$\boldsymbol{\theta}\ \textbf{update}: \quad \boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}} \ \texttt{VLBO}\,(q_{\boldsymbol{z}}^{(t+1)}; \boldsymbol{\theta}) \tag{3}$$

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

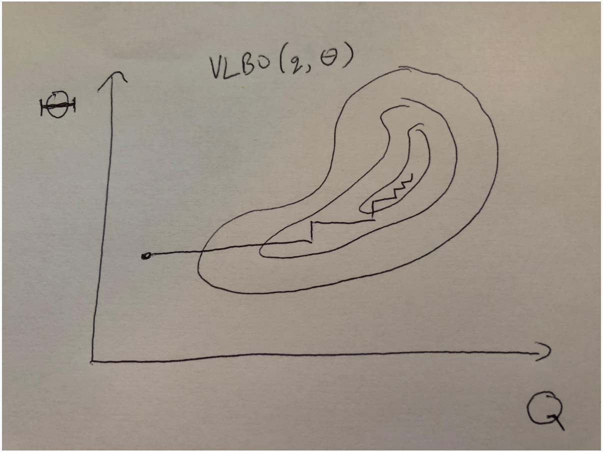
$$q_{\boldsymbol{z}}^{(t+1)} = p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}^{(t)})$$

in which case the $\boldsymbol{\theta}$ *update* becomes (what?)

$$\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}}\ \mathbb{E}_{p(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}^{(t)})}\Big[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big] \tag{4}$$

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:

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

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

$$\texttt{VLBO}\left(q_{\boldsymbol{z}}(\boldsymbol{z}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right) := \int \int q_{\boldsymbol{z}}(\boldsymbol{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \left(\frac{p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{x})}{q_{\boldsymbol{z}}(\boldsymbol{z}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) d\,\boldsymbol{\theta} \;\; d\,\boldsymbol{z} \tag{5}$$

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

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

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

VBEM: Derivation

See notes.

Variational Bayes Expectation Maximization (VBEM)

The coordinate ascent equations have the form

$$\textbf{VB-E step}: \quad q_{\boldsymbol{z}}^{(t+1)} \propto \exp\left(\mathbb{E}_{q_{\boldsymbol{\theta}}^{(t)}}\left[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\right]\right) \tag{6}$$

$$\textbf{VB-M step}: \quad q_{\boldsymbol{\theta}}^{(t+1)} \propto p(\boldsymbol{\theta}) \exp\left(\mathbb{E}_{q_{\boldsymbol{z}}^{(t)}}\left[\ln p(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\right]\right) \tag{7}$$

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)

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

$$q(u_1, ..., u_K) = \prod_{k=1}^{K} q_k(u_k) \tag{8}$$

Mean field coordinate ascent variational inference (MF-CAVI) is CAVI performed under the mean field assumption (8).

Update equations for MF-CAVI

To perform coordinate ascent on the `VLBO` under the mean field assumption (8), we iteratively update our variational factors $\{q_k\}_k$ via

$$q_k(u_k) \propto \exp\left\{\mathbb{E}_{q_{-k}}\left[\log p(u_k \mid \boldsymbol{u}_{-k}, \boldsymbol{x}, \boldsymbol{c})\right]\right\} \tag{9}$$

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

Update equations for MF-CAVI

To perform coordinate ascent on the `VLBO` under the mean field assumption (8), we iteratively update our variational factors $\{q_k\}_k$ via

$$q_k(u_k) \propto \exp\left\{\mathbb{E}_{q_{-k}}\left[\log p(u_k \mid \boldsymbol{u}_{-k}, \boldsymbol{x}, \boldsymbol{c})\right]\right\} \tag{9}$$

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

Rk: Note the connection to Gibbs sampling. In the MCMC literature, the distributions $p(u_k \mid \text{rest})$ are known as *full conditionals* or *complete conditionals*. Gibbs sampling involves successive draws from the full conditionals. In mean-field variational inference, we take expectations of the same distributions, in order to update our posterior approximations.

Example: Bayesian Mixture Model

Example: Bayesian Mixture Model

Why variational Bayesian mixture models?

Why variational Bayesian mixture models?

Why Bayes?

All the usual reasons – exploit prior knowledge, protect against overfitting, can use the evidence (or ELBO) for model selection, etc.

Why Variational Bayes?

Traditional MCMC becomes very burdensome for these types of models (mixture models, hidden mixture models, etc.) due to the multimodality in the posterior and the label switching.

See: A comparison of variational approximations for fast inference in mixed logit models

VB Predictive vs. ML Solution

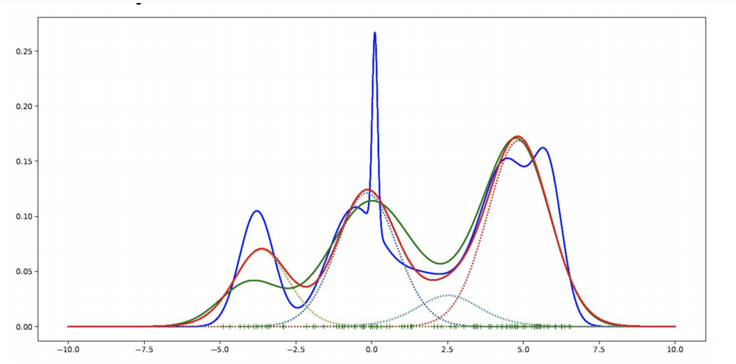


Image Credit: Lukas Burget

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

Extensions of Bayesian Mixture Models

The Bayesian framework can be used to endow mixture models with many nice properties.

Example: Dirichlet Process Mixture Models (DPMMs)

DPMM's have an unbounded number of mixture components.
The model automatically adapts its number of components.

- Click here for Demo 1.
- Click here for Demo 2.

Example: Bayesian Mixture Model

Inference algorithm

Example: Bayesian Gaussian Mixture Model

To see the mean field CAVI algorithm (9) 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 (9) 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 next slide reveals that the 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[2] 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:

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

[2] Indeed, in this course we have seen

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.

Overview

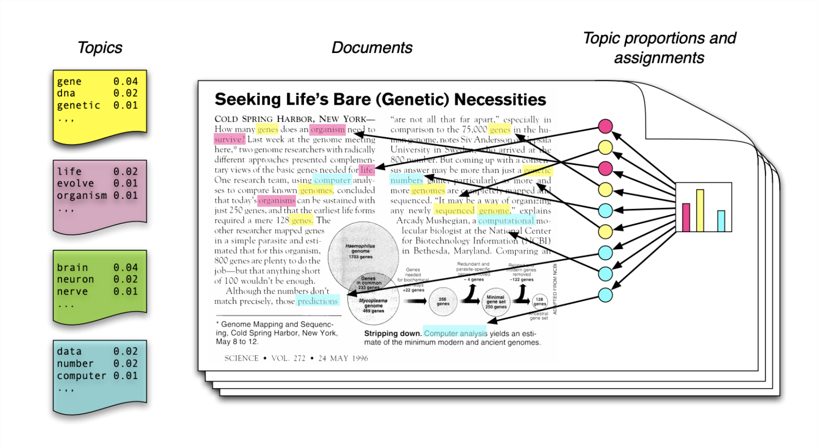
LDA is a generative probabilistic model of a corpus of documents of text.

LDA assumes:

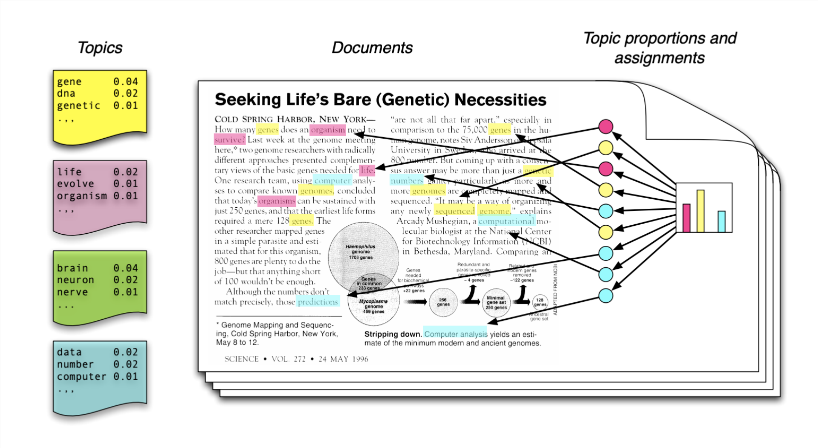
- There is a set of topics that describe the corpus
- Each document exhibits these topics to varying degrees.

So:

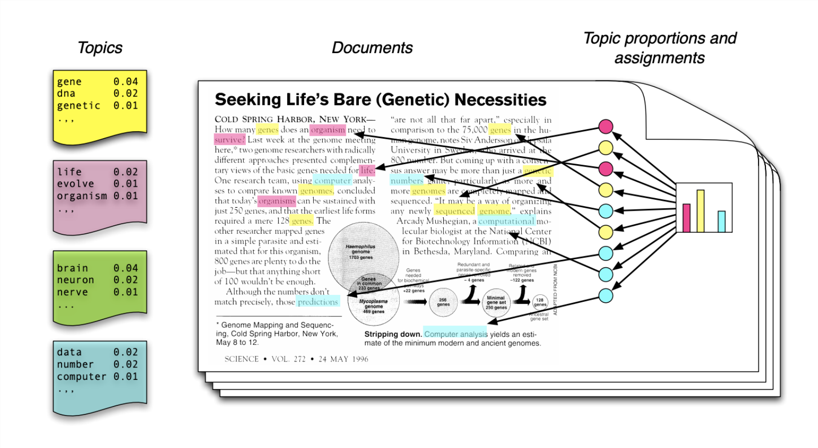
- The topics and how they relate to the documents are hidden structure
- The main computational problem is to infer this hidden structure



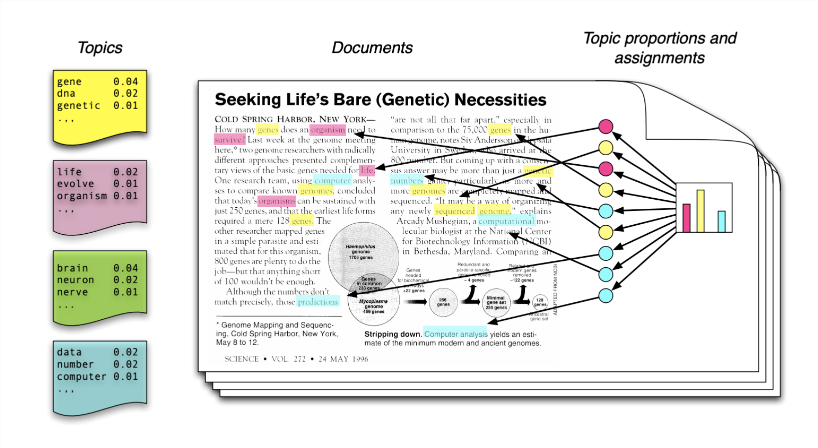
- Each **topic** is a distribution over words.



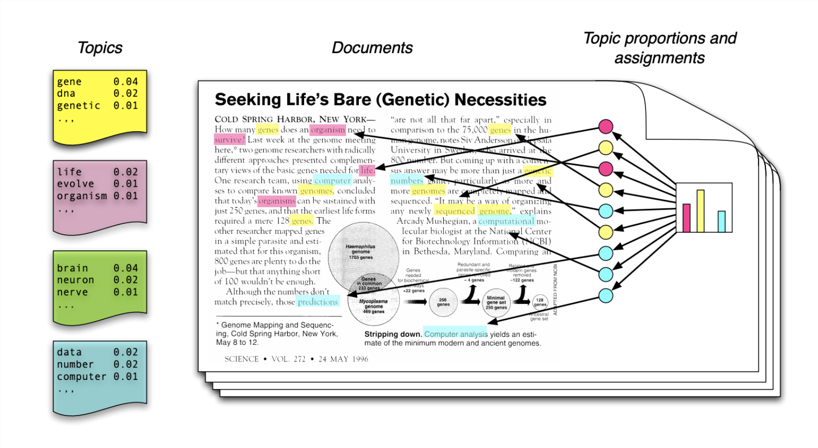
- Each **topic** is a distribution over words.
- Each **document** is a mixture of corpus-wide topics.



- 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.

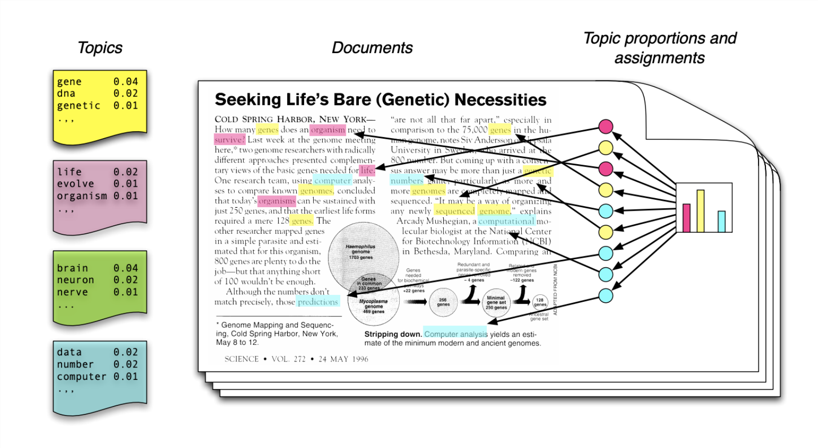


- Each **topic** is a distribution over words.
- Each **document** is a mixture of corpus-wide topics.
- Each **word** is drawn from one of those topics.



- 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.

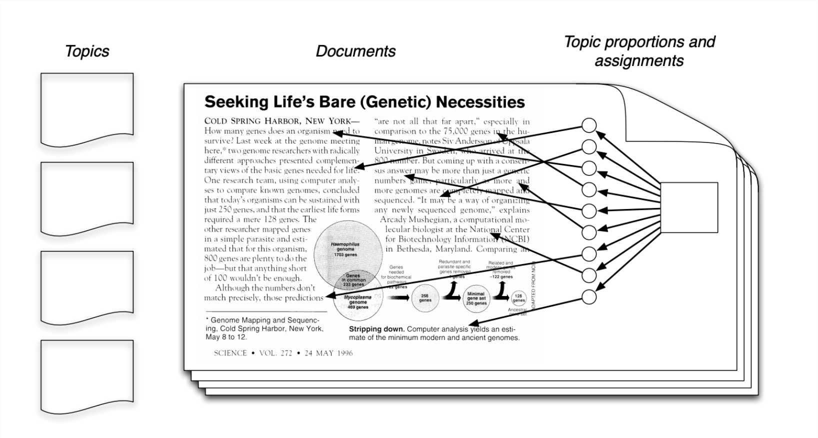
Rk: This is a "bag of words" model.



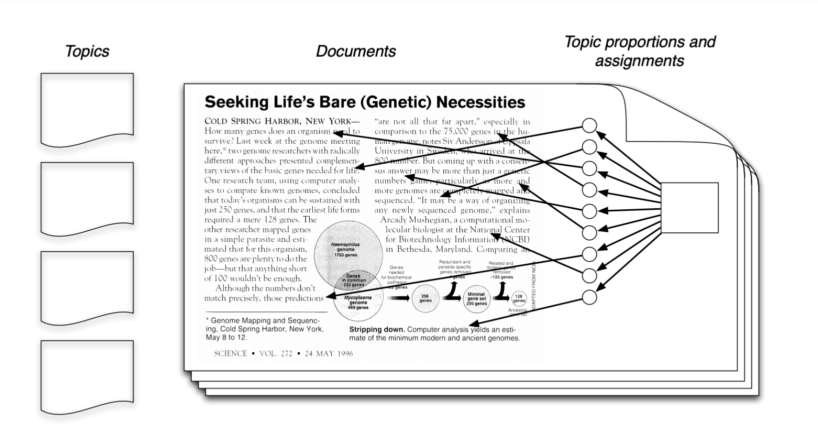
- Each **topic** is a distribution over words.
- Each **document** is a mixture of corpus-wide topics.
- Each **word** is drawn from one of those topics.

Rk: This is a "bag of words" model.

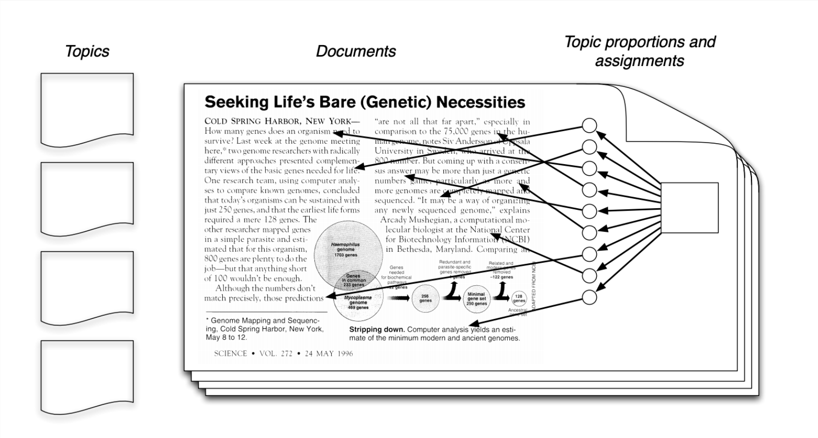
Source: David Blei, 2012 ICML Tutorial



- In reality, we only observe the documents.



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



- In reality, we only observe the documents.
- The model structure is **hidden**.
- Our goal is to **infer** the hidden variables; i.e. compute

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

Recall: Dirichlet Distribution

- Dirichlet distribution is *conjugate prior* of Categorical

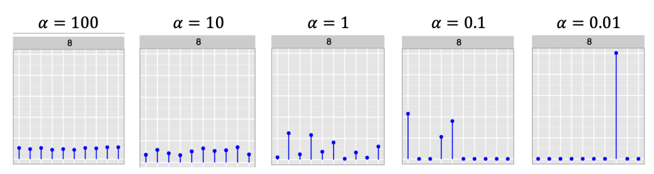
$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$$

Recall: Dirichlet Distribution

- Dirichlet distribution is *conjugate prior* of Categorical

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

- For symmetric Dirichlet distributions ($\alpha_1 = ... = \alpha_K$), a scalar hyperparameter $\alpha = \sum_k \alpha_k$ controls the shape and sparsity of the θ_d's. (per-document topic proportions) .
  - high α: typical θ_d (from the prior) will be uniform
  - small α: a typical θ_d (from the prior) will be sparse

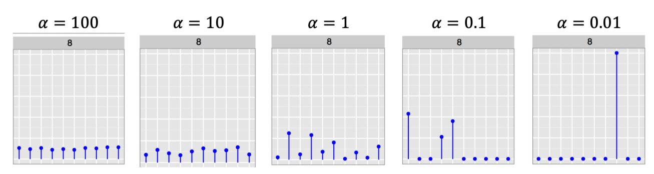


Recall: Dirichlet Distribution

- Dirichlet distribution is *conjugate prior* of Categorical

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

- For symmetric Dirichlet distributions ($\alpha_1 = \ldots = \alpha_K$), a scalar hyperparameter $\alpha = \sum_k \alpha_k$ controls the shape and sparsity of the θ_d's. (per-document topic proportions) .
    - high α: typical θ_d (from the prior) will be uniform
    - small α: a typical θ_d (from the prior) will be sparse



- Likewise, η controls the shape and sparsity of the β_k's (the topics – distribution over words)

LDA: Generative Process

The model is described by the following **generative process**:

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

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

LDA: Generative Process

The model is described by the following **generative process**:

- Fix a vocabulary of V words, and set the number of topics, K.

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

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

LDA: Generative Process

The model is described by the following **generative process**:

- Fix a vocabulary of V words, and set the number of topics, K.
- Set hyperparameters $\eta \in \mathbb{R}^V, \alpha \in \mathbb{R}^K$.

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

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

LDA: Generative Process

The model is described by the following **generative process**:

- Fix a 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 $(1, K)$:
    - Define the *topic* (a distribution over words), $\beta_k \in \mathbb{R}^V \sim \text{Dir}(\eta)$.[3]

---

[3] Interpretation of η: psuedocount of vocabulary words observed across prior topics.
[4] Interpretation of α: psuedocount of topics observed across prior documents.

LDA: Generative Process

The model is described by the following **generative process**:

- Fix a 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 $(1, K)$:
  - Define the *topic* (a distribution over words), $\beta_k \in \mathbb{R}^V \sim \text{Dir}(\eta)$.[3]
- For d in $(1, D)$:
  - Choose (per-document) *topic proportions* $\theta_d \in \mathbb{R}^K \sim \text{Dir}(\alpha)$.[4]

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

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

LDA: Generative Process

The model is described by the following **generative process**:

- Fix a 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 $(1, K)$:
  - Define the *topic* (a distribution over words), $\beta_k \in \mathbb{R}^V \sim \text{Dir}(\eta)$.[3]
- For d in $(1, D)$:
  - Choose (per-document) *topic proportions* $\theta_d \in \mathbb{R}^K \sim \text{Dir}(\alpha)$.[4]
  - For n in $(1, N_d)$:
    - Choose the *topic assignment* $z_{d,n} \sim \text{Categorical}_K(\theta_d)$

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

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

LDA: Generative Process

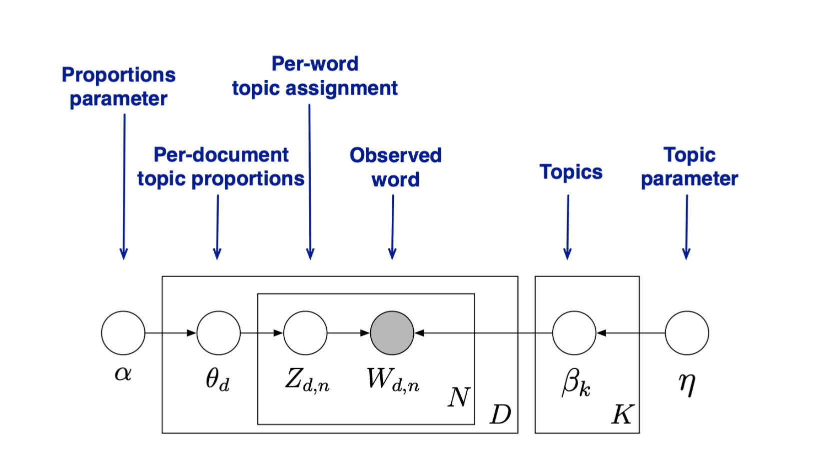
The model is described by the following **generative process**:

- Fix a 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 $(1, K)$:
  - Define the *topic* (a distribution over words), $\beta_k \in \mathbb{R}^V \sim \text{Dir}(\eta)$.[3]
- For d in $(1, D)$:
  - Choose (per-document) *topic proportions* $\theta_d \in \mathbb{R}^K \sim \text{Dir}(\alpha)$.[4]
  - For n in $(1, N_d)$:
    - Choose the *topic assignment* $z_{d,n} \sim \text{Categorical}_K(\theta_d)$
    - Choose *word* $w_{d,n} \sim \text{Categorical}_V(\beta_{z_{d,n}})$

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

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

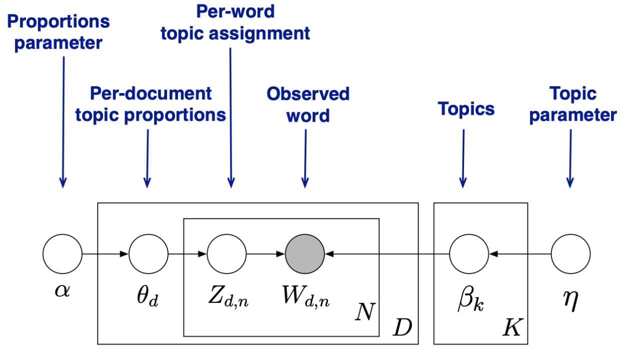
LDA as a graphical model



Recall:

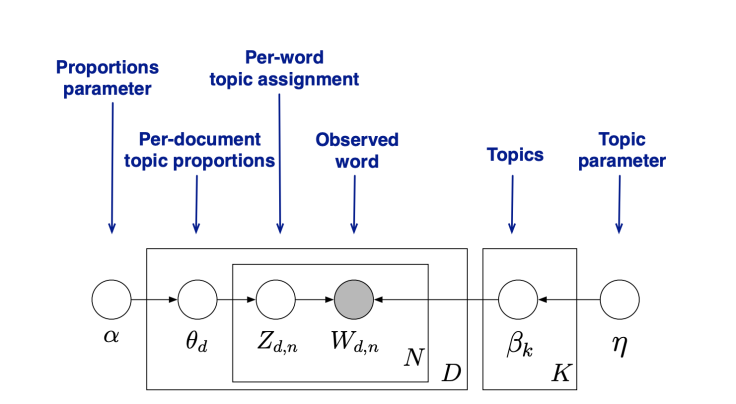
- Nodes are random variables.
- Shaded nodes are observed.
- Plates indicate replicated variables.
- Each node is conditionally independent from its non-descendants given its parents.

Joint distribution



$$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{\beta} \mid \alpha, \eta) = \prod_{k=1}^{K} \underbrace{p(\boldsymbol{\beta}_k \mid \eta)}_{\textit{Dirichlet}} \prod_{d=1}^{D} \underbrace{p(\boldsymbol{\theta}_d \mid \alpha)}_{\textit{Dirichlet}} \prod_{n=1}^{N} \underbrace{p(\boldsymbol{z}_{d,n} \mid \boldsymbol{\theta}_d)}_{\textit{Categorical}} \underbrace{p(\boldsymbol{w}_{d,n} \mid \boldsymbol{z}_{d,n}, \boldsymbol{\beta}_k)}_{\textit{Categorical}} \tag{10}$$

Joint distribution



$$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{w}, \boldsymbol{\beta} \mid \alpha, \eta) = \prod_{k=1}^{K} \underbrace{p(\boldsymbol{\beta}_k \mid \eta)}_{\text{Dirichlet}} \prod_{d=1}^{D} \underbrace{p(\boldsymbol{\theta}_d \mid \alpha)}_{\text{Dirichlet}} \prod_{n=1}^{N} \underbrace{p(\boldsymbol{z}_{d,n} \mid \boldsymbol{\theta}_d)}_{\text{Categorical}} \underbrace{p(\boldsymbol{w}_{d,n} \mid \boldsymbol{z}_{d,n}, \boldsymbol{\beta}_k)}_{\text{Categorical}} \tag{10}$$

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

Variational distribution

We approximate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{z}, \boldsymbol{w})$ using mean field variational inference (8). In particular, we assume that the variational family $\mathcal{Q}$ has a density which factorizes as

$$\begin{aligned} q &= q_{\boldsymbol{\delta}}(\boldsymbol{\theta})\, q_{\boldsymbol{\tau}}(\boldsymbol{z}) \\ &= \prod_{d=1}^{D} \underbrace{q_{\boldsymbol{\delta}}(\boldsymbol{\theta}_d)}_{\text{Dirichlet}} \prod_{n=1}^{N_d} \underbrace{q_{\boldsymbol{\tau}_n}(\boldsymbol{z}_{d,n})}_{\text{Categorical}} \end{aligned} \tag{11}$$

Note We are treating the topics β_k as a constant for simplicity. For a fuller treatment, see Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Update equations

LDA coordinate ascent update equations

$$\underbrace{\delta_{d,k}}_{\textit{var. dirichlet (topic proportion)}} = \underbrace{\alpha_k}_{\textit{prior counts}} + \sum_{n=1}^{N_d} \underbrace{\tau_{d,n,k}}_{\textit{var. expected assignment}} \tag{12}$$

$$\underbrace{\tau_{d,n,k}}_{\textit{var. categorical (topic assigment)}} \propto \exp\left\{ \underbrace{\mathbb{E}_{q_\delta(\theta)}}_{\textit{var. "prior"}} \underbrace{\left[\log \boldsymbol{\theta}_{d,k}\right]}_{\textit{topic assignment}} \right\} \underbrace{\beta_{k,[w_{d,n}]}}_{\textit{likelihood}}$$

$$= \left(\Psi(\delta_k) - \Psi(\sum_j \delta_j) \right) \beta_{k,[w_{d,n}]} \tag{13}$$

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

- Derivable via VBEM (see notes).
- Characteristic form: latent variable update depends on the data, global parameter update depends on the latent variable

Note We are treating β as a constant for brevity. We could fit also β to the data via VEM. (VEM does "empirical Bayes" for you.) More generally, we could model β as a random variable with VBEM. For a fuller treatment, see Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

The role of analytical computations

The varational categorical update crucially hinges on facts about the exponential family.

The role of analytical computations

The varational categorical update crucially hinges on facts about the exponential family.

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

$$\mathbb{E}_{q_\delta(\theta)}\left[\log \boldsymbol{\theta}_i\right] = \Psi(\delta_i) - \Psi(\sum_k \delta_k) \tag{14}$$

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).

The role of analytical computations

The varational categorical update crucially hinges on facts about the exponential family.

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

$$\mathbb{E}_{q_\delta(\theta)}\left[\log \boldsymbol{\theta}_i\right] = \Psi(\delta_i) - \Psi(\sum_k \delta_k) \tag{14}$$

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).

For many more complicated models (e.g. VAE), such expectations (even the variational ones) are intractable, and so we won't be able to use CAVI.

LDA Example Inference

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

Application: Anomaly Detection in Network Traffic Traces

LDA can be applied to *documents* that can be just about anything!

IP Addresses

An **ip address**, like 72.194.113.177, is (roughly) an address assigned to each device connected to the Internet. My laptop has one, my iphone has one, every website (Google, Apple, etc.) has one, etc.

One application (Newton, B. D. (2012). Anomaly Detection in Network Traffic Traces Using Latent Dirichlet Allocation.)

- "Documents" = the session of a specific IP address
- "words" = the full external IP address and port number combinations.
- The "words" in each "document" are counted and then this data set is processed by LDA to yield a compact model of the data.

Q: Ok, but how to perform anomaly detection? **Q:** What assumptions are being made by LDA?

I analyzed each of the anomalies detected in the last half hour of the trace [...]. The second anomaly with nearly 300 thousand messages exchanged with an SMTP server, is a bit more troubling. It is possible that this was actually a malicious client participating in a Mailbomb attack. According to the DARPA Intrusion Detection Attacks Database [8] a Mailbomb attack "is one in which the attacker sends many messages to a server, overflowing that server's mail queue and possibly causing a system failure."

Automatic Differentiation Variational Inference (ADVI)

Coordinate ascent, and its discontents

The ELBO (Evidence Lower Bound): Parametric View

$$L(\lambda) = \mathbb{E}_{q_\lambda(z)}\left[\log p(x,z) - \log q(z)\right]$$

Coordinate ascent, and its discontents

The ELBO (Evidence Lower Bound): Parametric View

$$L(\lambda) = \mathbb{E}_{q_\lambda(z)}\Big[\log p(x, z) - \log q(z)\Big]$$

Traditionally, we optimize via coordinate-ascent (CAVI).

$$\widehat{\lambda}_{t+1} = \widehat{\lambda}_t + \rho_t \nabla L(\widehat{\lambda}_t)$$

Coordinate ascent, and its discontents

The ELBO (Evidence Lower Bound): Parametric View

$$L(\lambda) = \mathbb{E}_{q_\lambda(z)}\Big[\log p(x,z) - \log q(z)\Big]$$

Traditionally, we optimize via coordinate-ascent (CAVI).

$$\widehat{\lambda}_{t+1} = \widehat{\lambda}_t + \rho_t \nabla L(\widehat{\lambda}_t)$$

Problem: Each model's VI algorithm is[5] a snowflake.

- Requires *model-specific* computations
- Must pick a good q.

[5] if it exists

Differentiable Probability Models

GOAL: Approximate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}) \propto p(\boldsymbol{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$

The class of probability models that ADVI supports

- Dataset $\boldsymbol{x} = x_{1:N}$ where each x_n is a realization of a discrete or continuous random variable
- Latent variables $\boldsymbol{\theta}$ are continuous
- $\nabla_\theta \log p(\boldsymbol{x}, \boldsymbol{\theta})$ exists within the support of the prior

$$\Theta := \text{supp}(p(\boldsymbol{\theta})) = \{\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^K \text{ and } p(\boldsymbol{\theta}) > 0\} \subseteq \mathbb{R}^K$$

INCLUDES

Generalized linear models
Mixture models
Topic models
Linear dynamic systems
Gaussian process models
Deep exponential families

EXCLUDES

Ising model
Sigmoid belief networks
Bayesian nonparametric models

Step 1: Transforming to unbounded support

We define a differentiable bijection to give the parameters unbounded support.

$$T : \Theta \to \mathbb{R}^K$$
$$\theta \mapsto \zeta$$

We use *change of variables* to express the joint density in the new space.

$$\underbrace{\mathrm{p}(\boldsymbol{x},\zeta)}_{\textit{new space}} = \underbrace{p\big(\boldsymbol{x},\, T^{-1}(\zeta)\big) \left| \det J_{T^{-1}}(\zeta) \right|}_{\textit{transformed original space}}$$

So in the new space, the ELBO becomes

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\zeta;\boldsymbol{\lambda})}\left[\log p\big(\boldsymbol{x},\, T^{-1}(\zeta)\big) + \log \left| \det J_{T^{-1}}(\zeta) \right| \right] + \mathbb{H}[q(\zeta;\boldsymbol{\lambda})]$$

Step 1: Transforming to unbounded support

We define a differentiable bijection to give the parameters unbounded support.

$$T : \Theta \to \mathbb{R}^K$$
$$\theta \mapsto \zeta$$

We use *change of variables* to express the joint density in the new space.

$$\underbrace{\mathrm{p}(\boldsymbol{x},\zeta)}_{\textit{new space}} = \underbrace{p\big(\boldsymbol{x},\, T^{-1}(\zeta)\big) \left|\det J_{T^{-1}}(\zeta)\right|}_{\textit{transformed original space}}$$

So in the new space, the ELBO becomes

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\zeta;\boldsymbol{\lambda})}\Big[\log p\big(\boldsymbol{x},\, T^{-1}(\zeta)\big) + \log \left|\det J_{T^{-1}}(\zeta)\right|\Big] + \mathbb{H}[q(\zeta;\boldsymbol{\lambda})]$$

✓ Model-independent variational factors.

✗ Cannot compute the gradient of the cost function

Step 1: Transforming to unbounded support

We define a differentiable bijection to give the parameters unbounded support.

$$T : \Theta \to \mathbb{R}^K$$
$$\theta \mapsto \zeta$$

We use *change of variables* to express the joint density in the new space.

$$\underbrace{\mathrm{p}(\boldsymbol{x},\zeta)}_{\text{new space}} = \underbrace{p\big(\boldsymbol{x},\, T^{-1}(\zeta)\big) \left|\det J_{T^{-1}}(\zeta)\right|}_{\text{transformed original space}}$$

So in the new space, the ELBO becomes

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\zeta;\boldsymbol{\lambda})}\Big[\log p\big(\boldsymbol{x},\, T^{-1}(\zeta)\big) + \log \left|\det J_{T^{-1}}(\zeta)\right|\Big] + \mathbb{H}[q(\zeta;\boldsymbol{\lambda})]$$

✓ Model-independent variational factors.

✗ Cannot compute the gradient of the cost function (Think: Why can't we just approximate the term by sampling?)

Step 2: The reparameterization trick

Example: We can *re-parameterize* the Gaussian $\boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, such that its dependence on the original parameter $\boldsymbol{\lambda} = (\boldsymbol{\mu}, \Sigma)$ is transferred to a (deterministic) standardization function, $S_{\boldsymbol{\lambda}}$

- Factorize $\Sigma = \boldsymbol{L}\boldsymbol{L}^T$.
- The standardized random variable is

$$\boldsymbol{\epsilon} := S_{\boldsymbol{\lambda}}(\zeta) = \boldsymbol{L}^{-1}(\zeta - \boldsymbol{\mu}), \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$$

- The unstandardized random variable can be recovered via

$$\zeta = S_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\epsilon})$$

ADVI ELBO

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon};0,\boldsymbol{I})}\left[\log p\left(\boldsymbol{x}, T^{-1}\left(S_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\epsilon})\right)\right) + \log \left|\det J_{T^{-1}}\left(S_{\boldsymbol{\lambda}}^{-1}(\boldsymbol{\epsilon})\right)\right|\right] + \mathbb{H}[q(\boldsymbol{\zeta};\boldsymbol{\lambda})]$$

Step 2: The reparameterization trick

Example: We can *re-parameterize* the Gaussian $\zeta \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, such that its dependence on the original parameter $\boldsymbol{\lambda} = (\boldsymbol{\mu}, \Sigma)$ is transferred to a (deterministic) standardization function, $S_{\boldsymbol{\lambda}}$

- Factorize $\Sigma = \boldsymbol{L}\boldsymbol{L}^T$.
- The standardized random variable is

$$\epsilon := S_{\boldsymbol{\lambda}}(\zeta) = \boldsymbol{L}^{-1}(\zeta - \boldsymbol{\mu}), \quad \text{where } \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$$

- The unstandardized random variable can be recovered via

$$\zeta = S_{\boldsymbol{\lambda}}^{-1}(\epsilon)$$

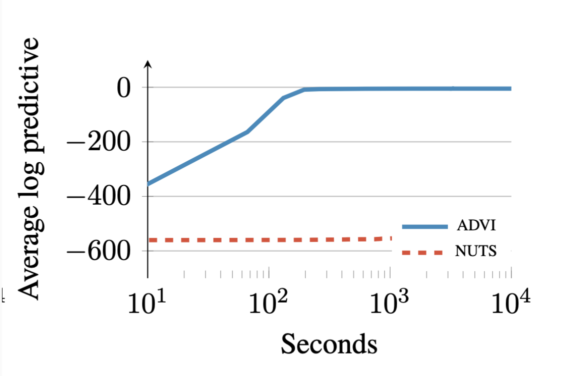
ADVI ELBO

$$\mathcal{L}(\boldsymbol{\lambda}) = \mathbb{E}_{\mathcal{N}(\boldsymbol{\epsilon};0,\boldsymbol{I})}\left[\log p\left(\boldsymbol{x},\, T^{-1}\left(S_{\boldsymbol{\lambda}}^{-1}(\epsilon)\right)\right) + \log\left|\det J_{T^{-1}}\left(S_{\boldsymbol{\lambda}}^{-1}(\epsilon)\right)\right|\right] + \mathbb{H}[q(\boldsymbol{\zeta}; \boldsymbol{\lambda})]$$

✓ Model-independent variational factors.

✓ Can estimate the gradient of the expectation.

Inference can be substantially more efficient



Results with a non-negative matrix factorization model applied to the Frey Faces dataset.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. The Journal of Machine Learning Research, 18(1), 430-474.

Questions?