Hierarchical Variational Models

Rajesh Ranganath Dustin Tran David M. Blei

Department of Computer Science, Princeton University
Department of Computer Science + Statistics, Columbia University

1. Review: Variational Inference

Let $p(\mathbf{z} \mid \mathbf{x})$ denote a posterior distribution, which is a distribution on d latent variables $\mathbf{z}_1, \ldots, \mathbf{z}_d$ conditioned on a set of observations $\mathbf{x}$.

In variational inference, one posits a family of distributions $q(\mathbf{z}; \lambda)$ and maximizes the Evidence Lower BOund (ELBO),

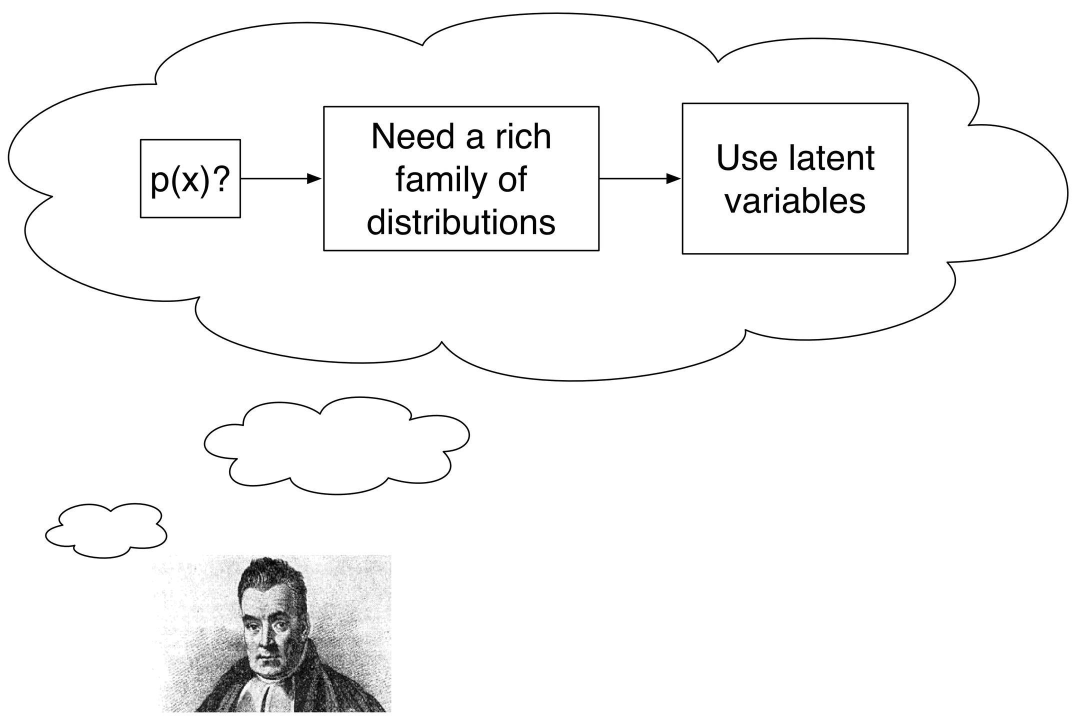
$$\mathcal{L}(\lambda) = \mathbb{E}_{q(\mathbf{z};\lambda)}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \lambda)].$$

Maximizing the ELBO minimize the KL to the posterior.

2. Variational Models

While black box variational methods expose variational inference algorithms to all probabilistic models, it remains an open problem to specify a variational distribution which both maintains high fidelity to arbitrary posteriors and is computationally tractable.

Practitioners add latent variables to form rich distributions over data:



Variational Models: View the variational distribution $q(\mathbf{z})$ as a "model" and use the same tools one uses to model data.

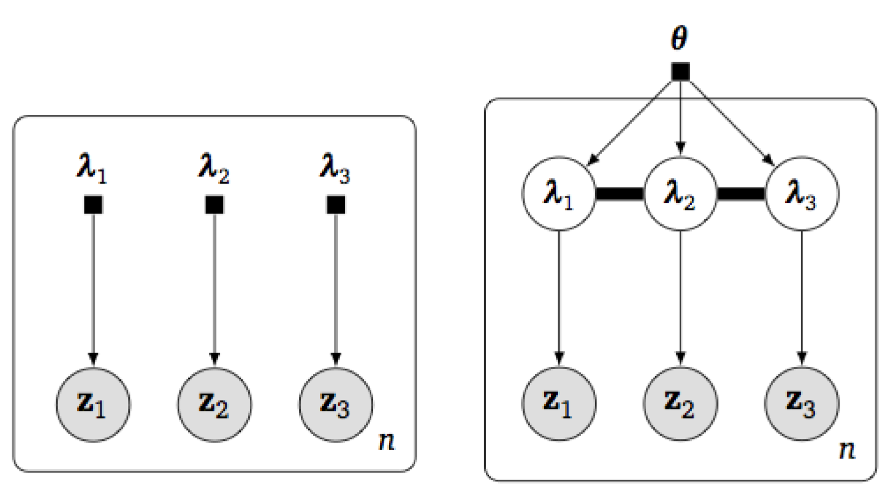
3. Hierarchical Variational Models

We construct hierarchical variational models by placing priors on tractable families of variational approximations. We focus on the mean-field family here.

Viewing the mean-field distribution plainly as a model of the posterior, a natural way to introduce more complexity is to construct it hierarchically. Adding a one layer hierarchical prior leads to the variational model

$$q_{\text{HVM}}(\mathbf{z}; \theta) = \int \left[\prod_{i=1}^{d} q(\mathbf{z}_i \mid \lambda_i)\right] q(\lambda; \theta)\, \mathrm{d}\lambda.$$

HVMs *provide richer approximations through the Bayesian hierarchical modeling framework.* Additional connections to: empirical Bayes, policy search methods, and annealing.



(a) MEAN-FIELD MODEL (b) HIERARCHICAL MODEL

4a. Example Hierarchical Variational Models

Specifying an HVM requires two components: the variational likelihood $q(\mathbf{z} \mid \lambda)$ and the prior $q(\lambda; \theta)$. The likelihood factors can be chosen in the same way that mean-field factors are typically chosen. The variational prior for a mixture of Gaussian is

$$q(\lambda; \theta) = \sum_{i=1}^{K} \pi_k \mathrm{N}(\mu_k, \sigma_k).$$

Higher order moments are capture by coocurrence in mixture components.

4b. Example Hierarchical Variational Models

We can construct variational priors by using normalizing flows [1]. Normalizing flows transform samples from a simple distribution in order to induce more complex representations.

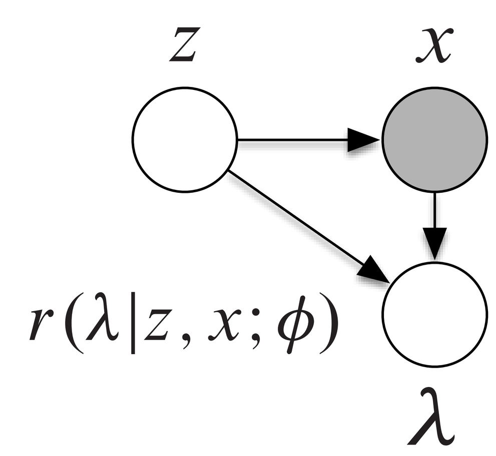
Formally, let q_0 be the distribution for λ_0 and λ be the result after k transformations. Then the log density of λ is

$$\log q(\lambda) = \log q(\lambda_0) - \sum_{k=1}^{K} \log \left(\left| \det\left(\frac{\partial f_k}{\partial z_k}\right) \right| \right).$$

HVMs extend the applicability of normalizing flows to discrete variables. We can also place a distribution over transformations to build an HVM without Jacobians [2].

5. Hierarchical ELBO

The entropy in hierarchical variational models is intractable. We can construct a tractable lower bound by expanding the model and doing variational inference.



This leads to the objective

$$\widetilde{\mathcal{L}}(\theta, \phi) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z}) + \log r(\lambda \mid \mathbf{x}, \mathbf{z}; \phi) - \log q(\mathbf{z}, \lambda; \theta)].$$

This is looser than marginal VB as variational latent variables imply a repeated application of Jensen's inequality.

6. Stochastic Gradients

The black-box gradient for the ELBO is

$$\nabla_\lambda \mathcal{L} = \mathbb{E}_q[\nabla_\lambda \log q(\mathbf{z}; \lambda)(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \lambda))].$$

Its variance scales with the learning signal. This can be improved for mean-field approximations using the structure of the model:

$$\nabla_{\lambda_i} \mathcal{L} = E_{q_{(i)}}[\nabla_{\lambda_i} \log q(\mathbf{z}_i; \lambda_i)(\log p_i(\mathbf{x}, \mathbf{z}_{(i)}) - \log q(\mathbf{z}_i; \lambda_i))].$$

The gradient of HVM with a differentiable prior is

$$\begin{aligned}\nabla_\theta \widetilde{\mathcal{L}}(\theta, \phi) &= \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon) \nabla_\lambda \mathcal{L}_{\text{MF}}(\lambda)] \\ &+ \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon) \nabla_\lambda [\log r(\lambda \mid \mathbf{z}; \phi) - \log q(\lambda; \theta)]] \\ &+ \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon) \mathbb{E}_{q(\mathbf{z} \mid \lambda)}[\nabla_\lambda \log q(\mathbf{z}; \lambda) \log r(\lambda \mid \mathbf{z}; \phi)]].\end{aligned}$$

If r factorizes in $\mathbf{z}$, we maintain computational efficiency. One example of such an r is defined via an inverse flow

$$\log r(\lambda \mid \mathbf{z}) = \log r(\lambda_0 \mid \mathbf{z}) + \sum_{k=1}^{K} \log \left(\left| \det\left(\frac{\partial g_k^{-1}}{\partial \lambda_k}\right) \right| \right),$$

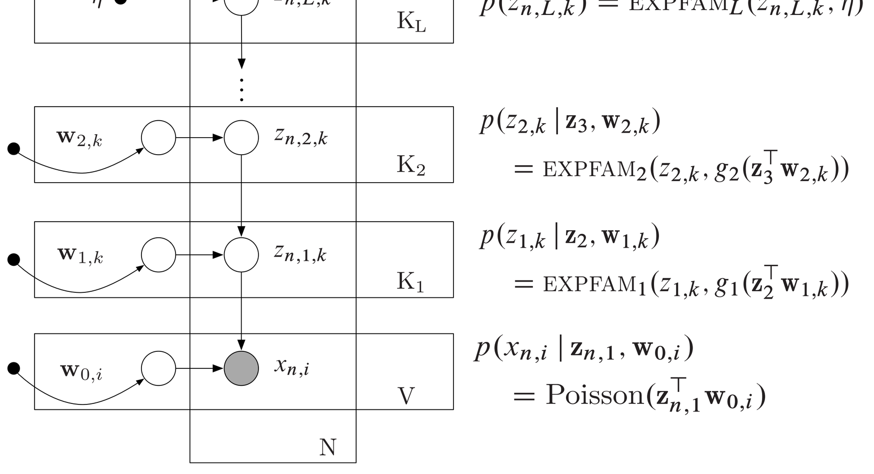
where

$$r(\lambda_0 \mid \mathbf{z}) = \prod_{i=1}^{d} r(\lambda_{0i} \mid \mathbf{z}_i).$$

Here r is a factorized regression under a parameterized transformation. Stochastic gradient updates are linear in the number of latent variables.

7. Results

We compare our method on deep exponential families [3] with multiple layers of Poisson latent variables.



	Model	HVM	Mean-Field
NYT	100	**3570**	**3570**
	100-30	**3460**	3660
	100-30-15	**3480**	3550
Science	100	**3360**	3377
	100-30	**3080**	3240
	100-30-15	**3110**	3190

We look at predictive perplexity. We get similar results on sigmoid belief networks.

References

1. Rezende + Mohamed, Variational Inference with Normalizing Flows, ICML, 2015.
2. Tran + Ranganath + Blei, Variational Gaussian Process, ArXiv, 2015.
3. Ranganath + Tang + Charlin + Blei, Deep Exponential Families, AISTATS, 2015.

Mail: rajeshr@cs.princeton.edu, dtran@g.harvard.edu, david.blei@columbia.edu