Hierarchical Variational Models

Rajesh Ranganath
Princeton University
rajeshr@cs.princeton.edu

Dustin Tran
Harvard University
dtran@g.harvard.edu

David M. Blei
Columbia University
david.blei@columbia.edu

Abstract

Black box inference allows researchers to easily prototype and evaluate an array of models. Recent advances in variational inference allow such algorithms to scale to high dimensions. However, a central question remains: How to specify an expressive variational distribution which maintains efficient computation? To address this, we develop hierarchical variational models. In a HIERARCHICAL VM, the variational approximation is augmented with a prior on its parameters, such that the latent variables are conditionally independent given this shared structure. This preserves the computational efficiency of the original approximation, while admitting hierarchically complex distributions for both discrete and continuous latent variables. We study HIERARCHICAL VM on a variety of deep discrete latent variable models. HIERARCHICAL VM generalizes other expressive variational distributions and maintains higher fidelity to the posterior.

1 Introduction

Black box variational inference (BBVI) is important to realizing the potential of modern applied Bayesian statistics. The promise of BBVI is that an investigator can specify any probabilistic model of hidden and observed variables, and then efficiently approximate the corresponding posterior without additional effort (Ranganath et al., 2014).

BBVI is a form of variational inference (Jordan et al., 1999; Wainwright and Jordan, 2008). It sets up a parameterized distribution over the latent variables and then optimizes the parameters to be close to the posterior distribution of interest. Typically this is done with the mean-field family, where each variable is independent and governed by its own variational parameters. Mean-field inference enables efficient BBVI algorithms, but it is limited by the strong factorization. It cannot capture dependencies between latent variables—this may be intrinsically important and also help improve the accuracy of the marginals.

In this paper we develop a black box variational method that goes beyond the mean-field and, indeed, beyond directly parameterized variational families in general. Our method remains tractable but uses a richer family of variational distributions, hierarchical variational models, and finds better approximations to the posterior. Like in traditional Bayesian methods, hierarchical variational models are built by placing a prior on the parameters of a tractable class of variational approximations, such as the mean-field, and integrating the prior out.

We develop a general algorithm for fitting HIERARCHICAL VM in the context of black box inference. Our algorithm maintains the computational efficiency of the original variational family. We demonstrate our methods with a study of approximate posteriors for several variants of deep exponential families (with Poisson layers) (Ranganath et al., 2015). In our study, hierarchical variational models always found better approximations to the exact posterior and formed better predictions on held out observations.

2 Variational Models

Black Box 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 $q(\mathbf{z}; \boldsymbol{\lambda})$, parameterized by $\boldsymbol{\lambda}$, and minimizes the KL divergence to the posterior distribution (Jordan et al., 1999; Bishop, 2006; Wainwright and Jordan, 2008).

This is equivalent to maximizing the Evidence Lower BOund (ELBO),

$$\mathscr{L}(\boldsymbol{\lambda}) = \mathbb{E}_{q(\mathbf{z};\boldsymbol{\lambda})}[\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}; \boldsymbol{\lambda})]. \tag{1}$$

Black box methods maximize the ELBO by constructing noisy gradients via samples from the variational approximation. This avoids model-specific computations, requiring only that the practictioner write a function to evaluate the model log-likelihood. Following the setting of black box inference, we now develop a framework for variational distributions.

Variational Models. While black box gradients 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 remains computationally tractable. The most common choice of variational model is the mean-field approximation

$$q_{\text{MF}}(\mathbf{z}; \boldsymbol{\lambda}) = \prod_{i=1}^{d} q(\mathbf{z}_i; \boldsymbol{\lambda}_i),$$

where $\boldsymbol{\lambda}_i$ denotes the parameters of the i^{th} latent variable. Mean-field approximations turn the computationally intractable problem of computing the posterior into an computationally feasible optimization problem. However, its marginal factorization compromises the expressivity of the variational model.

Hierarchical Variational Models. Viewing the mean-field distribution plainly as a model of the posterior, a natural way to introduce more complexity is to construct it hierarchically. That is, we place a prior distribution $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$ on the mean-field parameters and proceed to marginalize it out. Adding a one layer hierarchical prior leads to the variational model given by

$$q_{\text{HVM}}(\mathbf{z}; \boldsymbol{\theta}) = \int \left[\prod_{i=1}^{d} q(\mathbf{z}_i \mid \boldsymbol{\lambda}_i) \right] q(\boldsymbol{\lambda}; \boldsymbol{\theta}) d\boldsymbol{\lambda}. \tag{2}$$

The *variational prior* $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$ acts as a distribution over variational distributions, such that given its structure, each posterior variable $\mathbf{z}_i$ is conditionally independent. Thus we term q_{HVM} as the *hierarchical variational model*, which can either be a discrete or continuous distribution. For simplicity, we focus on one level hierarchies.

The variational prior $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$ is parameterized by a vector $\boldsymbol{\theta}$. These are the parameters we optimize over to find the optimal variational distribution within the class of hierarchical variational models. The ELBO using the hierarchical variational model is

$$\mathscr{L}(\boldsymbol{\theta}) = \mathbb{E}_{q_{\text{HVM}}(\mathbf{z};\boldsymbol{\theta})}[\log p(\mathbf{x}, \mathbf{z}) - \log q_{\text{HVM}}(\mathbf{z}; \boldsymbol{\theta})]. \tag{3}$$

We can lay out the properties required of variational models to ensure that the objective remains analytically tractable. The first term in the objective is tractable as long as we can sample from q and q has proper support. The second term with $\log q_{\text{HVM}}(\mathbf{z}; \boldsymbol{\theta})$, the entropy, contains an integral (Eq.2) that is in general analytically intractable.

We construct a bound on the entropy term by introducing a distribution $r(\boldsymbol{\lambda} \mid \mathbf{z}; \boldsymbol{\phi})$ with parameters $\boldsymbol{\phi}$, and applying the variational principle:

$$-\mathbb{E}_{q_{\text{HVM}}}[\log q_{\text{HVM}}(\mathbf{z})] \geq -\mathbb{E}_{q(\mathbf{z},\boldsymbol{\lambda})}[\log q(\boldsymbol{\lambda}) + \log q(\mathbf{z} \mid \boldsymbol{\lambda}) - \log r(\boldsymbol{\lambda} \mid \mathbf{z}; \boldsymbol{\phi})]. \tag{4}$$

We can view r as a recursive variational approximation. That is, it is a model for the posterior q of the mean-field parameters $\boldsymbol{\lambda}$ given a realization of the latent variables $\mathbf{z}$. Derivations and an alternative bound are supplied in the appendix.

Substituting the entropy bound (Eq.4) into the ELBO in Eq.3 gives a tractable lower bound which we call the *hierarchical* ELBO, denoted with $\widetilde{\mathscr{L}}$:

$$\widetilde{\mathscr{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q(\mathbf{z},\boldsymbol{\lambda};\boldsymbol{\theta})}\Big[\log p(\mathbf{x}, \mathbf{z}) + \log r(\boldsymbol{\lambda} \,|\, \mathbf{z}; \boldsymbol{\phi}) - \sum_{i=1}^{d} \log q(\mathbf{z}_i \,|\, \boldsymbol{\lambda}_i) - \log q(\boldsymbol{\lambda}; \boldsymbol{\theta})\Big]. \tag{5}$$

As all of the terms are known, this objective is tractable. We can fit q and r simultaneously by maximizing Eq.5 with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Maximizing this bound is equivalent to minimizing an upper bound on the KL-divergence of q_{HVM} to the black box posterior $p(\mathbf{z} \,|\, \mathbf{x})$. Similar to the EM-algorithm (Dempster et al., 1977), optimizing $\boldsymbol{\theta}$ improves the posterior approximation, while optimizing $\boldsymbol{\phi}$ tightens the upper bound on the KL divergence (improving the recursive variational approximation). We detail how to optimize this objective in the appendix.

3 Specifying the Hierarchical Variational Model

Hierarchical variational models are specified by a variational likelihood $q(\mathbf{z} \,|\, \boldsymbol{\lambda})$ and prior $q(\boldsymbol{\lambda})$. The variational likelihood can be considered as a typical mean-field distribution. We specify the variational prior based on normalizing flows, which was previously introduced as a variational approximation for differentiable probability models in Rezende and Mohamed (2015).

Normalizing Flows. Normalizing flows work by transforming a random variable $\boldsymbol{\lambda}_0$ with a known simple distribution such as the standard normal through a sequence of invertible functions f_1 to f_K. As each function is applied, the distribution of the output is a contorted version of the original distribution. This leads to very complicated variational families.

Consider normalizing flows for the variational prior. Formally, let q_0 be the distribution for $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}$ be the result after k transformations. Then the log density of $\boldsymbol{\lambda}$ is

$$\log q(\boldsymbol{\lambda}) = \log q(\boldsymbol{\lambda}_0) - \sum_{k=1}^{K} \log\left(\left|\det(\frac{\partial f_k}{\partial z_k})\right|\right). \tag{6}$$

Specifying r. The optimal r is the variational posterior $q(\boldsymbol{\lambda} \,|\, \mathbf{z})$. It is a continuous distribution, so we could again define r using a normalizing flow. The problem with this approach is that the intermediary $\boldsymbol{\lambda}$'s to $\boldsymbol{\lambda}_0$ required for the flow are unknown and generally require numeric inversions. Instead we define a normalizing flow where the *inverse flow* has a known parametric form. That is, let the distribution of $\boldsymbol{\lambda}$ in r be given by a sequence of invertible transforms g_1 to g_L of a simple distribution r_0. Now let $g^{-1}(\boldsymbol{\lambda})$ have a known form. The density of $\boldsymbol{\lambda}$ in r is

$$\log r(\boldsymbol{\lambda} \,|\, \mathbf{z}) = \log r(\boldsymbol{\lambda}_0 \,|\, \mathbf{z}) + \sum_{k=1}^{K} \log\left(\left|\det(\frac{\partial g_k^{-1}}{\partial \boldsymbol{\lambda}_k})\right|\right). \tag{7}$$

The sequence of intermediary $\boldsymbol{\lambda}$ can be computed quickly by applying the known inverse functions. This yields a flexible parameterization of r, which admits for easy computation of both the gradients of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. We specify $r(\boldsymbol{\lambda}_0 \,|\, \mathbf{z})$ to be a factorized regression, which allows for low variance gradients.

Optimizing the Hierarchical ELBO We maximize the hierarchical ELBO with stochastic optimization (Robbins and Monro, 1951), which follows noisy, yet unbiased, gradients of the objective. The stochastic gradients of the hierarchical ELBO maintain the variance properties of the original variational approximation. We describe this when the original variational approximation is mean-field.

To optimize Eq.5 we need to compute the stochastic gradient with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Due to the the choice of differentiable prior, we can use the reparameterization gradient on $q(\boldsymbol{\lambda})$. Let $\boldsymbol{\epsilon}$ be a distribution drawn from a standard distribution s such as the standard Gaussian. Then let $\boldsymbol{\lambda}$ be written as a function of $\boldsymbol{\epsilon}$ and $\boldsymbol{\theta}$ denoted $\boldsymbol{\lambda}(\boldsymbol{\epsilon}; \boldsymbol{\theta})$. Next we define V to be the score function

$$V = \nabla_{\lambda} \log q(\mathbf{z} \,|\, \boldsymbol{\lambda}).$$

	Model	HIERARCHICAL VM	Baseline
Perplexity	100	**3570**	**3570**
	100-30	**3460**	3660
	100-30-15	**3480**	3550
NLL	100	**3.55**	3.63
	100-30	**3.53**	3.58
	100-30-15	**3.60**	**3.60**

Table 1: *NY Times*. Perplexity, and negative log-likelihood in units of 10^6 nats (lower is better).

Then the gradient of the hierarchical ELBO with respect to $\boldsymbol{\theta}$ is

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \widetilde{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = {} & \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \nabla_{\boldsymbol{\lambda}} \mathscr{L}_{\text{MF}}(\boldsymbol{\lambda})] \\ & + \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \nabla_{\boldsymbol{\lambda}}[\log r(\boldsymbol{\lambda} \,|\, \mathbf{z}; \boldsymbol{\phi}) - \log q(\boldsymbol{\lambda}; \boldsymbol{\theta})]] \\ & + \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \mathbb{E}_{q(\mathbf{z} \,|\, \boldsymbol{\lambda})}[V \log r(\boldsymbol{\lambda} \,|\, \mathbf{z}; \boldsymbol{\phi})]]. \end{aligned} \tag{8}$$

The first term is the gradient of the mean-field variational approximation scaled by the chain rule gradient from reparameterization. Thus hierarchical variational models inherit the variance reduced gradient (Eq.10) from the mean-field factorization. The second and third terms try to match r and q. The second term is strictly based on reparameterization, and thus exhibits low variance. The third term involves potentially a high variance gradient due to the appearance of all the latent variables. Since the distribution $q(\mathbf{z} \,|\, \lambda(\boldsymbol{\epsilon}; \boldsymbol{\theta}))$ factorizes by definition, we can apply the same variance reduction for r as for done in the mean-field with p. Thus hierarchical variational models maintains the variance properties of the stochastic gradient of the original variational family. The best choice of auxiliary model is given by a factorized regression. We detail this with a more thorough discussion in the appendix.

4 Experimental Results

We have introduced a new class of variational models. See appendix for efficient black box algorithms for their computation. We compare our proposed variational approximation to the mean-field approximation on deep exponential families (Ranganath et al., 2015), a class of hierarchical models where each observation is represented by multiple layers of exponential family random variables. We benchmark on a collection of news articles from *the New York Times*.

We focus on Poisson deep exponential families (DEF) up to depth three in particular, a multifeature generalization of the sigmoid belief network introduced by Neal (1990). In the sigmoid belief network each observation either turns a feature on or off, whereas in a Poisson DEF each observation expresses each feature a positive integer number of times.

HIERARCHICAL VM achieves better performance on both perplexity and held-out likelihood across all models. The mean-field approximation appears to be more sensitive to the architecture used in the DEF as evidenced by the poor performance of the two layer Poisson DEF on *the New York Times*. More experiments can be found in the appendix.

5 Discussion

We present hierarchical variational models: a tractable mechanism to introduce a richer class of posterior approximation constructed by placing priors on existing tractable variational families. We achieve a tractable objective by estimating a model of the "posterior" of the latent parameters in the hierarchical variational model. We show recovery on a toy example and better posterior approximations on a deep-discrete latent variable model. There are several possible avenues for future work such as exploring alternative bounds, analyzing our approach in the empirical Bayes framework, and iteratively expanding our recursive posterior approximation to more than simply q and r, which yields a form of annealed variational approximation.

References

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Dayan, P. (2000). Helmholtz machines and wake-sleep learning. *Handbook of Brain Theory and Neural Network. MIT Press, Cambridge, MA*.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Gershman, S., Hoffman, M., and Blei, D. (2012). Nonparametric variational inference. In *International Conference on Machine Learning*.

Jaakkola, T. S. and Jordan, M. I. (1998). Improving the Mean Field Approximation Via the Use of Mixture Distributions. In *Learning in Graphical Models*, pages 163–173. Springer Netherlands, Dordrecht.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kingma, D. and Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR-14)*.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. (2015). Automatic Variational Inference in Stan. In *Neural Information Processing Systems*.

Lawrence, N. (2000). *Variational Inference in Probabilistic Models*. PhD thesis.

Mnih, A. and Gregor, K. (2014). Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*.

Neal, R. (1990). Learning stochastic feedforward networks. *Tech. Rep. CRG-TR-90-7: Department of Computer Science, University of Toronto*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). {Black Box Variational Inference}. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 814–822.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. (2015). Deep Exponential Families. In *Artificial Intelligence and Statistics*.

Rezende, D., Mohamed, S., and Wierstra, D. (2014). Stochastic back-propagation and variational infernece in deep latent gaussian models. *ArXiv e-prints*.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 31st International Conference on Machine Learning (ICML-15)*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):pp. 400–407.

Salimans, T., Kingma, D., and Welling, M. (2015). Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Artifical Intelligence and Statistics*.

Salimans, T., Knowles, D. A., et al. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882.

Stan Development Team (2015). Stan: A c++ library for probability and sampling, version 2.8.0.

Stuhlmüller, A., Taylor, J., and Goodman, N. (2013). Learning stochastic inverses. In *Advances in neural information processing systems*, pages 3048–3056.

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational bayes for non-conjugate inference. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979.

Titsias, M. K. (2015). Local Expectation Gradients for Doubly Stochastic Variational Inference. In *Neural Information Processing Systems*.

Tran, D., Blei, D. M., and Airoldi, E. M. (2015). Variational inference with copula augmentation. In *Neural Information Processing Systems*.

Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Machine Learning*, pages 229–256.

Related Work. Variational models have a rich history. Inspired by kernel density estimation, this was classically studied by Jaakkola and Jordan (1998), and later revisited by Gershman et al. (2012), in which one specifies a mixture of Gaussians (MIXTURE) as the variational distribution. Lawrence (2000) explores a rich class of variational approximations formed by both mixtures and Markov chains. Recently, the idea of latent variables in the variational approximation has received some recent interest by Salimans et al. (2013, 2015), but they are limited to cases where the posterior of the variational latent variables is known or differentiable. These posterior models capture dependencies that are lost in simpler approximations.

There has been a line of work for variational approximations that capture dependencies in differentiable probability models (Titsias and Lázaro-Gredilla, 2014; Rezende and Mohamed, 2015; Salimans et al., 2015; Kucukelbir et al., 2015), but there has been limited work in variational methods that capture dependencies with discrete latent variables. Such methods typically apply the score function estimator (Ranganath et al., 2014; Mnih and Gregor, 2014). These methods can be used to posit rich approximations, but are either limited by the noise in stochastic gradients, or are quadratic in the number of latent variables. For example, Mnih and Gregor (2014) apply such techniques for variational approximations in sigmoid belief networks, but this approach is limited to stochastic feed foreword networks. Additionally, the variance increases as the number of layers increase. Tran et al. (2015) propose copulas (COPULA VI) as a way of learning dependencies in factorized approximations. Copulas can be extended to the framework of hierarchical variational models, whereas the direct approach taken in Tran et al. (2015) requires computation squared in the number of latent variables.

A Optimizing the Hierarchical ELBO

We maximize the hierarchical ELBO with stochastic optimization (Robbins and Monro, 1951), which follows noisy, yet unbiased, gradients of the objective.

Stochastic Gradient of the ELBO. The score function estimator for the gradient of the ELBO applies to both discrete and continuous latent variable models. It has strong roots in the policy search literature and evolutionary algorithms, where it is more commonly known as the REINFORCE gradient (Williams, 1992). It is given by

$$\nabla_{\boldsymbol{\lambda}}^{score}\mathscr{L} = \mathbb{E}_{q(\mathbf{z}\,|\,\boldsymbol{\lambda})}[\nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}\,|\,\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}\,|\,\boldsymbol{\lambda}))]. \tag{9}$$

See Ranganath et al. (2014) for a full derivation. We can construct noisy gradients from Eq.9 by Monte Carlo approximating the expectation. Formally, let S be the number of samples, the Monte Carlo estimate is

$$\frac{1}{S}\sum_{s=1}^{S} \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}^s\,|\,\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z}^s) - \log q(\mathbf{z}^s\,|\,\boldsymbol{\lambda})),$$
$$\text{where } \mathbf{z}^s \sim q(\mathbf{z}\,|\,\boldsymbol{\lambda}).$$

In general, the score function estimator exhibits high variance. This is not surprising given that the score function estimator makes very few restrictions on the class of models, and requires access only to zero-order information. Roughly, the variance of this estimator scales with the number of random variables (Ranganath et al., 2014; Mnih and Gregor, 2014).

In mean-field models, the gradient of the ELBO with respect to $\boldsymbol{\lambda}_i$ can be written as

$$\nabla_{\boldsymbol{\lambda}_i}\mathscr{L}_{\text{MF}} = \mathbb{E}_{q(\mathbf{z}_i;\boldsymbol{\lambda}_i)}[\nabla_{\boldsymbol{\lambda}_i} \log q(\mathbf{z}_i;\boldsymbol{\lambda}_i)(\log p_i(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}_i;\boldsymbol{\lambda}_i))], \tag{10}$$

where $\log p_i(\mathbf{x},\mathbf{z})$ are the components in the joint distribution that contain $\mathbf{z}_i$. This update is not only local, but it drastically reduces the variance of Eq.9 to make it computationally efficient.

In the case of differentiable latent variables, one would like to take advantage of model gradients. One such estimator does this using reparameterization: the ELBO is written in terms of a random variable $\boldsymbol{\epsilon}$, whose distribution s is free of the variational parameters and such that the original $\mathbf{z}$ can be written as a deterministic function $\mathbf{z} = \mathbf{z}(\boldsymbol{\epsilon}; \boldsymbol{\lambda})$. This allows gradients with respect to the variational parameters to directly propagate inside the expectation:

$$\nabla_{\boldsymbol{\lambda}}^{rep} \mathscr{L} = \mathbb{E}_{s(\boldsymbol{\epsilon})}[(\nabla_{\mathbf{z}} \log p(\mathbf{x}, \mathbf{z}) - \nabla_{\mathbf{z}} \log q(\mathbf{z})) \nabla_{\boldsymbol{\lambda}} \mathbf{z}(\boldsymbol{\epsilon}; \boldsymbol{\lambda})].$$

Similar to the score gradient, we can construct noisy gradients from this expression via Monte Carlo. Empirically reparameterization gradients have been shown to have much lower variance than the score function gradient (Titsias, 2015).

Stochastic Gradient of the Hierarchical ELBO. As discussed in Section 3, the variational prior $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$ can be constructed from both discrete and continuous distributions. However, due to the efficiency of Monte Carlo estimates for differentiable probability models using the reparameterization gradient, we focus on differentiable priors such as the normalizing flow.

To optimize Eq.5 we need to compute the stochastic gradient with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Due to the the choice of differentiable prior, we can use the reparameterization gradient on $q(\boldsymbol{\lambda})$. Let $\boldsymbol{\epsilon}$ be a distribution drawn from a standard distribution s such as the standard Gaussian. Then let $\boldsymbol{\lambda}$ be written as a function of $\boldsymbol{\epsilon}$ and $\boldsymbol{\theta}$ denoted $\boldsymbol{\lambda}(\boldsymbol{\epsilon}; \boldsymbol{\theta})$. Next we define V to be the score function

$$V = \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z} \,|\, \boldsymbol{\lambda}).$$

Then the gradient of the hierarchical ELBO with respect to $\boldsymbol{\theta}$ is

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \widetilde{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \nabla_{\boldsymbol{\lambda}} \mathscr{L}_{\text{MF}}(\boldsymbol{\lambda})] \\ + \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \nabla_{\boldsymbol{\lambda}}[\log r(\boldsymbol{\lambda} \,|\, \mathbf{z}; \boldsymbol{\phi}) - \log q(\boldsymbol{\lambda}; \boldsymbol{\theta})]] \\ + \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \mathbb{E}_{q(\mathbf{z} \,|\, \boldsymbol{\lambda})}[V \log r(\boldsymbol{\lambda} \,|\, \mathbf{z}; \boldsymbol{\phi})]]. \end{aligned} \tag{11}$$

The first term is the gradient of the mean-field variational approximation scaled by the chain rule gradient from reparameterization. Thus hierarchical variational models inherit the variance reduced gradient (Eq.10) from the mean-field factorization. The second and third terms try to match r and q. The second term is strictly based on reparameterization, and thus exhibits low variance. The third term involves potentially a high variance gradient due to the appearance of all the latent variables. Since the distribution $q(\mathbf{z} \,|\, \lambda(\boldsymbol{\epsilon}; \boldsymbol{\theta}))$ factorizes by definition, we can apply the same variance reduction for r as for done in the mean-field with p. We examine this below.

Local Learning with r. Let r_i be the terms $\log r(\boldsymbol{\lambda} \,|\, \mathbf{z}_i)$ containing $\mathbf{z}_i$, and define V_i to be the local score

$$V_i = \nabla_{\boldsymbol{\lambda}} \log q(\mathbf{z}_i \,|\, \boldsymbol{\lambda}_i).$$

Then the last term in Eq.11 can be transformed as

$$\mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}; \boldsymbol{\theta}) \mathbb{E}_{q(\mathbf{z} \,|\, \boldsymbol{\lambda})}[V \log r(\lambda \,|\, \mathbf{z}; \boldsymbol{\phi})]] = \mathbb{E}_{s(\boldsymbol{\epsilon})}\left[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}; \boldsymbol{\theta}) \mathbb{E}_{q(\mathbf{z} \,|\, \boldsymbol{\lambda})}\left[\sum_{i=1}^{d} V_i \log r_i(\boldsymbol{\lambda} \,|\, \mathbf{z}; \boldsymbol{\phi})\right]\right].$$

When r_i does not depend on too many variables, this gradient effectively combines both the computational efficiency of the mean-field and reparameterizations for a hierarchical variational models for both discrete and continuous latent variables. For example, in the variational model based on normalizing flows, the term r_i only depends on $\mathbf{z}_i$ as it is factorized. Positing the inverse model as a factorized regression the optimal choice in maintaining the learning signal of the mean-field.

Stochastic Gradient with respect to $\boldsymbol{\phi}$. Finally, as the expectation in the hierarchical ELBO (Eq.5) does not depend on $\boldsymbol{\phi}$, we can simply pass the gradient operator inside the expectation to obtain

$$\nabla_{\boldsymbol{\phi}} \widetilde{\mathscr{L}} = \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\lambda})}[\nabla_{\boldsymbol{\phi}} \log r(\boldsymbol{\lambda} \,|\, \mathbf{z}, \boldsymbol{\phi})]. \tag{12}$$

Algorithm 1: Black box inference with HIERARCHICAL VM

```
Input  : Model $\log p(\mathbf{x}, \mathbf{z})$,
         Variational model $q(z \mid \boldsymbol{\lambda}) q(\boldsymbol{\lambda}; \boldsymbol{\theta})$,
         Auxiliary Distribution $r(\boldsymbol{\lambda} \mid \mathbf{x}, \mathbf{z}; \boldsymbol{\phi})$
Output : Variational Parameters: $\boldsymbol{\theta}$
         Auxiliary Parameters: $\boldsymbol{\phi}$
Initialize $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$ randomly;
while change in ELBO is above some threshold do
    Compute unbiased estimate of $\nabla_{\boldsymbol{\theta}} \mathscr{L}$ using Eq.11 ;
    Compute unbiased estimate of $\nabla_{\boldsymbol{\phi}} \mathscr{L}$ using Eq.12 ;
    Update $\boldsymbol{\phi}$ and $\boldsymbol{\lambda}$ using stochastic gradient ascent ;
end
```

Algorithm. The inference procedure is outlined in Algorithm 1, where we evaluate noisy estimates of both gradients by sampling from the joint $q(\mathbf{z}, \boldsymbol{\lambda})$. In general, these gradients can be computed via automatic differentiation systems such as those available in Stan and Theano (Stan Development Team, 2015; Bergstra et al., 2010); this removes the need for model-specific computations, and moreover we note that no assumption has been made on $\log p(\mathbf{x}, \mathbf{z})$ other than the ability to calculate it.

Table 2 outlines black box methods and their complexity requirements. Hierarchical variational models equipped with a normalizing flow prior has complexity linear in the number of latent variables scaled by the length of the flow used to represent r and q.

Black box methods	Computation	Storage	Dependency	Class of models
BBVI (Ranganath et al., 2014)	$\mathcal{O}(D)$	$\mathcal{O}(D)$	✗	discrete/continuous
DSVI (Titsias and Lázaro-Gredilla, 2014)	$\mathcal{O}(D^2)$	$\mathcal{O}(D^2)$	✓	differentiable
COPULA VI (Tran et al., 2015)	$\mathcal{O}(D^2)$	$\mathcal{O}(D^2)$	✓	discrete/continuous
MIXTURE (Jaakkola and Jordan, 1998)	$\mathcal{O}(KD)$	$\mathcal{O}(KD)$	✓	discrete/continuous
NF (Rezende and Mohamed, 2015)	$\mathcal{O}(KD)$	$\mathcal{O}(KD)$	✓	differentiable
HIERARCHICAL VM w/ NF	$\mathcal{O}(KD)$	$\mathcal{O}(KD)$	✓	discrete/continuous

Table 2: A summary of black box inference methods. D is the number of latent variables; for MIXTURE, K is the number of mixture components; for NF procedures, K is the number of transformations.

Multi-level $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$. Multi-level hierarchical variational models can contain both discrete and differentiable latent variables. Higher level differentiable variables can be addressed by repeated use of the reparameterization trick. Discrete variables in the prior pose a difficulty as the learning signal used to estimate them has high variance. Local expectation gradients (Titsias, 2015) provide an efficient gradient estimator for variational approximations over discrete variables with small support, done by analytically marginalizing over each discrete variable individually. This approach can be combined with the gradient in Eq.11 to form an efficient gradient estimator.

Inference Networks. Classically, variational inference on models with random variables associated with each data point requires finding the optimal variational parameters for each data point's variational factor. This process can be computationally prohibitive especially at test time. The use of inference networks (Dayan, 2000; Stuhlmüller et al., 2013; Kingma and Welling, 2014; Rezende et al., 2014) amortizes the cost of estimating these local variational parameters by tying them together through a neural network. Specifically, the data point specific variational parameters are outputs to a neural network with the data point as input. The parameters of the neural network ζ then become the variational parameters; this reduces the cost of estimating the parameters of all the data points to estimating parameters of the inference network. Inference networks can be applied to hierarchical variational models by making both the parameters of the variational model and auxiliary distribution functions of their conditioning sets.

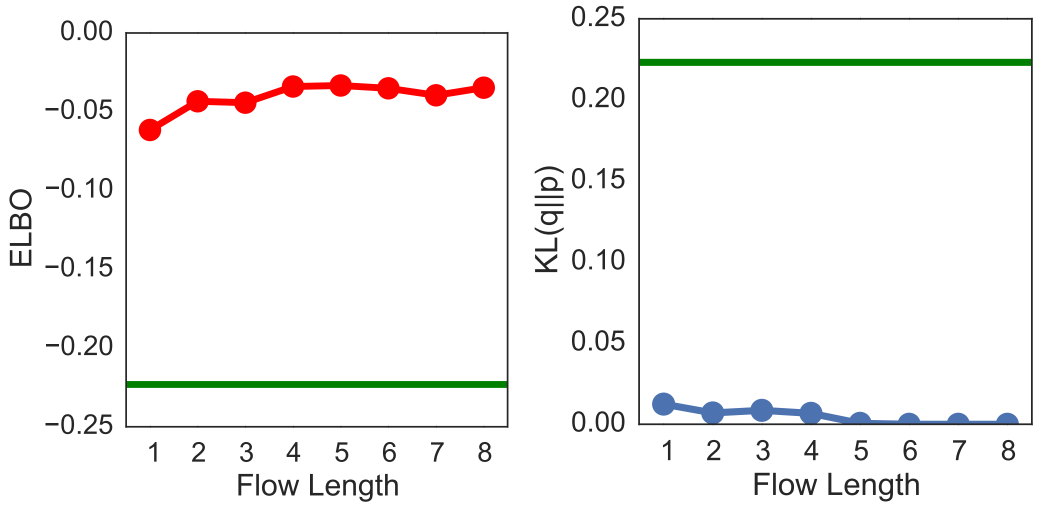

Figure 1: Expressivity of the hierarchical variational models, measured by ELBO and KL divergence after 10,000 iterations following Algorithm 1. Higher ELBOs and lower KLs are better. Green is the mean-field solution. The true KL divergence and the bound are better with HIERARCHICAL VM.

A.1 Toy Example: Correlated Bernoullis

Consider a toy model with no observations and two binary variables, defined by the following probability distribution:

	$\mathbf{z}_1 = 0$	$\mathbf{z}_1 = 1$
$\mathbf{z}_0 = 0$	0.1	0.4
$\mathbf{z}_1 = 1$	0.4	0.1

The mean-field approximation has a hard time capturing this distribution due to both the presence of negative correlation and that the negative correlation is not strong enough to bifurcate the optimization problem into two modes. Thus the optimal mean-field approximation is uniform.

Figure 1 plots the ELBO and KL-divergence for the hierarchical variational models and mean-field, where we specify a hierarchical variational models with prior and auxiliary distribution given by a normalizing flow. The KL is improved by over an order of magnitude, and a flow length of 8 exactly recovers the probability table up to numerical precision.