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 and finds better approximations to the posterior.

The key idea is this: we treat the original variational family as a "model" of the latent variables, and then expand it hierarchically as is commonly done in Bayesian statistics. The resulting variational model is a two-stage distribution that first draws variational parameters from a "prior" and then draws latent variables from the resulting variational distribution ("likelihood"). The "posterior" will play a role during inference, i.e., the conditional distribution of the variational parameters given a realization of the hidden variables. With a hierarchical variational model, rather than fit the mean-field parameters directly, one fits the hyperparameters of the prior. This construction maintains the computational efficiency

Seen as a traditional variational method, hierarchical variational models transform the variational family to the distribution obtained by marginalizing out the variational parameters under a prior. This expands the family from those that are directly parameterized, and thus can better approximate the posterior. Note the prior is a choice—for example, we will show how to use the recently-proposed normalizing flows (Rezende and Mohamed, 2015) as a prior to a mean-field distribution.

In this paper we define hierarchical variational models and develop a general algorithm for fitting them 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.

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 speci-

fies 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 forward 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 quadratic in the number of latent variables.

2 Black Box Variational Inference

Black box variational methods provide a mechanism to approximate posterior distributions with minimal analytic work. Let $p(\mathbf{z}\,|\,\mathbf{x})$ 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),

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

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 practitioner write a function to evaluate the model log-likelihood. Following the setting of black box inference, we now develop a framework for richer variational families.

3 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.

Fundamentally, black box methods also expose the loose requirements necessary for specifying an approximating distribution to the posterior. Such a model must satisfy three requirements: 1. a generating process for obtaining posterior samples $\mathbf{z}$, 2. computation of its log density function $\log q(\mathbf{z})$, and 3. proper support within the posterior.

Mean-field Variational Models The most common choice satisfying these properties 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 ith latent variable.

Mean-field approximations turn the intractable problem of computing the posterior into an computationally feasible optimization problem. However, its marginal factorization compromises the expressivity of the variational model: it abandons recovery of any dependence structure of the posterior, and it cannot in general capture all marginal information unless correctly specified. This predicates the need for more expressive variational models, which apply to both discrete and continuous variables yet remain computationally tractable. We develop hierarchical variational models to this extent.

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

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

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

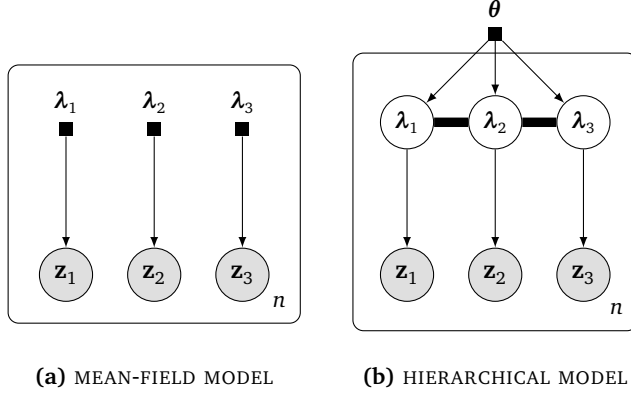**(a)** MEAN-FIELD MODEL   **(b)** HIERARCHICAL MODEL

Figure 1: Graphical model representation. **(a)** In mean-field models, the latent variables are strictly independent. **(b)** In hierarchical variational models, the latent variables are governed by a prior distribution on their parameters, which induces arbitrarily complex structure.

can either be a discrete or continuous distribution. Figure 1 displays this graphically. For simplicity, we focus on one level hierarchies.

The augmentation with a variational prior has strong ties to empirical Bayesian methods, which use data to estimate hyperparameters of a prior distribution (Robbins, 1964; Efron and Morris, 1973). In general, empirical Bayes considers the fully Bayesian treatment of a hyperprior on the original prior—here, the variational prior on the original mean-field—and proceeds to integrate it out. As this is analytically intractable, much work has been on parametric estimation, which seek point estimates rather than the whole distribution encoded by the hyperprior. We avoid this at the level of the hyperprior (variational prior) based on techniques discussed later; however, our procedure can still be viewed in this framework at one higher level. That is, we seek a point estimate of the "variational hyperprior" which governs the parameters on the variational prior.

A similar methodology also arises in the policy search literature (Rückstieß et al., 2008; Sehnke et al., 2008). Policy search methods aim to maximize the expected reward for a sequential decision-making task, by positing a distribution over trajectories and proceeding to learn its parameters. This distribution is known as the policy, and an upper-level policy considers a distribution over the original policy. This encourages exploration in the latent variable space and can be seen as a form of annealing.

The variational prior $q(\lambda; \theta)$ is parameterized by a vector θ. 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

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

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 $\log q_{\text{HVM}}(\mathbf{z}; \theta)$, 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 $r(\lambda \mid \mathbf{z}; \phi)$ with parameters ϕ, and applying the variational principle:

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

It can be shown that the bound is exact when $r(\lambda \mid \mathbf{z}; \phi)$ matches the *variational posterior* $q(\lambda \mid \mathbf{z}; \theta)$. From this perspective, we can view r as a recursive variational approximation. That is, it is a model for the posterior q of the mean-field parameters λ given a realization of the latent variables $\mathbf{z}$.

The bound is derived by introducing a term $\text{KL}(q\|r)$. Due to the asymmetry of KL-divergence, r could also be substituted into the first rather than the second argument of the KL divergence; this produces an alternative bound to Eq.4. The bound can also be extended to multi-level hierarchical variational models, where now we model the posterior distribution q of the higher levels using higher levels in r. Derivations of the bound and more details are available in the appendix.

We note that Eq.4 is tighter than the trivial conditional entropy bound of $\mathbb{H}[q_{\text{HVM}}] \geq \mathbb{H}[q \mid \lambda]$ (Cover and Thomas, 2012). This bound is attained when specifying the recursive approximation to be the prior, i.e., it is the special case when $r(\lambda \mid \mathbf{z}; \phi) = q(\lambda; \theta)$.

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 $\widetilde{\mathcal{L}}$:

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

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 θ and ϕ. Maximizing this bound is equivalent to minimizing an upper bound on the KL-divergence of q_{HVM} to the black box posterior $p(\mathbf{z} \mid \mathbf{x})$. Similar to the EM-algorithm (Dempster et al., 1977), optimizing θ improves the posterior approximation, while optimizing ϕ

tightens the upper bound on the KL divergence (improving the recursive variational approximation).

We can analyze Eq.5 by rewriting it in terms of the mean-field lower bound $\mathcal{L}_{\text{MF}}(\boldsymbol{\lambda})$,

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

This shows that Eq.5 is a sum of two terms: a Bayesian model average of mean-field objectives, with weights given by the variational prior $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$; and a correction term that's a function of both the auxiliary distribution r and the variational prior. Since mixtures (i.e. convex combinations) cannot be sharper than their components, r must not be conditionally independent of $\mathbf{z}$, in order for this bound to be better than the mean-field's.

4 Specifying the Hierarchical Variational Model

Specifying a hierarchical variational model requires two components: the variational likelihood $q(\mathbf{z} \mid \boldsymbol{\lambda})$ and the prior $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$. The likelihood factors $q(\mathbf{z}_i \mid \boldsymbol{\lambda}_i)$ can be chosen in the same way that mean-field factors are typically chosen. This leaves the specification of the prior distribution $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$. The prior distribution should satisfy two criteria. First, the components of the vector $\boldsymbol{\lambda}$ should not be independent as this reduces to the mean-field. Additionally, we want our choice of $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$ to allow for black box inference over models with both discrete and continuous latent variables. This criteria is discussed more directly in the next section when we discuss optimization with stochastic gradients.

Most variational models which move beyond mean-field can be extended to this framework. In general, one can take any tractable variational family typically used in variational inference, and specify it as the variational prior $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$. We outline several examples.

Mixture of Gaussians. A mixture of Gaussians can approximate arbitrary distributions given enough components and has been considered in traditional literature (Jaakkola and Jordan, 1998; Lawrence, 2000; Gershman and Blei, 2012). Thus they form an interesting class of prior distributions $q(\boldsymbol{\lambda}; \boldsymbol{\theta})$. Formally, let K be the number of components, $\boldsymbol{\pi}$ be a probability vector, $\boldsymbol{\mu}_k$, and $\boldsymbol{\sigma}_k$ be the parameters of a d-dimensional multivariate Gaussian with diagonal covariance. The variational prior is

$$q(\boldsymbol{\lambda}; \boldsymbol{\theta}) = \sum_{i=1}^{K} \pi_k \text{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k).$$

The parameters $\boldsymbol{\theta}$ consist of the probability vector $\boldsymbol{\pi}$ as well as the component means $\boldsymbol{\mu}_k$ and variances $\boldsymbol{\sigma}_k$. Relationships between different latent variables are captured by the mixture locations $\boldsymbol{\mu}_k$. For example, a two component mixture with two latent variables can capture that the latent variables are either very positive or very negative.

Normalizing Flows. Normalizing flows, which was previously introduced as a variational approximation for differentiable probability models in Rezende and Mohamed (2015), 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

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

The sequence of Jacobians describe the transport of probability mass through the flow. Functions that lead to complicated densities with efficient linear-time Jacobians can be created (Rezende and Mohamed, 2015). In this paper we focus on the planar flows $f(\boldsymbol{\lambda}) = \boldsymbol{\lambda} + uh(w^{\top}\boldsymbol{\lambda} + b)$, where u, w, and b are (variational) parameters and h is a non-linearity such as the hyperbolic tangent function. Let $\psi(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}}h(\boldsymbol{\lambda})$; then the flow density is

$$\log q(\boldsymbol{\lambda}) = \log q(\boldsymbol{\lambda}_0) - \sum_{k=1}^{K} \log\left(\left|1 + u_k^{\top}\psi(\boldsymbol{\lambda}_k)\right|\right).$$

We focus on this variational model in our experiments, so we describe the approximation r when the variational prior $q(\boldsymbol{\lambda})$ is a normalizing flow. The optimal r is the variational posterior $q(\boldsymbol{\lambda} \mid \mathbf{z})$. 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. Then the density of $\boldsymbol{\lambda}$ in r is

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

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 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:

$$r(\boldsymbol{\lambda}_0 \,|\, \mathbf{z}) = \prod_{i=1}^{d} r(\boldsymbol{\lambda}_{0_i} \,|\, \mathbf{z}_i). \qquad (8)$$

In words, the distribution r parameterizes a transformation under which the initial distribution factorizes.

Other Variational Models. Variational models can be constructed from many other classes of distributions. For example, copulas explicitly introduce dependence in a collection of d random variables with parameterized marginals, using joint distributions on the d-dimensional hypercubes (Nelsen, 2006). These can used as priors on either point mass or traditional mean-field likelihoods. A class of variational models can also be constructed by replacing the mixture model with a factorial mixture (Ghahramani, 1995). This leads to a much richer posterior approximation with the same number of parameters and latent variables $\mathbf{z}$.

5 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 REIN-FORCE gradient (Williams, 1992). It is given by

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

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

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

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). In the appendix, we explore an analytic equality for both gradients to explain the empirically observed difference in variance.

Stochastic Gradient of the Hierarchical ELBO. As discussed in Section 4, 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

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

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 $q(\mathbf{z} \,|\, \lambda(\epsilon; \theta))$ 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 $\log r(\lambda \,|\, \mathbf{z}_i)$ containing $\mathbf{z}_i$, and define V_i to be the local score

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

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

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

We derive this expression along with the full gradient in the appendix. 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 model 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 (Eq.8).

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

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

Algorithm. The inference procedure is outlined in Algorithm 1, where we evaluate noisy estimates of both gradients by sampling from the joint $q(\mathbf{z}, \lambda)$. 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 1 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.

Algorithm 1: Black box inference with HIERARCHICAL VM

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

Multi-level $q(\lambda; \theta)$. 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. We detail this in the appendix.

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.

6　Experimental Results

We have introduced a new class of variational models and developed efficient black box algorithms for their computation. We focus on hierarchical variational models defined by normalizing flows. We first consider a toy ex-

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; Lawrence, 2000)	$\mathcal{O}(Kd)$	$\mathcal{O}(Kd)$	✓	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 1: 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.
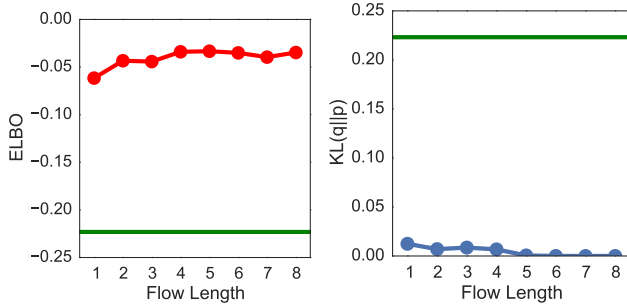

Figure 2: Expressivity of the hierarchical variational model, measured by ELBO and KL divergence after 10,000 iterations following Algorithm 1. Higher ELBOs and lower KLs are better. Green is the mean-field solution. The true KL divergence and the bound are better with HIERARCHICAL VM.

ample. Then we compare our proposed variational approximations to more standard ones 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.

6.1 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 2 plots the ELBO and KL-divergence for the hierarchical variational model and mean-field, where we specify

a hierarchical variational model 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.

6.2 Deep Exponential Families

The deep exponential family (DEF) (Ranganath et al., 2015) forms a class of probabilistic models built from exponential families (Brown, 1986) whose latent structure parallels the architectures used in deep neural networks.

Model. Exponential families are parameterized by a set of natural parameters. We denote a draw from an unspecified exponential family with natural parameter η as EXPFAM(η). The natural parameter in deep exponential families are constructed from an inner product of the previous layer with shared weights passed through a link function g.

Let $\mathbf{x}_i$ be an observation, $\mathbf{z}_{i,\ell}$ be a vector of latent variables with $\mathbf{z}_{i,\ell,k}$ as an element of that vector, L be the total number of layers and $\mathbf{W}_{\ell,k}$ be a shared vector of random variables across observations, then deep exponential families are

$$\mathbf{W}_{\ell,k} \sim \text{EXPFAM}_W(\xi)$$
$$\mathbf{z}_{i,L,k} \sim \text{EXPFAM}_L(\eta)$$
$$\mathbf{z}_{i,\ell,k} \sim \text{EXPFAM}_\ell(g_\ell(\mathbf{z}_{i,l+1}^\top \mathbf{W}_{l,k})$$
$$\mathbf{x}_{i,n} \sim \text{Poisson}(\mathbf{W}_{0,n}^\top \mathbf{z}_{i,1}).$$

Here, we specialized the likelihood on $\mathbf{x}$ for count data which we focus on in our experiments.

We focus on Poisson DEFS in particular. In a sigmoid belief network introduced by Neal (1990), each observation either turns a feature on or off, while in a Poisson DEF each observation expresses each feature a positive integer number of times. This means Poisson DEFS are a multifeature generalization of the classic sigmoid belief network.

The Poisson DEF has Poisson distributed latent variables and normal weights with the log-softplus link function. This means the conditional distribution is

$$p(\mathbf{z}_{i,\ell,k} \mid \mathbf{z}_{i,l+1}, \mathbf{W}_{l,k}) \sim \text{Poisson}(\log(1 + \exp(\mathbf{z}_{i,l+1}^\top \mathbf{W}_{l,k}))).$$

Variational Models   We consider the variational approximation that adds dependence to the $\mathbf{z}'s$. We parameterize each variational prior $q(\boldsymbol{\lambda}_{\mathbf{z}_i})$ with a normalizing planar flow of length 2, and use the inverse flow of length 10 for $r(\boldsymbol{\lambda}_{\mathbf{z}_i})$. In a pilot study, we found little improvement with longer flow lengths. We compare to the mean-field approximation from Ranganath et al. (2015).

Data.   We consider two text corpora of news and scientific articles: *The New York Times* and *Science*. Both datasets have 11K documents. *The New York Times* consists of 8K terms and *Science* consists of 5.9K terms.

Evaluation.   We examine both held out perplexity and the bound on the evidence for held out data.

To compute held-out perplexity, we use ten percent of each held-out observation to compute the variational parameters associated with that observation, then compute the perplexity on the remaining 90 percent. Held out perplexity is

$$\exp\left(\frac{-\sum_d \sum_{w \in d} \log p(w \mid \# \text{ held out in } d)}{N_{\text{held out}}}\right)$$

This is a document complete evaluation metric (Wallach et al., 2009) where the words are tested independently.

The bound on the held out evidence is the ELBO for the mean-field approximation, and the hierarchical ELBO for the hierarchical variational model. Note that the held out evidence directly measures how well the posterior is approximated by the variational model. On the other hand, held out perplexity may be affected by the fact that good posterior estimates might overfit on test data.

Hyperparameters.   We study one, two, and three layer DEFS with 100, 30, and 15 units respectively and set prior hyperparameters following Ranganath et al. (2015). We use Nesterov's accelerated gradient with momentum parameter of 0.9, combined with RMSProp with a scaling factor of 10^{-2}, to maximize the lower bound.

Results.   HIERARCHICAL VM achieves better performance on both perplexity and held-out likelihood across all data sets and models. The two-layer Poisson DEF achieves the lowest perplexities for both data sets, which we conjecture is due to the need for a wider number of latent variables available in the third layer (more than 15). The mean-field approximation appears to be more sensitive to the

	Model	HIERARCHICAL VM	Mean-Field
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 2: *New York Times*. Perplexity, and negative log-likelihood in units of 10^6 nats (lower is better). HIERARCHICAL VM provides a more accurate representation of the posterior and predicts better than mean-field variational methods.

	Model	HIERARCHICAL VM	Mean-Field
Perplexity	100	**3360**	3377
	100-30	**3080**	3240
	100-30-15	**3110**	3190
NLL	100	**0.64**	0.71
	100-30	**2.07**	2.89
	100-30-15	**2.78**	2.89

Table 3: *Science*. Perplexity, and negative log-likelihood in units of 10^6 nats (lower is better). HIERARCHICAL VM outperforms mean-field on both metrics.

architecture used in the DEF as evidenced by the poor performance of the two layer Poisson DEF on *The New York Times*. We also observe similar performance when training sigmoid belief networks with varying layers on both data sets.

7   Discussion

We present hierarchical variational models: a Bayesian framework for introducing a richer class of posterior approximations, constructed by placing priors on existing tractable variational families. We achieve a tractable objective by estimating a model of the "posterior" of the variational parameters in the hierarchical variational model. We show recovery on a toy example and better posterior approximations on deep discrete latent variable models. There are several avenues for future work such as exploring alternative bounds and analyzing our approach in the empirical Bayes framework. Lastly, new kinds of hierarchical variational model can be explored by placing priors on other tractable variational families and iteratively expanding our recursive posterior approximation to more than simply q and r; this 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.

Brown, L. (1986). *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, CA.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

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.

Efron, B. and Morris, C. (1973). Combining possibly related estimation problems. *Journal of the Royal Statistical Society, Series B*, 35(3):379–421.

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

Gershman, S. J. and Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1 – 12.

Ghahramani, Z. (1995). Factorial learning and the em algorithm. In *Advances in Neural Information Processing Systems*, pages 617–624.

Hoffman, M. D. and Blei, D. M. (2015). Structured Stochastic Variational Inference. In *Artificial Intelligence and Statistics*.

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., editor (1999). *Learning in Graphical Models*. MIT Press, Cambridge, MA.

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.

Minka, T. (2001). *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology.

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*.

Nelsen, R. B. (2006). *An Introduction to Copulas (Springer Series in Statistics)*. Springer-Verlag New York, Inc.

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. (1964). The empirical bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, pages 1–20.

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

Rückstieß, T., Felder, M., and Schmidhuber, J. (2008). State-dependent exploration for policy gradient methods. In *Machine Learning and Knowledge Discovery in Databases*, pages 234–249. Springer.

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.

Sehnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., and Schmidhuber, J. (2008). Policy gradients with parameter-based exploration for control. In *Artificial Neural Networks-ICANN 2008*, pages 387–396. Springer.

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). Copula variational inference. 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.

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *International Conference on Machine Learning (ICML)*.

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

A Appendix

Tractable bound on the entropy.   Deriving an analytic expression for the entropy of q_{HVM} is generally intractable due to the integral in the definition of q_{HVM}. However, it is tractable when we know the distribution $q(\lambda \,|\, \mathbf{z})$. This can be seen by noting from standard Bayes' rule that

$$q(\mathbf{z})q(\lambda \,|\, \mathbf{z}) = q(\lambda)q(\mathbf{z} \,|\, \lambda). \tag{13}$$

The right hand side is specified by the construction of the hierarchical variational model. Note also that $q(\lambda \,|\, \mathbf{z})$ can be interpreted as the posterior distribution of the original variational parameters λ given the model, thus we will denote it as $q_{\text{POST}}(\lambda \,|\, \mathbf{z})$.

In general, computing $q_{\text{POST}}(\lambda \,|\, \mathbf{z})$ from the specification of the hierarchical variational model is as hard as the integral needed to compute the entropy from Eq.3. Instead, we approximate q_{POST} with an auxiliary distribution $r(\lambda \,|\, \mathbf{z}; \phi)$ parameterized by ϕ. This yields a bound on the entropy in terms of the analytically known distributions $r(\lambda \,|\, \mathbf{z})$, $q(\mathbf{z} \,|\, \lambda)$, and $q(\lambda)$.

First note that the KL-divergence between two distributions is greater than zero, and is precisely zero only when the two distributions are equal. This means the entropy can be bounded as follows:

$$
\begin{aligned}
&-\mathbb{E}_{q_{\text{HVM}}}[\log q_{\text{HVM}}(\mathbf{z})] \\
&\quad = -\mathbb{E}_{q_{\text{HVM}}}[\log q_{\text{HVM}}(\mathbf{z}) - \text{KL}(q_{\text{POST}}(\lambda \,|\, \mathbf{z})||q_{\text{POST}}(\lambda \,|\, \mathbf{z}))] \\
&\quad \geq -\mathbb{E}_{q_{\text{HVM}}}[\log q_{\text{HVM}}(\mathbf{z}) + \text{KL}(q_{\text{POST}}(\lambda \,|\, \mathbf{z})||r(\lambda \,|\, \mathbf{z}; \phi))]] \\
&\quad = -\mathbb{E}_{q_{\text{HVM}}}[\mathbb{E}_{q_{\text{POST}}}[\log q_{\text{HVM}}(\mathbf{z}) + \log q_{\text{POST}}(\lambda \,|\, \mathbf{z}) \\
&\quad\quad\quad - \log r(\lambda \,|\, \mathbf{z}; \phi)]] \\
&\quad = -\mathbb{E}_{q(\mathbf{z},\lambda)}[\log q_{\text{HVM}}(\mathbf{z}) + \log q_{\text{POST}}(\lambda \,|\, \mathbf{z}) - \log r(\lambda \,|\, \mathbf{z}; \phi)].
\end{aligned}
$$

Then by Eq.13, the bound simplifies to

$$
\begin{aligned}
&-\mathbb{E}_{q_{\text{HVM}}}[\log q_{\text{HVM}}(\mathbf{z})] \\
&\quad \geq -\mathbb{E}_{q(\mathbf{z},\lambda)}[\log q(\lambda) + \log q(\mathbf{z} \,|\, \lambda) - \log r(\lambda \,|\, \mathbf{z}; \phi)].
\end{aligned}
$$

A similar bound in derived by Salimans et al. (2015) directly for $\log p(x)$.

In the above derivation, the approximation r to the variational posterior $q_{\text{POST}}(\lambda \,|\, \mathbf{z})$ is placed as the second argument of a KL-divergence term. Replacing the first argument instead yields a different tractable upper bound as

well.

$$
\begin{aligned}
&-\mathbb{E}_{q_{\text{HVM}}}[\log q(\mathbf{z})] \\
&\quad = \mathbb{E}_{q_{\text{HVM}}}[-\log q(\mathbf{z}) + \text{KL}(q_{\text{POST}}(\lambda \,|\, \mathbf{z})||q_{\text{POST}}(\lambda \,|\, \mathbf{z}))] \\
&\quad \leq \mathbb{E}_{q_{\text{HVM}}}[-\log q(\mathbf{z}) + \text{KL}(r(\lambda \,|\, \mathbf{z}; \phi)||q_{\text{POST}}(\lambda \,|\, \mathbf{z}))]] \\
&\quad = \mathbb{E}_{q_{\text{HVM}}}[\mathbb{E}_r[-\log q(\mathbf{z}) - \log q_{\text{POST}}(\lambda \,|\, \mathbf{z}) + \log r(\lambda \,|\, \mathbf{z}; \phi)]] \\
&\quad = \mathbb{E}_{q_{\text{HVM}}}\left[\mathbb{E}_r\left[-\log q(\mathbf{z}) - \log \frac{q(\mathbf{z} \,|\, \lambda)q(\lambda)}{q(\mathbf{z})} + \log r(\lambda \,|\, \mathbf{z}; \phi)\right]\right] \\
&\quad = \mathbb{E}_{q_{\text{HVM}}}[\mathbb{E}_r[-\log q(\lambda) - \log q(\mathbf{z} \,|\, \lambda) + \log r(\lambda \,|\, \mathbf{z}; \phi)]].
\end{aligned}
$$

The bound is also tractable when r and q_{HVM} can be sampled and all distributions are analytic. The derivation of these two bounds parallels the development of expectation propagation (Minka, 2001) and variational Bayes (Jordan, 1999) which are based on alternative forms of the KL-divergence[1]. Exploring the role and relative merits of both bounds we derive in the context of variational models will be an important direction in the study of variational models with latent variables.

Gradient Derivation.   We derive the gradient of the hierarchical ELBO using its mean-field representation:

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

Using the reparameterization $\lambda(\epsilon; \theta)$, where $\epsilon \sim s$, this is

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

We now differentiate the three additive terms with respect to θ. As in the main text, we suppress θ in the definition of λ when clear and define the score function

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

By the chain rule the derivative of the first term is

$$\nabla_\theta \mathbb{E}_{s(\epsilon)}[\mathscr{L}(\lambda(\epsilon; \theta))] = \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon) \nabla_\lambda \mathscr{L}(\lambda)].$$

We now differentiate the second term:

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{s(\epsilon)}[\mathbb{E}_{q(\mathbf{z}|\lambda)}[\log r(\lambda(\epsilon; \theta) \,|\, \mathbf{z}; \phi)]] \\
&\quad = \nabla_\theta \mathbb{E}_{s(\epsilon)}\left[\left[\int q(\mathbf{z} \,|\, \lambda) \log r(\lambda(\epsilon; \theta) \,|\, \mathbf{z}; \phi)\, \mathrm{d}\mathbf{z}\right]\right] \\
&\quad = \mathbb{E}_{s(\epsilon)}\left[\nabla_\theta\left[\left[\int q(\mathbf{z} \,|\, \lambda) \log r(\lambda(\epsilon; \theta) \,|\, \mathbf{z}; \phi)\, \mathrm{d}\mathbf{z}\right]\right]\right] \\
&\quad = \mathbb{E}_{s(\epsilon)}\left[\nabla_\theta \lambda(\epsilon) \nabla_\lambda\left[\left[\int q(\mathbf{z} \,|\, \lambda) \log r(\lambda(\epsilon; \theta) \,|\, \mathbf{z}; \phi)\, \mathrm{d}\mathbf{z}\right]\right]\right].
\end{aligned}
$$

[1]Note that the first bound, which corresponds to the objective in expectation propagation (EP), directly minimizes $\text{KL}(q\|r)$ whereas EP only minimizes this locally.

Applying the product rule to the inner derivative gives

$$\nabla_\lambda \left[\int q(\mathbf{z}\,|\,\lambda) \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)\,d\mathbf{z} \right]$$

$$= \int \nabla_\lambda q(\mathbf{z}\,|\,\lambda) \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)\,d\mathbf{z}$$

$$+ \int q(\mathbf{z}\,|\,\lambda)\nabla_\lambda \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)\,d\mathbf{z}$$

$$= \int \nabla_\lambda \log q(\mathbf{z}\,|\,\lambda)q(\mathbf{z}\,|\,\lambda) \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)\,d\mathbf{z}$$

$$+ \int q(\mathbf{z}\,|\,\lambda)\nabla_\lambda \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)\,d\mathbf{z}$$

$$= \mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[V \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)]$$

$$+ \mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[\nabla_\lambda \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)].$$

Substituting this back into the previous expression gives the gradient of the second term

$$\mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon)\mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[V \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)]]$$

$$+ \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon)\mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[\nabla_\lambda \log r(\lambda(\epsilon;\theta)\,|\,\mathbf{z};\phi)]]$$

The third term also follows by the chain rule

$$\nabla_\theta \mathbb{E}_{s(\epsilon)}[\log q(\lambda(\epsilon;\theta);\theta)]$$

$$= \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon)\nabla_\lambda \log q(\lambda;\theta) + \nabla_\theta \log q(\lambda;\theta)]$$

$$= \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon)\nabla_\lambda \log q(\lambda;\theta)]$$

where the last equality follows by

$$\mathbb{E}_{s(\epsilon)}[\nabla_\theta \log q(\lambda;\theta)] = \mathbb{E}_{q(\lambda;\theta)}[\nabla_\theta \log q(\lambda;\theta)] = \mathbf{0}.$$

Combining these together gives the total expression for the gradient

$$\nabla_\theta \widetilde{L}(\theta,\phi) = \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon)\nabla_\lambda \mathcal{L}_{\mathrm{MF}}(\lambda)]$$

$$+ \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon)\nabla_\lambda[\log r(\lambda\,|\,\mathbf{z};\phi) - \log q(\lambda;\theta)]]$$

$$+ \mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon)\mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[V \log r(\lambda\,|\,\mathbf{z};\phi)]].$$

Introducing r_i to the gradient. One term of the gradient involves the product of the score function with all of r,

$$\mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon)\mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[V \log r(\lambda\,|\,\mathbf{z};\phi)]].$$

Localizing (Rao-Blackwellizing) the inner expectation as in Ranganath et al. (2014); Mnih and Gregor (2014) can drastically reduce the variance. Recall that

$$q(\mathbf{z}\,|\,\lambda) = \prod_{i=1}^d q(\mathbf{z}_i\,|\,\lambda_i).$$

Next, we define V_i to be the score functions of the factor. That is

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

This is a vector with nonzero entries corresponding to λ_i. Substituting the factorization into the gradient term yields

$$\mathbb{E}_{s(\epsilon)} \left[\nabla_\theta \lambda(\epsilon) \sum_{i=1}^d \mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[V_i \log r(\lambda\,|\,\mathbf{z};\phi)] \right]. \quad (14)$$

Now we define r_i to be the terms in $\log r$ containing $\mathbf{z}_i$ and r_{-i} to be the remaining terms. Then the inner expectation in the gradient term is

$$\sum_{i=1}^d \mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[V_i(\log r_i(\lambda\,|\,\mathbf{z};\phi) + \log r_{-i}(\lambda\,|\,\mathbf{z};\phi))]$$

$$= \sum_{i=1}^d \mathbb{E}_{q(\mathbf{z}_i\,|\,\lambda)}[V_i\mathbb{E}_{q(\mathbf{z}_{-i}\,|\,\lambda)}[\log r_i(\lambda\,|\,\mathbf{z};\phi) + \log r_{-i}(\lambda\,|\,\mathbf{z};\phi)]],$$

$$= \sum_{i=1}^d \mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[V_i \log r_i(\lambda\,|\,\mathbf{z};\phi)],$$

where the last equality follows from the expectation of the score function of a distribution is zero. Substituting this back into Eq.14 yields the desired result

$$\mathbb{E}_{s(\epsilon)}[\nabla_\theta \lambda(\epsilon;\theta)\mathbb{E}_{q(\mathbf{z}\,|\,\lambda)}[V \log r(\lambda\,|\,\mathbf{z};\phi)]]$$

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

Equality of Two Gradients. We now provide a direct connection between the score gradient and the reparameterization gradient. We carry this out in one-dimension for clarity, but the same principle holds in higher dimensions. Let Q be the cumulative distribution function (CDF) of q and let $z = T(\mathbf{z}_0;\lambda)$ be reparameterizable in terms of a uniform random variable $\mathbf{z}_0$ (inverse-CDF sampling). We focus on the one dimensional case for simplicity. Recall integration by parts computes a definite integral as

$$\int_{supp(\mathbf{z})} w(\mathbf{z})\,dv(\mathbf{z}) = |w(\mathbf{z})v(\mathbf{z})|_{supp(\mathbf{z})} - \int_{supp(\mathbf{z})} v(\mathbf{z})\,dw(\mathbf{z}),$$

where the $|\cdot|$ notation indicates evaluation of a portion of the integral. In the subsequent derivation, we let $w(\mathbf{z}) = \log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z})$, and let $dv(\mathbf{z}) = \nabla_\lambda \log q(\mathbf{z})q(\mathbf{z}) = \nabla_\lambda q(\mathbf{z})$.

Recall that we assume that we can CDF-transform $\mathbf{z}$ and that the transformation is differentiable. That is, when $\mathbf{u}$ is a standard uniform random variable, $\mathbf{z} = Q^{-1}(\mathbf{u},\lambda)$.

Then

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

$$= \int_{supp(\mathbf{z})} \nabla_\lambda q(\mathbf{z}|\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda}))]\,d\mathbf{z}$$

$$= \left| \left[\int_{\mathbf{z}} \nabla_\lambda q(\mathbf{z}|\boldsymbol{\lambda})\,d\mathbf{z} \right] (\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda})) \right|_{supp(\mathbf{z})}$$

$$\quad - \int \left[\int_{\mathbf{z}} \nabla_\lambda q(\mathbf{z}|\boldsymbol{\lambda})\,d\mathbf{z} \right] \nabla_{\mathbf{z}}[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda})]\,d\mathbf{z}$$

$$= \left| \nabla_\lambda Q(\mathbf{z}|\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda})) \right|_{supp(\mathbf{z})}$$

$$\quad - \int \nabla_\lambda \left[Q(\mathbf{z}|\boldsymbol{\lambda}) \right] \nabla_{\mathbf{z}}[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda})]\,d\mathbf{z}$$

$$= \left| \nabla_\lambda Q(\mathbf{z}|\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda})) \right|_{supp(\mathbf{z})}$$

$$\quad + \int q(\mathbf{z}|\boldsymbol{\lambda})\nabla_\lambda \left[\mathbf{z} \right] \nabla_{\mathbf{z}}[\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda})]\,d\mathbf{z}$$

$$= \left| \nabla_\lambda Q(\mathbf{z}|\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z}) - \log q(\mathbf{z}|\boldsymbol{\lambda})) \right|_{supp(\mathbf{z})}$$

$$\quad + \nabla_\lambda^{rep} \mathscr{L},$$

where the second to last equality follows by the derivative of the CDF function (Hoffman and Blei, 2015). By looking at the Monte Carlo expression of both sides, we can see the reduction in variance that the reparameterization gradient has over the score gradient comes from the analytic computation of the gradient of the definite integral (which has value **0**).

Optimizing with Discrete Variables in the Variational Prior. In the setting where the prior has discrete variables, optimization requires a little more work. First we note that in a non-degenerate mean-field setup that the $\boldsymbol{\lambda}$'s are differentiable parameters of probability distributions. This means they will always, conditional on the discrete variables, be differentiable in the variational prior. This means that we can both compute the gradients for these parameters using the technique from above and that the discrete variables exist at a higher level of the hierarchical variational model. The gradients of discrete variables can be computed using the score gradient, but Monte Carlo estimates of this will have high variance due to no simplification of the learning signal (like in the mean-field). We can step around this issue by using local expectation gradients (Titsias, 2015) which marginalize out one variable at a time to get low variance stochastic gradients. This is generally tractable when the discrete variables have small support such as in binary variables.