Hierarchical Variational Models (Appendix)

Rajesh Ranganath RAJESHR@CS.PRINCETON.EDU
Princeton University, 35 Olden St., Princeton, NJ 08540

Dustin Tran DUSTIN@CS.COLUMBIA.EDU
David M. Blei DAVID.BLEI@COLUMBIA.EDU
Columbia University, 500 W 120th St., New York, NY 10027

Relationship to empirical Bayes and RL. 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 & 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) via the hierarchical ELBO; however, our procedure can 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.

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 $q(\boldsymbol{\lambda} \mid \mathbf{z})$. This can be seen by noting from standard Bayes' rule that

$$q(\mathbf{z})q(\boldsymbol{\lambda} \mid \mathbf{z}) = q(\boldsymbol{\lambda})q(\mathbf{z} \mid \boldsymbol{\lambda}), \tag{1}$$

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

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

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:

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

Then by Eq. 1, the bound simplifies to

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

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 $q_{\text{POST}}(\boldsymbol{\lambda} \mid \mathbf{z})$ is placed as the second argument of a KL-divergence term. Replacing the first argument instead yields a different tractable upper bound as well.

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

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.

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

Gradient Derivation. We derive the gradient of the hierarchical Evidence Lower BOund (ELBO) using its mean-field representation:

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

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

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

We now differentiate the three additive terms with respect to $\boldsymbol{\theta}$. As in the main text, we suppress $\boldsymbol{\theta}$ in the definition of $\boldsymbol{\lambda}$ when clear and define the score function

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

By the chain rule the derivative of the first term is

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

We now differentiate the second term:

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

Applying the product rule to the inner derivative gives

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

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

$$\begin{aligned}&\mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \mathbb{E}_{q(\mathbf{z} \mid \boldsymbol{\lambda})}[V \log r(\boldsymbol{\lambda}(\boldsymbol{\epsilon}; \boldsymbol{\theta}) \mid \mathbf{z}; \boldsymbol{\phi})]] \\ &\quad + \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \mathbb{E}_{q(\mathbf{z} \mid \boldsymbol{\lambda})}[\nabla_{\boldsymbol{\lambda}} \log r(\boldsymbol{\lambda}(\boldsymbol{\epsilon}; \boldsymbol{\theta}) \mid \mathbf{z}; \boldsymbol{\phi})]]\end{aligned}$$

The third term also follows by the chain rule

$$\begin{aligned}&\nabla_{\boldsymbol{\theta}} \mathbb{E}_{s(\boldsymbol{\epsilon})}[\log q(\boldsymbol{\lambda}(\boldsymbol{\epsilon}; \boldsymbol{\theta}); \boldsymbol{\theta})] \\ &\quad = \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\lambda}; \boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} \log q(\boldsymbol{\lambda}; \boldsymbol{\theta})] \\ &\quad = \mathbb{E}_{s(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\theta}} \boldsymbol{\lambda}(\boldsymbol{\epsilon}) \nabla_{\boldsymbol{\lambda}} \log q(\boldsymbol{\lambda}; \boldsymbol{\theta})]\end{aligned}$$

where the last equality follows by

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

Combining these together gives the total expression for the gradient

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

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

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

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

$$q(\mathbf{z} \mid \boldsymbol{\lambda}) = \prod_{i=1}^{d} q(z_i \mid \boldsymbol{\lambda}_i).$$

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

$$V_i = \nabla_{\boldsymbol{\lambda}} \log q(z_i \mid \boldsymbol{\lambda}_i).$$

[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.

This is a vector with nonzero entries corresponding to $\boldsymbol{\lambda}_i$. Substituting the factorization into the gradient term yields

$$\mathbb{E}_{s(\boldsymbol{\epsilon})}\left[\nabla_{\boldsymbol{\theta}}\boldsymbol{\lambda}(\boldsymbol{\epsilon})\sum_{i=1}^{d}\mathbb{E}_{q(\mathbf{z}\mid\boldsymbol{\lambda})}[V_i\log r(\boldsymbol{\lambda}\mid\mathbf{z};\boldsymbol{\phi})]\right]. \quad (2)$$

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

$$\begin{aligned}
&\sum_{i=1}^{d}\mathbb{E}_{q(\mathbf{z}\mid\boldsymbol{\lambda})}[V_i(\log r_i(\boldsymbol{\lambda}\mid\mathbf{z};\boldsymbol{\phi})+\log r_{-i}(\boldsymbol{\lambda}\mid\mathbf{z};\boldsymbol{\phi}))]\\
&\quad=\sum_{i=1}^{d}\mathbb{E}_{q(z_i\mid\boldsymbol{\lambda})}[V_i\mathbb{E}_{q(\mathbf{z}_{-i}\mid\boldsymbol{\lambda})}[\log r_i(\boldsymbol{\lambda}\mid\mathbf{z};\boldsymbol{\phi})+\log r_{-i}(\boldsymbol{\lambda}\mid\mathbf{z};\boldsymbol{\phi})]]\\
&\quad=\sum_{i=1}^{d}\mathbb{E}_{q(\mathbf{z}\mid\boldsymbol{\lambda})}[V_i\log r_i(\boldsymbol{\lambda}\mid\mathbf{z};\boldsymbol{\phi})],
\end{aligned}$$

where the last equality follows from the expectation of the score function of a distribution is zero. Substituting this back into Eq. 2 yields the desired result

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

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 $z=T(\mathbf{z}_0;\boldsymbol{\lambda})$ 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

$$\begin{aligned}
&\int_{\text{supp}(\mathbf{z})}w(\mathbf{z})\,\mathrm{d}v(\mathbf{z})\\
&\qquad=|w(\mathbf{z})v(\mathbf{z})|_{\text{supp}(\mathbf{z})}-\int_{\text{supp}(\mathbf{z})}v(\mathbf{z})\,\mathrm{d}w(\mathbf{z}),
\end{aligned}$$

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 $dv(\mathbf{z})=\nabla_{\boldsymbol{\lambda}}\log q(\mathbf{z})q(\mathbf{z})=\nabla_{\boldsymbol{\lambda}}q(\mathbf{z})$.

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, $\mathbf{z}=Q^{-1}(\mathbf{u},\boldsymbol{\lambda})$. Then

$$\begin{aligned}
\nabla^{score}_{\boldsymbol{\lambda}}\mathcal{L}&=\mathbb{E}_{q(\mathbf{z}\mid\boldsymbol{\lambda})}[\nabla_{\boldsymbol{\lambda}}\log q(\mathbf{z}\mid\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda}))]\\
&=\int_{supp(\mathbf{z})}\nabla_{\boldsymbol{\lambda}}q(\mathbf{z}\mid\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda}))]\,\mathrm{d}\mathbf{z}\\
&=\left|\left[\int_{\mathbf{z}}\nabla_{\boldsymbol{\lambda}}q(\mathbf{z}\mid\boldsymbol{\lambda})\,\mathrm{d}\mathbf{z}\right](\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda}))\right|_{supp(\mathbf{z})}\\
&\quad-\int\left[\int_{\mathbf{z}}\nabla_{\boldsymbol{\lambda}}q(\mathbf{z}\mid\boldsymbol{\lambda})\,\mathrm{d}\mathbf{z}\right]\nabla_{\mathbf{z}}[\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda})]\,\mathrm{d}\mathbf{z}\\
&=|\nabla_{\boldsymbol{\lambda}}Q(\mathbf{z}\mid\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda}))|_{supp(\mathbf{z})}\\
&\quad-\int\nabla_{\boldsymbol{\lambda}}\left[Q(\mathbf{z}\mid\boldsymbol{\lambda})\right]\nabla_{\mathbf{z}}[\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda})]\,\mathrm{d}\mathbf{z}\\
&=|\nabla_{\boldsymbol{\lambda}}Q(\mathbf{z}\mid\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda}))|_{supp(\mathbf{z})}\\
&\quad+\int q(\mathbf{z}\mid\boldsymbol{\lambda})\nabla_{\boldsymbol{\lambda}}\left[\mathbf{z}\right]\nabla_{\mathbf{z}}[\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda})]\,\mathrm{d}\mathbf{z}\\
&=|\nabla_{\boldsymbol{\lambda}}Q(\mathbf{z}\mid\boldsymbol{\lambda})(\log p(\mathbf{x},\mathbf{z})-\log q(\mathbf{z}\mid\boldsymbol{\lambda}))|_{supp(\mathbf{z})}\\
&\quad+\nabla^{rep}_{\boldsymbol{\lambda}}\mathcal{L},
\end{aligned}$$

where the second to last equality follows by the derivative of the CDF function (Hoffman & 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 $\mathbf{0}$).

Hyperparameters and Convergence. We study one, two, and three layer deep exponential familys (DEFS) with 100, 30, and 15 units respectively and set prior hyperparameters following Ranganath et al. (2015). For hierarchical variational models (HVMS), we use Nesterov's accelerated gradient with momentum parameter of 0.9, combined with RMSProp with a scaling factor of 10^{-3}, to maximize the lower bound. For the mean-field family, we use the learning rate hyperparameters from the original authors'. The HVMS converge faster on Poisson models relative to Bernoulli models. The one layer Poisson model was the fastest to infer.

Multi-level $q(\lambda;\theta)$ and Optimizing with Discrete Variables in the Variational Prior. As mentioned in the main text Hierarchical variational models with multiple layers can contain both discrete and differentiable latent variables. Higher level differentiable variables follow directly from our derivation above. Discrete variables in the prior pose a difficulty due to high variance, as the learning signal contains the entire model. 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 Equation 8 of the main text to form an efficient gradient estimator.

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; these discrete variables can be added to r conditional on everything else. 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 the binary variables in the factorial mixture

References

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. John Wiley & Sons, 2012.

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

Hoffman, Matthew and Blei, David. Stochastic structured variational inference. In *International Conference on Artifical Intelligence and Statistics*, 2015.

Jordan, M. (ed.). *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

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

Mnih, Andriy and Gregor, Karol. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, 2014.

Ranganath, Rajesh, Gerrish, Sean, and Blei, David. Black box variational inference. In *Artificial Intelligence and Statistics*, 2014.

Ranganath, Rajesh, Tang, Linpeng, Charlin, Laurent, and Blei, David M. Deep exponential families. In *Artificial Intelligence and Statistics*, 2015.

Robbins, Herbert. The empirical bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, pp. 1–20, 1964.

Rückstieß, Thomas, Felder, Martin, and Schmidhuber, Jürgen. State-dependent exploration for policy gradient methods. In *Machine Learning and Knowledge Discovery in Databases*, pp. 234–249. Springer, 2008.

Salimans, Tim, Kingma, Diederik, and Welling, Max. Markov chain Monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, 2015.

Sehnke, Frank, Osendorfer, Christian, Rückstieß, Thomas, Graves, Alex, Peters, Jan, and Schmidhuber, Jürgen. Policy gradients with parameter-based exploration for control. In *Artificial Neural Networks-ICANN 2008*, pp. 387–396. Springer, 2008.

Titsias, Michalis K. Local expectation gradients for doubly stochastic variational inference. In *Neural Information Processing Systems*, 2015.