Hierarchical Importance Weighted Autoencoders

Chin-Wei Huang [1 2] **Kris Sankaran** [1] **Eeshan Dhekane** [1] **Alexandre Lacoste** [2] **Aaron Courville** [1 3]

Abstract

Importance weighted variational inference (Burda et al., 2015) uses multiple i.i.d. samples to have a tighter variational lower bound. We believe a joint proposal has the potential of reducing the number of redundant samples, and introduce a hierarchical structure to induce correlation. The hope is that the proposals would coordinate to make up for the error made by one another to reduce the variance of the importance estimator. Theoretically, we analyze the condition under which convergence of the estimator variance can be connected to convergence of the lower bound. Empirically, we confirm that maximization of the lower bound does implicitly minimize variance. Further analysis shows that this is a result of negative correlation induced by the proposed hierarchical meta sampling scheme, and performance of inference also improves when the number of samples increases.

1. Introduction

Recent advance in variational inference (Kingma & Welling, 2014; Rezende et al., 2014) makes it efficient to model complex distribution using latent variable model with an intractable marginal likelihood $p(\boldsymbol{x}) = \int_{\boldsymbol{z}} p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$, where $\boldsymbol{z}$ is an unobserved vector distributed by a prior distribution, e.g. $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The use of an inference network, or encoder, allows for amortization of inference by directly conditioning on the data $q(\boldsymbol{z}|\boldsymbol{x})$, to approximate the true posterior $p(\boldsymbol{z}|\boldsymbol{x})$. This is known as the **Variational Autoencoder** (VAE).

Normally, learning is achieved by maximizing a lower bound on the marginal likelihood, since the latter is not tractable in general. Naturally, one would be interested in reducing the gap between the bound and the desired objective. Burda et al. (2015) devises a new family of lower bounds with progressively smaller gap using multiple i.i.d. samples from the variational distribution $q(\boldsymbol{z}|\boldsymbol{x})$, which they call the **Importance Weighted Autoencoder** (IWAE). Cremer et al. (2017); Bachman & Precup (2015) notice that IWAE can be interpretted as using a corrected variational distribution in the normal variational lower bound. The proposal is corrected towards the true posterior by the importance weighting, and approaches the latter with an increasing number of samples.

Intuitively, when only one sample is drawn to estimate the variational lower bound, the loss function highly penalizes the drawn sample, and thus the encoder. The decoder will be adjusted accordingly to maximize the likelihood in a biased manner, as it treats the sample as the real, observed data. In the IWAE setup, the inference model is allowed to make mistakes, as a sample corresponding to a high loss is penalized less owing to the importance weight.

Drawing multiple samples also allows us to represent the distribution at a higher resolution. This motivates us to construct a joint sampler such that the empirical distribution drawn from the joint sampler can better represent the posterior distribution. More specifically, we consider a hierarchical sampler, whose latent variable $\boldsymbol{z}_0$ acts at a meta-level to decide the salient points of the true posterior, which we call the **Hierarchical Importance Weighted Autoencoder** (H-IWAE). Doing so allows for (1) summarizing the distribution using the latent variable (Edwards & Storkey, 2017), and (2) counteracting bias induced by each proposal. To analyze the latter effect, we look at the variance of the Monte Carlo estimate of the lower bound, and show that maximizing the lower bound implicitly reduces the variance. Our main contributions are as follows:

- We propose a hierarchical model to induce dependency among the samples for approximate inference, and derive a hierarchical importance weighted lower bound.
- We analyze the convergence of the variance of the Monte Carlo estimate of the lower bound, and draw a connection to the convergence of the bound itself.
- Empirically, we explore different weighting heuristics, validate the hypothesis that the joint samples tend to be negatively correlated, and perform experiments on a suite of standard density estimation tasks.

[1] Mila, University of Montreal [2] Element AI [3] CIFAR member. Correspondence to: Chin-Wei Huang <chin-wei.huang@umontreal.ca>.

Proceedings of the 36^{th} *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

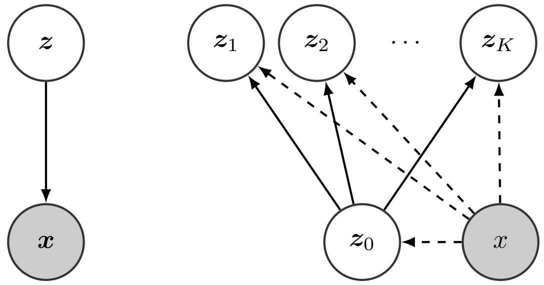


Figure 1. Latent variable model (left) with hierarchical proposals as the inference model (right). Dashed lines indicate amortization.

2. Background

In a typical setup of latent variable model, we assume a joint density function that factorizes as $p(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$, where $\boldsymbol{x},\boldsymbol{z}$ are the observed and unobserved random variables, respectively. Learning is achieved by maximizing the marginal log-likelihood of the data $\log p(\boldsymbol{x}) = \log \int_{\boldsymbol{z}} p(\boldsymbol{x},\boldsymbol{z})d\boldsymbol{z}$, which is in general intractable due to the integration, since a neural network is usually used to parameterize the likelihood function $p(\boldsymbol{x}|\boldsymbol{z})$.

2.1. Variational Inference

One way to estimate the marginal likelihood is via the variational lower bound. Concretely, we introduce a variational distribution $q(\boldsymbol{z})$ [1], and maximize the bound on the RHS:

$$\begin{aligned}\log p(\boldsymbol{x}) &= \log \mathbb{E}_{q(\boldsymbol{z})}\left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] \\ &\geq \mathbb{E}_{q(\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}\right] := \mathcal{L}(q)\end{aligned}$$

which is known as the evidence lower bound (ELBO). The tightness of the ELBO can be described by the reverse Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(q(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x}))$, which is equal to 0 if and only if $q(\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{x})$.

2.2. Importance Weighted Auto-Encoder

Burda et al. (2015) noticed the likelihood ratio in the ELBO, $p(\boldsymbol{x},\boldsymbol{z})/q(\boldsymbol{z})$, resembles the importance weight in importance sampling, and introduced a family of lower bounds called the Importance Weighted Lower Bound (IWLB):

$$\begin{aligned}\log p(\boldsymbol{x}) &= \log \mathbb{E}_{\boldsymbol{z}_j\sim q(\boldsymbol{z}_j)}\left[\frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{z}_j)}{q(\boldsymbol{z}_j)}\right] \\ &\geq \mathbb{E}_{\boldsymbol{z}_j\sim q(\boldsymbol{z}_j)}\left[\log \frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{z}_j)}{q(\boldsymbol{z}_j)}\right] := \mathcal{L}_K(q)\end{aligned}$$

with $\mathcal{L}_K(q) = \mathcal{L}(q)$ when $K = 1$. In practice, the expectation over the product measure is estimated by sampling $\boldsymbol{z}_j \sim q(\boldsymbol{z})$ for $j = 1, ..., K$, and setting

$$\tilde{\mathcal{L}}_K(q) = \log \frac{1}{K}\sum_{j=1}^{K}\frac{p(\boldsymbol{x},\boldsymbol{z}_j)}{q(\boldsymbol{z}_j)}$$

One property of this family of bounds is monotonicity; i.e. $\mathcal{L}_M(q) \geq \mathcal{L}_N(q)$ for $M > N$. Strong consistency of $\tilde{\mathcal{L}}_K(q)$ can also be proved since $\boldsymbol{z}_j$'s are independently and identically distributed (by law of $q(\boldsymbol{z})$), in which case $\tilde{\mathcal{L}}_K(q) \xrightarrow{K\to\infty} \log p(\boldsymbol{x})$ almost surely.

This asymptotic property motivates the use of large number of samples in practice, but there are two practical concerns: First, even though evaluation of $\boldsymbol{z}_j$ under the generative model $p(\boldsymbol{x},\boldsymbol{z})$ can be done in parallel (sublinear rate in general), memory scales in $\mathcal{O}(K)$. An online algorithm can be adopted to reduce memory to $\mathcal{O}(1)$, but at the cost of $\mathcal{O}(n)$ evaluation (Huang & Courville, 2017). Second, for any finite K, for any choice of proposal such that $q(\boldsymbol{z})$ is not proportional to $p(\boldsymbol{x},\boldsymbol{z})$ [2], $\mathcal{L}_K(q)$ is strictly a lower bound on the marginal likelihood. This motivates the research in improving the surrogate loss functions (such as $\mathcal{L}_K(q)$) for the intractable $\log p(\boldsymbol{x})$.

Nowozin (2018), for instance, interpreted $\tilde{\mathcal{L}}_K(q)$ as a biased estimator of $\log p(\boldsymbol{x})$, and introduced a family of estimators with reduced bias. The bias of the new estimator can be shown to be reduced to $\mathcal{O}(K^{-(m+1)})$, for $m < K$ (compared to $\mathcal{O}(K^{-1})$ for $\tilde{\mathcal{L}}_K(q)$). However, the variance of the estimator can be high and is no longer a lower bound.

Alternatively, one can look at the dispersion [3] of the likelihood ratio $w_j = \frac{p(\boldsymbol{x},\boldsymbol{z}_j)}{q(\boldsymbol{z}_j)}$, which itself is a random variable. The gap between $\log p(\boldsymbol{x})$ and $\mathcal{L}_K(q)$ can be explained by Jensen's inequality and the fact that log is a strictly concave function: $\log \mathbb{E}[w] \geq \mathbb{E}[\log w]$ where w is a positive random variable. The equality holds if and only if w is almost surely a constant. This can be approximately true when w is taken to be the likelihood ratio $\frac{p(\boldsymbol{x},\boldsymbol{z})}{q(\boldsymbol{z})}$ and when $q(\boldsymbol{z}) \approx p(\boldsymbol{z}|\boldsymbol{x})$. Instead, if we take $w = \frac{1}{K}\sum_{j=1}^{K} w_j$, we see a direct reduction in variance if w_j's are all uncorrelated and identically distributed: $\mathrm{Var}(w) = \frac{1}{K}\mathrm{Var}(w_1)$. Intuitively, the shape of the distribution of w becomes sharper with larger K, resulting in a smaller gap when Jensen's inequality is applied. This idea was also noticed by (Domke & Sheldon, 2018) and explored by (Klys et al., 2018).

[1] For notational convenience we omit the conditioning on $\boldsymbol{x}$ for amortized inference.

[2] which is a reasonable assumption given the finite approximating capacity of the chosen family of q and the finiteness of the recognition model for amortization.

[3] By dispersion, we mean how "spread out" or "stretched" the distribution of the random variable is. Common statistical dispersion indices include variance, mean absolute difference, entropy, etc.

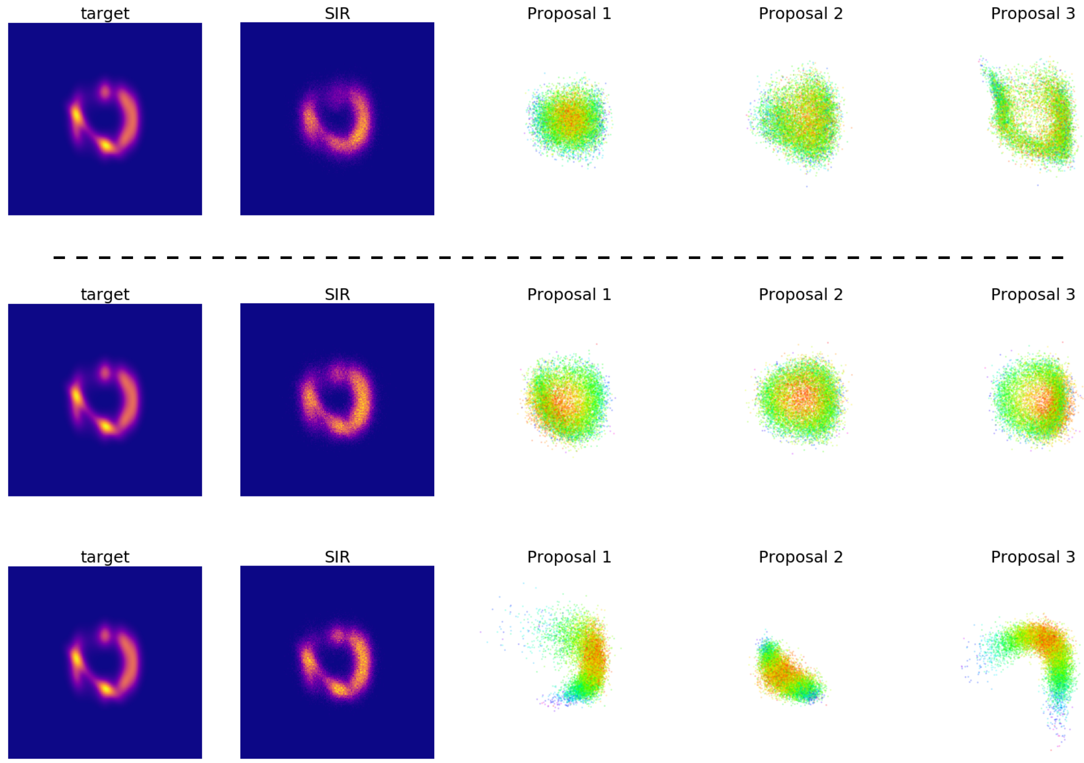


Figure 2. Learned hierarchical importance sampling proposals with uniform weighting (2nd row) and balanced heuristic (3rd row). From left to right are the target distribution, heat map of samples drawn from a *sampling importance resampling* procedure, and the proposals. The color scheme corresponds to the norm of $\boldsymbol{z}_0$, indicating the correlation among the samples. A proposal with Markov transition is presented in the 1st row as a comparison: the dependency among the samples is visually much weaker.

However, a clear connection between how well $\log p(x)$ is approximated and the minimization of the variance of w and/or the variance of $\log w$ is lacking. We defer the discussion to Section 4 wherein we also provide a theoretical analysis.

2.3. Variance Reduction via Negative Correlation

Let $w = \frac{1}{K}\sum_{i=1}^{K}\pi_i w_i$ where w_i's are random variables and π_i's sum to one. The variance of w can be written as:

$$\mathrm{Var}(w) = \sum_{i=1}^{K}\pi_i^2\mathrm{Var}(w_i) + 2\sum_{i<j}\pi_i\pi_j\mathrm{Cov}(w_i, w_j)$$

This suggests that the variance of w is smaller if w_i's are negatively correlated with each other. The intuition of this is as follows: if w_1 deviates from its mean from below, the error it makes can be canceled out by w_2 if the latter is above its mean.

For example, *antithetic variate* (Owen, 2013) is a classic technique that relies on negative correlation. Assume we want to estimate $\mathbb{E}_q[w(\boldsymbol{z})]$, where $q(\boldsymbol{z})$ is a symmetric density. Given $\boldsymbol{z} \sim q(\boldsymbol{z})$, we can augment our estimate with an $\boldsymbol{z}'$ that is opposite to $\boldsymbol{z}$ by reflecting through some center point, and set $\bar{w} = \frac{w(\boldsymbol{z})+w(\boldsymbol{z}')}{2}$. $\bar{w}$ is still an unbiased estimator of $\mathbb{E}_q[w(\boldsymbol{z})]$, but has variance $\mathrm{Var}(\bar{w}) = \frac{\sigma_w^2}{2}(1+\rho)$, where $\sigma_w^2 = \mathrm{Var}(w(\boldsymbol{z}))$ and ρ is the correlation between $w(\boldsymbol{z})$ and $w(\boldsymbol{z}')$. If $w(\cdot)$ is monotonic, then the correlation is negative, so we achieved variance reduction at a rate faster than averaging two samples. Thus, Wu et al. (2018) propose to train an antithetic sampler. But in general, there's no guarantee that the random function $w(\cdot)$ is monotonic, so we propose to train a joint sampler via latent variable.

3. Hierarchical Importance Weighted Auto-Encoder

In order to achieve anticorrelation just described, we consider a joint sampling scheme, with a hierarchical structure that admits fast sampling and allows us to approximate marginal densities required to form a valid lower bound.

3.1. Joint Importance Sampling

We first consider a joint density of K $\boldsymbol{z}_j$'s, denoted by $Q(\boldsymbol{z}_1, ..., \boldsymbol{z}_K)$. Let $q_j(\boldsymbol{z}_j) = \int_{\boldsymbol{z}_{\neg j}} Q(\boldsymbol{z}_1, ..., \boldsymbol{z}_K) d\boldsymbol{z}_{\neg j}$ be the marginal. Let $\pi_j(\boldsymbol{z})$ be a weighting factor such that $\sum_{j=1}^K \pi_j(\boldsymbol{z}) = 1$ for all $\boldsymbol{z}$. Then Jensen's inequality gives

$$\log p(\boldsymbol{x}) = \log \int_{\boldsymbol{z}_1,...,\boldsymbol{z}_K} \sum_{j=1}^K \pi_j(\boldsymbol{z}_j) \frac{p(\boldsymbol{x}, \boldsymbol{z}_j)}{q_j(\boldsymbol{z}_j)} dQ(\boldsymbol{z}_1, ..., \boldsymbol{z}_K)$$

$$\geq \mathbb{E}_Q \left[\log \sum_{j=1}^K \pi_j(\boldsymbol{z}_j) \frac{p(\boldsymbol{x}, \boldsymbol{z}_j)}{q_j(\boldsymbol{z}_j)} \right] := \mathcal{L}_K(Q)$$

which we call the Joint Importance Weighted Lower Bound (J-IWLB) (see Appendix A for a detailed derivation). This allows us to generalize $\mathcal{L}_K(q)$ in two ways:

① The flexibility to have different marginals

② The dependency among $\boldsymbol{z}_j$'s

Point ① with $K > 1$ allows us to relax the necessary condition for optimality when $K = 1$, i.e. $q \propto p$. For example, let $\pi_j(\boldsymbol{z})$ be the "posterior probability" of the random index j, $\pi_j(\boldsymbol{z}) = \frac{q_j(\boldsymbol{z})}{\sum_i q_i(\boldsymbol{z})}$. Then $\mathcal{L}_K(Q)$ is equivalent to using a mixture proposal $\sum_{i=1}^K \frac{q_i}{K}$. In order for the proposals to be optimal, it is sufficient if p can be decomposed as a mixture density (to wit, q_j's do not all need to be equal to p).

Point ② allows us to leverage the correlation among $\boldsymbol{z}_j$'s to further reduce the variance. With $\boldsymbol{z}_1$ making a positive deviation from the mean $\frac{p(\boldsymbol{x}, \boldsymbol{z}_j)}{q_1(\boldsymbol{z}_j)} > p(\boldsymbol{x})$, one would hope $\boldsymbol{z}_2$ has a higher chance of making a negative deviation to cancel the error. This can be thought of as a soft version of antithetic variates.

One difficulty in optimizing $\mathcal{L}_K(Q)$ lies in estimating the marginal density q_j. Particular choices of Q, such as multivariate normal distribution, allows for exact evaluation, since q_j is still normally distributed; this was explored by Klys et al. (2018). Another option is to define $K - 1$ set of invertible transformations, $\mathcal{T}_j(\cdot)$ for $j = 2, ..., K$, and then apply the mapping $\boldsymbol{z}_j \leftarrow \mathcal{T}_j(\boldsymbol{z}_{j-1})$, where $\boldsymbol{z}_1 \sim q_1(\boldsymbol{z})$. The density of z_j under q_j can then be evaluated via change of variable transformation. This is known as the *normalizing flows* (Rezende & Mohamed, 2015). Other more general forms of the joint Q, however, does not have tractable marginal densities, and thus require further approximation.

3.2. Hierarchical Importance Sampling

In order to induce correlation among $\boldsymbol{z}_j$'s, we consider a hierarchical proposal:

$$Q(\boldsymbol{z}_1, ..., \boldsymbol{z}_K) = \int_{\boldsymbol{z}_0} Q(\boldsymbol{z}_1, ..., \boldsymbol{z}_K | \boldsymbol{z}_0) dq_0(\boldsymbol{z}_0)$$

$$= \int_{\boldsymbol{z}_0} \prod_{j=1}^K q_j(\boldsymbol{z}_j | \boldsymbol{z}_0) dq_0(\boldsymbol{z}_0)$$

with the conditional independence assumption $z_i \perp z_j \mid z_0$ for $i \neq j$ and $1 \leq i, j \leq K$ (see Figure 1). First, notice that the marginals are not identical since each $\boldsymbol{z}_j$ is sampled from a different conditional $q_j(\boldsymbol{z}_j|\boldsymbol{z}_0)$. More specifically, each marginal is a latent variable model, which can be thought of as an infinite mixture (Ranganath et al., 2016): $q(\boldsymbol{z}_j) = \int_{\boldsymbol{z}_0} q_j(\boldsymbol{z}_j|\boldsymbol{z}_0) dq_0(\boldsymbol{z}_0)$. Second, $\boldsymbol{z}_1, ..., \boldsymbol{z}_K$ are entangled through sharing the same *common random number* $\boldsymbol{z}_0$. The smaller the conditional entropy $H(\boldsymbol{z}_j|\boldsymbol{z}_0)$ is, the more mutually dependent $\boldsymbol{z}_j$'s are. We analyze the effect of this common random number in Section 6.1.

While there are different ways to model the joint proposal, we emphasize the following properties of the hierarchical proposals that make it an appealing choice:

1. Sampling can be parallelized.

2. Empirically, we found that optimization behaves better than a sequential model (we also tested a Markov joint proposal, i.e. $\boldsymbol{z}_j$ depends only on $\boldsymbol{z}_{j-1}$, and found the learned proposals do not compensate for each other; see top rule of Figure 2).

3. $\boldsymbol{z}_0$ can be interpreted as a summary (Edwards & Storkey, 2017) of the empirical distribution.

Optimization Objective In the spirit of the variational formalism, we need to define an objective to optimize Q. However, the marginal densities are no longer tractable due to the integration over the prior measure q_0. We introduce an auxiliary random variable (Agakov & Barber, 2004) $r(\boldsymbol{z}_0|\boldsymbol{z}_j)$ to approximate $q_j(\boldsymbol{z}_0|\boldsymbol{z}_j)$ [4], and maximize the following objective:

$$\mathcal{L}_K(Q^0) := \mathbb{E}_{Q^0} \left[\log \sum_{j=1}^K \pi_j(\boldsymbol{z}_j, \boldsymbol{z}_0) \frac{p(\boldsymbol{x}, \boldsymbol{z}_j) r(\boldsymbol{z}_0|\boldsymbol{z}_j)}{q_j(\boldsymbol{z}_j|\boldsymbol{z}_0) q_0(\boldsymbol{z}_0)} \right]$$

where Q^0 is the joint of $\boldsymbol{z}_0, ..., \boldsymbol{z}_K$, and $\pi_j(\boldsymbol{z}, \boldsymbol{z}_0)$ is a partition of unity for all $\boldsymbol{z}, \boldsymbol{z}_0$. First, if we choose $\pi_j = \frac{1}{K}$ and let r depend on j, $\mathcal{L}_K(Q^0)$ boils down to the J-IWLB $\mathcal{L}_K(Q)$ if $r_j(\boldsymbol{z}_0|\boldsymbol{z}_j) = q_j(\boldsymbol{z}_0|\boldsymbol{z}_j)$. Second, with $K = 1$, it is simply variational inference with a hierarchical model (Ranganath et al., 2016). Lastly, $\mathcal{L}_K(Q^0)$ is a lower bound on $\log p(\boldsymbol{x})$ (we relegate the proof to Appendix B). We call it the Hierarchical Importance Weighted Lower Bound (H-IWLB).

This training criterion is chosen also because the inference model and generative model can have a unified objective. It is also possible to consider training the inference model with different objectives. For instance, direct minimization of the variance of the importance ratio amounts to minimization

[4] $q_j(\boldsymbol{z}_0|\boldsymbol{z}_j)$ is the posterior of $\boldsymbol{z}_0$ given $\boldsymbol{z}_j$; that is $q_j(\boldsymbol{z}_0|\boldsymbol{z}_j) \propto q_j(\boldsymbol{z}_j|\boldsymbol{z}_0) q_0(\boldsymbol{z}_0)$.

of the χ^2-divergence (see Dieng et al. (2017); Müller et al. (2018)) [5]. We discuss the connection between vanishing variance and convergence of the log estimator in Section 4.

Weighting Heuristics A family of weighting heuristics commonly used in practice are the *power heuristics*:

$$\pi_j^\alpha(\boldsymbol{z}, \boldsymbol{z}_0) = \frac{q_j(\boldsymbol{z}|\boldsymbol{z}_0)^\alpha}{\sum_{i=1}^K q_i(\boldsymbol{z}|\boldsymbol{z}_0)^\alpha}$$

With larger α, the heuristic tends to emphasize the proposal under which $\boldsymbol{z}$ has larger likelihood more. When $\alpha = 0$ and 1, π_j boils down to uniform probability (arithmetic average of the likelihood ratio) and the posterior probability of the index j, respectively. We compare different choices of weighting heuristics qualitatively on a toy example in Section 6.2 and quantitatively on a suite of standard density modeling tasks with latent variable model in Section 6.3.

Training Procedure In our experiments, all q_j and q_0 are conditional Gaussian distributions (also conditioned on $\boldsymbol{x}$ in the amortized inference setting). This allows us to use the reparameterized gradient (Kingma & Welling, 2014; Rezende et al., 2014). However, as noted by Rainforth et al. (2018), the signal-to-noise ratio of the gradient estimator for IWAE's encoder decreases as K increases, large K would render the learning of the inference model inefficient. Thus, we consider the recently proposed doubly reparameterized gradient by Tucker et al. (2018). Empirically, we find the algorithm converges more stably.

4. Connecting Variance and Lower Bound

Ideally, using a joint proposal allows for modelling the dependency among $\boldsymbol{z}_j$'s and thus the negative correlation among the likelihood ratios, but we did not choose to optimize (minimize) variance directly. Instead we maximize the lower bound. We are interested in how the two are connected, i.e. if convergence of one implies another. The following is the main takeaway of this section:

Convergence happening in one mode implies the convergence in the other mode "if values of w_j and $\log w_j$ cannot be extreme" (we will assume these two are bounded in some sense). However, this is in general not true in the case of importance sampling, where the likelihood ratio is sensitive to the choice of proposal distribution. This suggests the density q needs to be controlled in some way (e.g. smoothing the standard deviation of a Gaussian to avoid over-confidence).

We work in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We write $w \in \mathcal{L}^p$ for $p \geq 1$ if w is a random variable and $\mathbb{E}[|w|^p] < \infty$. A random sequence $\{w_n\}$ converges in $\mathcal{L}^p$ to w if $\mathbb{E}[|w_n - w|^p] \to 0$ as $n \to \infty$. Convergence in probability and in $\mathcal{L}^p$ are denoted by $\xrightarrow{P}$ and $\xrightarrow{\mathcal{L}^p}$, respectively.

Let $\{w_n\}$ be a sequence of random variables, such that $\mathbb{E}[w_n] = p(\boldsymbol{x})$ and thus $\mathbb{E}[\log w_n] \leq \log p(\boldsymbol{x})$. w_n can be thought of as the likelihood ratio $p(\boldsymbol{x}, \boldsymbol{z})/q(\boldsymbol{z})$ where $\boldsymbol{z} \sim q(\boldsymbol{z})$, and its expectation wrt $\boldsymbol{z}$ is exactly $p(\boldsymbol{x})$ [6]. We'd like to know if convergence of the lower bound $\mathbb{E}[\log w_n]$ to $\log p(\boldsymbol{x})$ (e.g. via maximization of H-IWLB) implies vanishing variance under any assumption, which justifies the use of a joint/hierarchical proposal to reduce variance at a potentially faster rate. However, it is in general untrue that smaller $\mathrm{Var}(w_n)$, smaller $\mathrm{Var}(\log w_n)$ and larger $\mathbb{E}[\log w_n]$ can imply one another [7]. Instead, we discuss their limiting behavior, and conclude that the condition $\mathrm{Var}(\log w_n)$ converges to zero sits somewhere between $\mathcal{L}^1$-convergence and $\mathcal{L}^2$-convergence of $\log w_n$. To do so, we require a bit more control over the sequences w_n and $\log w_n$, such as boundedness. The first implication of the conclusion is that if the variance of $\log w_n$ vanishes and the sequences are bounded in some sense, $\mathbb{E}[\log w_n]$ also converges to $\log p(\boldsymbol{x})$ (see below). The second implication is that if $\log w_n$ converges to $\log p(\boldsymbol{x})$ "more uniformly" ($\mathcal{L}^2$-convergence), then the variance of $\log w_n$ also converges to zero. We will discuss why boundedness control is required at the end of this section from the f-divergence perspective.

Now we turn to the analysis on the variance of $\log w_n$. Doing so allows us to interpret $\log w_n$ as an estimator for $\log p(\boldsymbol{x})$, and we would like to answer the following question: if the variance of $\log w_n$ converges to zero, does $\mathbb{E}[\log w_n]$ converge to $\log p(\boldsymbol{x})$? The answer is positive given some boundedness condition, and the result suggests that if the variance of the estimator is infinitesimally small, so is the bias. We now state the result:

Proposition 1. *Assume $w_n \in \mathcal{L}^1$ with $w_n > 0$, $\log w_n \in \mathcal{L}^2$ and $\mathbb{E}[w_n] = c$ for some $c > 0$, for all $n \geq 1$. If $\{w_n\}$ is uniformly integrable and $\{\log w_n\}$ is bounded in $\mathcal{L}^1$,*

$$\lim_{n\to\infty} \mathrm{Var}(\log w_n) = 0 \;\Rightarrow\; \log w_n \xrightarrow{P} \log c$$

The proposition tells us that if we look at $\log w_n$ as an estimator for some constant $\log c$ (e.g. $\log p(\boldsymbol{x})$), if $\mathbb{E}[w_n] = c$, then the vanishing variance of $\log w_n$ implies consistency.

With the same integrability condition on $\log w_n$, we can conclude $\mathbb{E}[\log w_n]$ converges to $\log p(\boldsymbol{x})$.

Corollary 1. *With the same condition as Proposition 1, if*

[5] We have also tried to minimize the variance of the estimator (average of importance ratio), with an estimate of the first moment. But we found even the reparameterized gradient suffers from noisy signal and usually converges to a suboptimal solution.

[6] The index n can be thought of as an indicator of a sequence of parameterizations of the joint Q^0.

[7] Klys et al. (2018) also attempt to derive a bound, but their result does not hold without making any assumption. Maddison et al. (2017) also require uniform integrability to establish consistency.

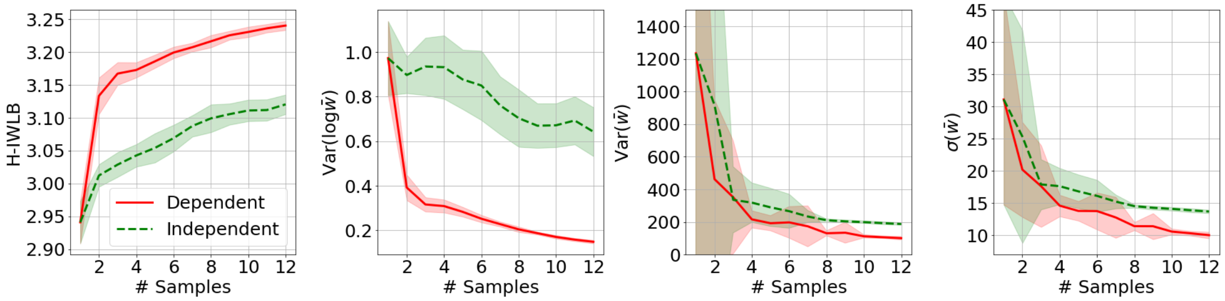
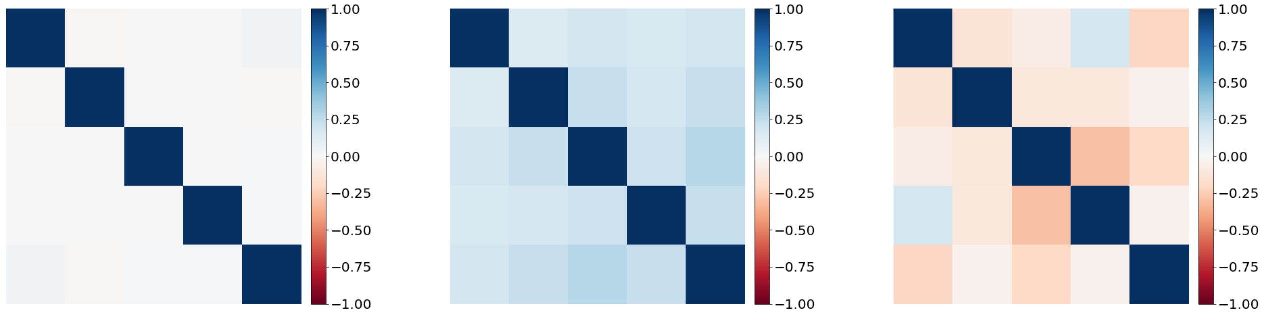


Figure 3. H-IWLB trained with (dependent) and without (independent) common random number. The hierarchical proposal trained with the common factor learns to avoid redundant sampling and effectively reduces dispersion (lower variance and standard deviation).



(a) Uncorrelated w_j

(b) Trained with independent $\boldsymbol{z}_0$

(c) Trained with common $\boldsymbol{z}_0$

Figure 4. Correlation matrix of $\bar{w}_j$ with independent $\boldsymbol{z}_0$ (a) and common $\boldsymbol{z}_0$ (b,c). When trained with common $\boldsymbol{z}_0$ (c), the proposals coordinate to make up for the bias made by one another, resulting in negative correlation.

we further assume $\{\log w_n\}$ *is uniformly integrable,*

$$\lim_{n\to\infty} \mathrm{Var}(\log w_n) = 0 \Rightarrow \log w_n \overset{\mathcal{L}^1}{\to} \log c$$

In particular, $\mathbb{E}[\log w_n] \to \log p(\boldsymbol{x})$ *as* $n \to \infty$*.*

Finally, the following result shows that $\mathcal{L}^2$-convergence is sufficient for the convergence of variance of $\log w_n$ to zero.

Proposition 2. *With the same condition as Corollary 1:*

$$\log w_n \overset{\mathcal{L}^2}{\to} \log c \Rightarrow \lim_{n\to\infty} \mathrm{Var}(\log w_n) = 0$$

Now, we turn back to the case of variational inference where $w = p/q$ (assume p is normalized for the ease of exposition) and discuss the difficulty in analyzing the relationship between variance of w and the lower bound $\mathbb{E}[\log w]$. It is because variance is more sensitive to large likelihood ratio, whereas lower density region under q does not take a heavy toll on the lower bound. More concretely, since $\mathbb{E}_q[w] = 1$, $\mathrm{Var}(w) = \mathbb{E}_q[(p/q)^2] - 1 = D_{\chi^2}(p||q)$, where D_{χ^2} is the χ^2-divergence. Our insight is that the χ^2-divergence is an upper bound on the forward KL divergence $D_{\mathrm{KL}}(p||q)$ (see appendix for a proof). This means when variance of w decreases, the χ^2-divergence and the forward KL-divergence also decrease. However, decrease in the forward KL does not impliy decrease in the reverse KL, $D_{\mathrm{KL}}(q||p)$, which is what maximization of the lower bound $\mathbb{E}_q[\log w]$ is equivalent to. This can be explained by the characteristic function f that is used in the f-divergence family: $D_f(p||q) := \int f\left(\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\right) q(\boldsymbol{z})d\boldsymbol{z}$, where f is a convex function such that $f(1) = 0$. For the forward KL and reverse KL, the corresponding convex functions are $f_F(w) = w \log w$ and $f_R(w) = -\log w$, respectively. When $w \to 0$, $f_F(w) \to 0$ and $f_R(w) \to \infty$. When $w \to \infty$, $f_F(w) \to \infty$ and $f_R(w) \to -\infty$. This means the forward KL will not reflect it when $p \ll q$, but will be high when $p \gg q$. The reverse KL on the other hand will explode when $p \gg q$ and can tolerate $p \ll q$. These are known as the inclusive and exclusive properties of the two KLs (Minka et al., 2005). χ^2-divergence is defined by the convex function $f_\chi(w) = w^2 - 1$, which asymptotically bounds f_F: $\lim_{w\to\infty} f_\chi(w)/f_F(w) = \infty$. This means it

Table 1. Variational Autoencoders trained on the statically binarized MNIST dataset (Larochelle & Murray, 2011). Each hyperparameter is run 5 times to carry out mean and standard deviation (uncertainties are reported in Appendix E for readability). $\tilde{\mathcal{L}}_*$ stands for the lower bound on log likelihood of the dataset (training, validation and test). NLL$_*$ is the negative log likelihood estimated with 2000 samples.

mode	IWAE	H-IWAE	IWAE	H-IWAE			IWAE	H-IWAE			IWAE	H-IWAE		
α	-	0	-	0	1	3	-	0	1	3	-	0	1	3
K	1	1	2	2	2	2	5	5	5	5	10	10	10	10
$-\tilde{\mathcal{L}}_{tr}$	83.26	**82.92**	82.36	82.19	**82.15**	82.43	81.48	82.25	**81.28**	81.32	**80.85**	82.30	80.89	81.17
$-\tilde{\mathcal{L}}_{va}$	86.57	**85.75**	85.40	85.05	**85.03**	85.18	84.45	85.05	**84.04**	84.12	83.84	85.11	**83.77**	83.95
$-\tilde{\mathcal{L}}_{te}$	86.36	**85.50**	85.16	84.79	**84.76**	84.90	84.25	84.85	**83.79**	83.90	83.62	84.86	**83.56**	83.72
NLL$_{va}$	82.64	**82.24**	82.03	81.93	**81.88**	81.96	81.63	82.19	81.42	**81.39**	81.37	82.42	**81.28**	81.33
NLL$_{te}$	82.37	**81.96**	81.77	81.65	**81.60**	81.66	81.37	81.93	81.16	**81.13**	81.13	82.17	**81.04**	81.08

Table 2. Variational Autoencoders trained on the dynamically binarized OMNIGLOT dataset (Lake et al., 2015). Each hyperparameter is run 5 times to carry out the mean and standard deviation (uncertainties are reported in Appendix E for readability).

mode	IWAE	H-IWAE	IWAE	H-IWAE			IWAE	H-IWAE			IWAE	H-IWAE		
α	-	0	-	0	1	3	-	0	1	3	-	0	1	3
K	1	1	2	2	2	2	5	5	5	5	10	10	10	10
$\tilde{\mathcal{L}}_{tr}$	106.40	**104.57**	**102.66**	103.24	102.88	103.17	102.01	**101.35**	101.37	102.78	**99.71**	99.75	100.81	101.29
$\tilde{\mathcal{L}}_{va}$	109.05	**106.48**	105.85	105.27	**105.18**	105.56	104.48	103.76	**103.72**	104.70	**102.67**	103.21	103.49	103.66
$\tilde{\mathcal{L}}_{te}$	109.90	**107.36**	106.70	106.12	**105.93**	106.34	105.37	104.68	**104.62**	105.56	**103.53**	104.11	104.29	104.45
NLL$_{va}$	102.98	**100.36**	100.14	99.82	**99.68**	99.87	99.43	**98.75**	98.85	99.38	**97.97**	98.48	98.58	98.78
NLL$_{te}$	103.28	**100.79**	100.48	100.16	**100.00**	100.18	99.90	99.23	**99.21**	99.80	**98.48**	98.80	99.04	99.22

will incur an even higher cost than the forward KL if $p \gg q$. See the Figure 6 in the appendix for an illustration. The boundedness assumption on $\log w_n$ and w_n made in this section makes sure it is less likely that w_n will be arbitrarily large or close to zero.

5. Related Work

Much work has been done on improving the variational approximation, e.g. by using a more complex form of q (Rezende & Mohamed, 2015; Kingma et al., 2016; Huang et al., 2018; Ranganath et al., 2016; Maaløe et al., 2016; Miller et al., 2016) in place of the conventional normal distribution. Along side the work of Burda et al. (2015), Mnih & Rezende (2016) and Bornschein & Bengio (2014); Le et al. (2018) also apply importance sampling to learning discrete latent variables and modifying the *wake-sleep algorithm*, respectively. Cremer et al. (2017); Bachman & Precup (2015) interpret IWAE as using a corrected proposal, whereas Nowozin (2018) interprets the IWLB as a biased estimator of the marginal log likelihood and propose to reduce the bias using the *Jackknife method*. However, Rainforth et al. (2018) realize the *signal-to-noise ratio* of the gradient of the inference model vanishes as K increase, as the magnitude of the expected gradient decays faster than variance. Tucker et al. (2018) propose to use a doubly reparameterized gradient to mitigate this problem.

Closest to out work is that of Klys et al. (2018), where they propose to explore multivariate normal distribution as the joint proposal. Domke & Sheldon (2018) integrate defensive sampling into variational inference, and Wu et al. (2018) propose to use a differentiable antithetic sampler. Yin & Zhou (2018); Molchanov et al. (2018); Sobolev & Vetrov (2018) propose to use multiple samples to better estimate the marginal distribution of a hierarchical proposal.

6. Experiments

We first demonstrate the effect of sharing a common random number on the dependency among the multiple proposals, and then apply the amortized version of hierarchical importance sampling (that is, H-IWAE) to learning a deep latent Gaussian models. The details of how the inference model $Q^0(\cdot|\boldsymbol{x})$ is parameterized can be found in Appendix D.

6.1. Effect of Common Random Number

In this section, we analyze the effect of training with common random number, $\boldsymbol{z}_0$. As mentioned in Section 3, the correlation among $\boldsymbol{z}_j$'s is induced by tying $\boldsymbol{z}_0$ as a common factor. If for each index j, we draw $\boldsymbol{z}_0$ independently from q_0, $\boldsymbol{z}_j$ will be rendered independent. Doing so allows us to test if inference models trained with a common random number tend to output correlated $\boldsymbol{z}_j$'s. Let the target in Figure 2 be the posterior distribution $\tilde{p}(\boldsymbol{z})$ (unnormalized), and let $\bar{w}_j = \tilde{p}(\boldsymbol{z}_j) r_j(\boldsymbol{z}_0|\boldsymbol{z}_j)/q_j(\boldsymbol{z}_j|\boldsymbol{z}_0)q_0(\boldsymbol{z}_0)$ and $\bar{w} = \sum \bar{w}_j/K$. We consider maximizing H-IWLB with two settings: with and without a common $\boldsymbol{z}_0$. We repeat the experiment 25 times with different random seeds, record the corresponding H-IWLB, variance of $\log \bar{w}$, variance and standard deviation of $\bar{w}$ with common $\boldsymbol{z}_0$, as plotted in Figure 3.

Qualitatively, we visualize the correlation matrix of $\bar{w}_j$ in Figure 4. This shows that $\bar{w}_j$'s are indeed negatively correlated with each other when trained with common $\boldsymbol{z}_0$. When trained with independent $\boldsymbol{z}_0$ for each $\boldsymbol{z}_j$, the hierarchical sampler tends to find similar solutions and are prone to in-

Table 3. Variational Autoencoders trained on the Caltech101 Silhouettes dataset (Marlin et al., 2010). Each hyperparameter is run 3 times to carry out the statistics (μ:σ). (*) indicates the encoder is updated twice on the same minibatch before the decoder is updated once.

	model	α	K	$\tilde{\mathcal{L}}_{tr}$	$\tilde{\mathcal{L}}_{va}$	$\tilde{\mathcal{L}}_{te}$	NLL$_{va}$	NLL$_{te}$
	IWAE	-	1	123.74:1.56	128.87:1.47	129.71:1.47	117.88:1.50	118.61:1.44
	IWAE	-	2	118.67:2.16	124.78:2.20	125.56:2.12	114.24:2.10	114.79:2.11
	IWAE	-	5	112.50:1.16	120.12:0.43	120.90:0.40	109.87:0.68	110.49:0.64
	IWAE (h)	-	1	113.54:3.06	121.20:1.13	121.81:1.23	109.28:1.38	109.71:1.35
	IWAE (h)	-	2	111.94:2.46	118.59:1.24	119.51:1.13	107.52:1.36	108.18:1.38
	IWAE (h)	-	5	108.20:0.58	115.58:1.08	116.40:1.11	105.16:0.77	105.70:0.78
	H-IWAE	0	2	108.69:1.14	118.35:0.69	119.18:1.01	106.93:0.59	107.51:0.74
	H-IWAE	0	5	108.97:1.15	116.84:0.92	117.55:0.79	106.41:0.88	106.83:0.63
	H-IWAE	1	2	111.57:0.88	118.03:0.74	118.78:0.71	107.01:0.79	107.62:0.71
	H-IWAE	1	5	108.96:0.32	116.27:0.39	117.05:0.15	106.03:0.21	106.62:0.09
	H-IWAE	3	2	111.02:0.72	118.02:0.63	119.05:0.45	107.21:0.46	107.90:0.35
	H-IWAE	3	5	109.66:1.07	116.96:0.66	117.80:0.73	106.41:0.35	107.13:0.41
(*)	H-IWAE	1	2	108.31:1.68	115.31:1.08	116.07:0.94	104.27:0.93	104.88:0.87
(*)	H-IWAE	1	5	108.03:1.18	113.59:0.70	114.07:0.75	103.74:0.59	104.16:0.64

curring positively correlated biases (deviation from mean).

6.2. Effect of Weighting Heuristics

We also analyze the effect of using different weighting heuristics. We repeat the experiment in the last subsection and apply the power heuristic with $\alpha = 0$ and $\alpha = 1$ as the weighting scheme, both with common $\boldsymbol{z}_0$. We find that the weighting function π_j has a major effect on the shape of the learned proposals q_j (see Figure 2). With $\alpha = 1$, the proposals behave more like a mixture, and the negative correlation is stronger as the proposals "specialize" in different regions. We also explore training the weighting function in Appendix E.

6.3. Variational Autoencoder

Our final experiment was to apply hierarchical proposals to learning variational autoencoders on a set of standard datasets, including binarized MNIST (Larochelle & Murray, 2011), binarized OMNIGLOT (Lake et al., 2015) and Caltech101 Silhouettes (Marlin et al., 2010), using the same architecture as described in Huang et al. (2018).

Results on the binarized MNIST and OMNIGLOT are in Table 1 and 2, respectively. We compare with Gaussian IWAE as a baseline. The hyperparameters are fixed as follows: minibatch size 64, learning rate 5×10^{-5}, linear annealing schedule with 50,000 iterations for the log density terms except $p(\boldsymbol{x}|\boldsymbol{z})$ (i.e. KL between $q(\boldsymbol{z}|\boldsymbol{x})$ and $p(\boldsymbol{z})$ for VAE), polyak averaging with exponential averaging coefficient 0.998. For the MNIST dataset, with $\alpha = 0$, arithmetic averaging does not provide monotonic improvement in terms of negative log-likelihood (NLL) when K increases, whereas with $\alpha = 1$ or 2 the performance is better with larger K. For the OMNIGLOT dataset, the performance is consistently better with larger K.

Table 3 summarizes the experiment on the Caltech101 dataset. We compare with the Gausssian IWAE and IWAE with one single hierarchical proposal, dubbed IWAE (h), and perform grid search on a set of hyperparameters (Appendix D), each repeated three times to calculate the mean and standard deviation. We report the performance of the models by selecting the hyperparameters that correspond to the lowest averaged NLL on the validation set. We see that with larger K, H-IWAE has an improved performance, but is outperformed by the IWAE with a single hierarchical proposal. We speculate this is due to the increased difficulty in optimizing the inference model with the extra parameters for each q_j, resulting in a larger amortization bias in inference (Cremer et al., 2018). To validate the hypothesis, we repeat the experiment of H-IWAE with the balanced heuristic $\alpha = 1$, but update the encoder twice on the same minibatch before updating the decoder once. This gives us a significant improvement over the learned model (see the last two rows).

7. Conclusion

In order to approximate the posterior distribution in variational inference with a more representative empirical distribution, we propose to use a hierarchical meta-sampler. We derive a variational lower bound as a training objective. Theoretically, we provide sufficient condition on boundedness to connect the convergence of the variance of the Monte Carlo estimate of the lower bound with the convergence of the bound itself. Empirically, we show that maximizing the lower bound implicitly reduces the variance of the estimate. Our analysis shows that learning dependency among the joint samples can induce negative correlation and improve the performance of inference.

References

Agakov, F. V. and Barber, D. An auxiliary variational method. In *Neural Information Processing*, 2004.

Bachman, P. and Precup, D. Training deep generative models: Variations on a theme. In *NIPS Approximate Inference Workshop*, 2015.

Bornschein, J. and Bengio, Y. Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751*, 2014.

Burda, Y., Grosse, R., and Salakhutdinov, R. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2015.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *International Conference on Learning Representations*, 2016.

Cremer, C., Morris, Q., and Duvenaud, D. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

Cremer, C., Li, X., and Duvenaud, D. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, 2018.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, 2017.

Domke, J. and Sheldon, D. R. Importance weighting and variational inference. In *Advances in Neural Information Processing Systems*, pp. 4475–4484, 2018.

Edwards, H. and Storkey, A. Towards a neural statistician. In *International Conference on Learning Representations*, 2017.

Huang, C.-W. and Courville, A. Sequentialized sampling importance resampling and scalable iwae. *NIPS Bayesian Deep Learning Workshop*, 2017.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. In *International Conference on Machine Learning*, 2018.

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

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, 2016.

Klys, J., Bettencourt, J., and Duvenaud, D. Joint importance sampling for variational inference. 2018.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Larochelle, H. and Murray, I. The neural autoregressive distribution estimator. In *International Conference on Artificial Intelligence and Statistics*, 2011.

Le, T. A., Kosiorek, A. R., Siddharth, N., Teh, Y. W., and Wood, F. Revisiting reweighted wake-sleep. *arXiv preprint arXiv:1805.10469*, 2018.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. Auxiliary deep generative models. In *International Conference on Machine Learning*, 2016.

Maddison, C. J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., and Teh, Y. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pp. 6573–6583, 2017.

Marlin, B., Swersky, K., Chen, B., and Freitas, N. Inductive principles for restricted boltzmann machine learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 509–516, 2010.

Miller, A. C., Foti, N., and Adams, R. P. Variational boosting: Iteratively refining posterior approximations. *arXiv preprint arXiv:1611.06585*, 2016.

Minka, T. et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

Mnih, A. and Rezende, D. J. Variational inference for monte carlo objectives. *arXiv preprint arXiv:1602.06725*, 2016.

Molchanov, D., Kharitonov, V., Sobolev, A., and Vetrov, D. Doubly semi-implicit variational inference. *arXiv preprint arXiv:1810.02789*, 2018.

Müller, T., McWilliams, B., Rousselle, F., Gross, M., and Novák, J. Neural importance sampling. *arXiv preprint arXiv:1808.03856*, 2018.

Nowozin, S. Debiasing evidence approximations: On importance-weighted autoencoders and jackknife variational inference. In *International Conference on Learning Representations*, 2018.

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.

Rainforth, T., Kosiorek, A. R., Le, T. A., Maddison, C. J., Igl, M., Wood, F., and Teh, Y. W. Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, 2018.

Ranganath, R., Tran, D., and Blei, D. Hierarchical variational models. In *International Conference on Machine Learning*, 2016.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

Sobolev, A. and Vetrov, D. Importance weighted hierarchical variational inference. 2018.

Tucker, G., Lawson, D., Gu, S., and Maddison, C. J. Doubly reparameterized gradient estimators for monte carlo objectives. *arXiv preprint arXiv:1810.04152*, 2018.

Wu, M., Goodman, N., and Ermon, S. Differentiable antithetic sampling for variance reduction in stochastic variational inference. *arXiv preprint arXiv:1810.02555*, 2018.

Yin, M. and Zhou, M. Semi-implicit variational inference. *arXiv preprint arXiv:1805.11183*, 2018.

A. Detailed J-IWLB Derivation

We provide an alternative to the derivation in Section 3.1. Note that if we can show that

$$\mathbb{E}_{\mathcal{Q}}\left[\sum_{j=1}^{K}\pi_j\left(\boldsymbol{z}_j\right)\frac{p\left(\boldsymbol{x},\boldsymbol{z}_j\right)}{q_j\left(\boldsymbol{z}_j\right)}\right]$$

is equal to $p\left(\boldsymbol{x}\right)$, then an application of Jensen's inequality to the log of this quantity gives the desired bound.

First, since linearity holds for any collection of variables, whether or not they are independent, the summation can be taken outside, and since the resulting expectations involve only one $\boldsymbol{z}_j$ at a time, the expectations can be taken with respect to their marginals q_j. That is,

$$\mathbb{E}_{\mathcal{Q}}\left[\sum_{j=1}^{K}\pi_j\left(\boldsymbol{z}_j\right)\frac{p\left(\boldsymbol{x},\boldsymbol{z}_j\right)}{q_j\left(\boldsymbol{z}_j\right)}\right]=\sum_{j=1}^{K}\mathbb{E}_{q_j}\left[\pi_j\left(\boldsymbol{z}_j\right)\frac{p\left(\boldsymbol{x},\boldsymbol{z}_j\right)}{q_j\left(\boldsymbol{z}_j\right)}\right].$$

Each of these K expectations has a simple form,

$$\mathbb{E}_{q_j}\left[\pi_j\left(\boldsymbol{z}_j\right)\frac{p\left(\boldsymbol{x},\boldsymbol{z}_j\right)}{q_j\left(\boldsymbol{z}_j\right)}\right]=\int\pi_j\left(\boldsymbol{z}\right)p\left(\boldsymbol{x},\boldsymbol{z}\right)d\boldsymbol{z},$$

where it's okay to drop the index j previously appended to the z's, because each integral refers to all values each z_j could possibly take on (not any individual sample of their values, see Figure 5). Since $\sum_j\pi_j\left(\boldsymbol{z}\right)=1$ pointwise for all z, we can then swap summation and integration (assuming dominated-convergence-style regularity) and find

$$\begin{aligned}\sum_{j=1}^{K}\int\pi_j\left(\boldsymbol{z}\right)p\left(\boldsymbol{x},\boldsymbol{z}\right)d\boldsymbol{z}&=\int\sum_{j=1}^{K}\pi_j\left(\boldsymbol{z}\right)p\left(\boldsymbol{x},\boldsymbol{z}\right)d\boldsymbol{z}\\&=\int p\left(\boldsymbol{x},\boldsymbol{z}\right)d\boldsymbol{z}\\&=p\left(\boldsymbol{x}\right).\end{aligned}$$

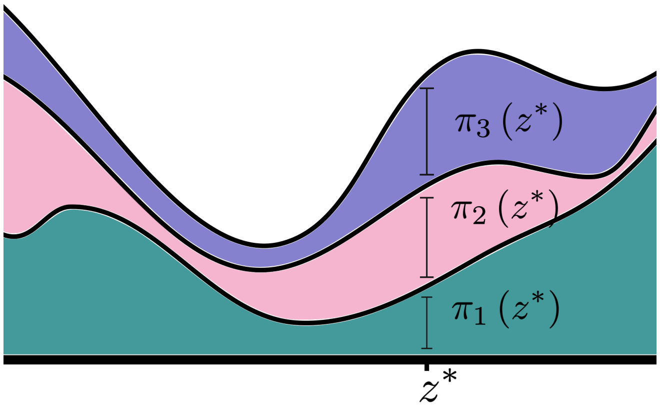


Figure 5. The integral $\int p\left(x,z\right)dz$ can be split into K terms, where at each z the different terms have a proportion $\pi_j\left(z\right)$ of the original function value.

B. Detailed H-IWLB Derivation

This derivation follows the derivation from the last section, and differs in the last step where we also marginalize out the auxiliary variable $\boldsymbol{z}_0$.

$$\begin{aligned}
\mathcal{L}_K(Q^0) &\leq \log \mathbb{E}_{Q^0}\left[\sum_{j=1}^{K} \pi_j(\boldsymbol{z}_j, \boldsymbol{z}_0)\frac{p(\boldsymbol{x}, \boldsymbol{z}_j)r(\boldsymbol{z}_0|\boldsymbol{z}_j)}{q_j(\boldsymbol{z}_j|\boldsymbol{z}_0)q_0(\boldsymbol{z}_0)}\right] && \text{(Jensen's Inequality)}\\
&= \log \sum_{j=1}^{K} \mathbb{E}_{Q^0}\left[\pi_j(\boldsymbol{z}_j, \boldsymbol{z}_0)\frac{p(\boldsymbol{x}, \boldsymbol{z}_j)r(\boldsymbol{z}_0|\boldsymbol{z}_j)}{q_j(\boldsymbol{z}_j|\boldsymbol{z}_0)q_0(\boldsymbol{z}_0)}\right] && \text{(Linearity of expectation)}\\
&= \log \sum_{j=1}^{K} \int_{\boldsymbol{z}_0, \boldsymbol{z}_j} \pi_j(\boldsymbol{z}_j, \boldsymbol{z}_0)q_j(\boldsymbol{z}_j|\boldsymbol{z}_0)q_0(\boldsymbol{z}_0)\frac{p(\boldsymbol{x}, \boldsymbol{z}_j)r(\boldsymbol{z}_0|\boldsymbol{z}_j)}{q_j(\boldsymbol{z}_j|\boldsymbol{z}_0)q_0(\boldsymbol{z}_0)} d\boldsymbol{z}_0 d\boldsymbol{z}_j && \text{(Marginalization of } \boldsymbol{z}_{\neg(j\wedge 0)}\text{)}\\
&= \log \int_{\boldsymbol{z}_0, \boldsymbol{z}} \left(\sum_{j=1}^{K} \pi_j(\boldsymbol{z}, \boldsymbol{z}_0)\right) p(\boldsymbol{x}, \boldsymbol{z})r(\boldsymbol{z}_0|\boldsymbol{z}) d\boldsymbol{z}_0 d\boldsymbol{z} && \text{(Identity)}\\
&= \log p(\boldsymbol{x}) && \text{(Marginalization of } \pi_j, \boldsymbol{z}_0 \text{ and } \boldsymbol{z}\text{)}
\end{aligned}$$

C. Proofs of Section 4

Setup. We work in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We write $X \in \mathcal{L}^p$ for $p \geq 1$ if X is a random variable and $\mathbb{E}[|X|^p] < \infty$. Convergence in probability and in $\mathcal{L}^p$ are denoted by $\overset{P}{\to}$ and $\overset{\mathcal{L}^p}{\to}$, respectively.

We start with a classic result. Assume $\mathrm{Var}(w_n) \to 0$ as $n \to \infty$, then by Chebyshev's inequality, $w_n \to p(\boldsymbol{x})$ in probability, and $\log w_n \to \log p(\boldsymbol{x})$ in probability by the continuity of $\log$. To get $\mathcal{L}^1$-convergence, i.e. $\mathbb{E}[|\log w_n - \log p(\boldsymbol{x})|] \to 0$, the missing piece that is both necessary and sufficient is the uniform integrability of $\log w_n$, which is a form of boundedness condition (see also Proposition 1 of Maddison et al. (2017)). It is clear from now that to say something about the expected value we need to bound the random variable in some way. With the $\mathcal{L}^1$-convergence, we conclude the expected lower bound (in our case, some form of ELBO) converges to the marginal log-likelihood:

$$|\mathbb{E}[\log w_n] - \log p(\boldsymbol{x})| \leq \mathbb{E}[|\log w_n - \log p(\boldsymbol{x})|] \overset{n\to\infty}{\longrightarrow} 0$$

Definition 1. *A family of random variables $\{X_n\}$ is bounded in probability if for any ϵ, there exists $M \geq 0$ such that*

$$\sup_{n\geq 1} \mathbb{P}(|X_n| > M) < \epsilon$$

Note that if $\{X_n\}$ is bounded in $\mathcal{L}^1$ (i.e. $\sup_{n\geq 1} \mathbb{E}[|X_n|] < \infty$), $\{X_n\}$ is also bounded in probability, since by Markov's inequality,

$$\mathbb{P}(|X_n| > M) < \frac{\mathbb{E}[|X_n|]}{M} \leq \frac{\sup_n \mathbb{E}[|X_n|]}{M}$$

Setting $M = \frac{\sup_n \mathbb{E}[|X_n|]}{\epsilon}$ gives us the uniform bound.

Below, we extend the well known Continuous Mapping Theorem to the difference of random sequences $X_n - Y_n$.

Lemma 1. ***(Extended Continuous Mapping Theorem)*** *Let $\{X_n\}$ and $\{Y_n\}$ be random variables bounded in probability. If f is a continuous function, then*

$$X_n - Y_n \overset{P}{\to} 0 \Rightarrow f(X_n) - f(Y_n) \overset{P}{\to} 0$$

Proof. Fix $\epsilon > 0$. Choose $r > 0$. There exists some positive value $T_{r,\epsilon}$ such that $\mathbb{P}(|X_n| > T_{r,\epsilon}) < \frac{r}{2}\epsilon$ and $\mathbb{P}(|Y_n| > T_{r,\epsilon}) < \frac{r}{2}\epsilon$. Within the interval $[-T_{r,\epsilon}, T_{r,\epsilon}]$, g is uniformly continuous, so there exists $\delta_{\epsilon,T_{r,\epsilon}} > 0$ such that

$$|x - y| \leq \delta_{\epsilon,T_{r,\epsilon}} \Rightarrow |g(x) - g(y)| \leq \epsilon$$

Now by subadditivity of measure,

$$\begin{aligned}\mathbb{P}(|g(X)-g(Y)|>\epsilon) &\leq \mathbb{P}(|X_n|>T_{r,\epsilon})+\mathbb{P}(|Y_n|>T_{r,\epsilon}) \\ &\quad+\mathbb{P}(\{|g(X_n)-g(Y_n)|>\epsilon\}\cap\{|X_n|\leq T_{r,\epsilon}\}\cap\{|Y_n|\leq T_{r,\epsilon}\}) \\ &\leq r\epsilon+\mathbb{P}(|X_n-Y_n|>\delta_{\epsilon,T_{r,\epsilon}})\end{aligned}$$

The second term goes to 0 as $n\to\infty$. Taking r to 0 yields the result. □

Note that the continuity assumption of f can be weakened to assuming the set of discontinuity points of f has measure zero.

Now we restate Proposition 1 below.

Proposition 1. *Assume $w_n\in\mathcal{L}^1$ with $w_n>0$, $\log w_n\in\mathcal{L}^2$ and $\mathbb{E}[w_n]=c$ for some $c>0$, for all $n\geq 1$. If $\{w_n\}$ is uniformly integrable and $\{\log w_n\}$ is bounded in $\mathcal{L}^1$,*

$$\lim_{n\to\infty}\text{Var}(\log w_n)=0\ \Rightarrow\ \log w_n\xrightarrow{P}\log c$$

Proof. Let $C_n=\mathbb{E}[\log w_n]$ and $C=\log c$. By Chebyshev's inequality, for any $\epsilon>0$,

$$\lim_{n\to\infty}\mathbb{P}(|\log w_n-C_n|>\epsilon)=0$$

which means $\log w_n-C_n\xrightarrow{P}0$. Due to $\mathcal{L}^1$-boundedness, $\log w_n$ is also bounded in probability, and $|C_n|\leq\mathbb{E}[|\log w_n|]\leq\sup_n\mathbb{E}[|\log w_n|]<\infty$. By the *Extended Continuous Mapping Theorem*, $w_n-\exp(C_n)\xrightarrow{P}0$. Also, $\exp(C_n)=\exp(\mathbb{E}[\log w_n])\leq\exp(|\mathbb{E}[\log w_n]|)$ is bounded, implying $w_n-\exp(C_n)$ is uniformly integrable, so $w_n-\exp(C_n)\xrightarrow{\mathcal{L}^1}0$, and

$$\lim_{n\to\infty}|\mathbb{E}[w_n]-\exp(C_n)|\leq\lim_{n\to\infty}\mathbb{E}[|w_n-\exp(C_n)|]=0$$

Thus, $\lim_{n\to\infty}\exp(C_n)=c$ and $\lim_{n\to\infty}C_n=\log c=C$.

The rest can follow naturally by applying continuous mapping again. As an alternative, we consider an elementary proof. Fix $\epsilon>0$ and let $A_n=\{|\log w_n-C|>\epsilon\}$. Assume $\mathbb{P}(A_n)\nrightarrow 0$. Then there exists $d'\in(0,1]$ such that $\mathbb{P}(A_n)\geq d'$ for infinitely many n. Let n_k be such a subsequence; $\forall\, k\geq 1$,

$$\mathbb{P}(|\log w_{n_k}-C_{n_k}|>\frac{\epsilon}{2})+\mathbb{P}(|C_{n_k}-C|>\frac{\epsilon}{2})\geq\mathbb{P}(|\log w_{n_k}-C|>\epsilon)\geq d'$$

Since the first term on the LHS converges to zero, $\mathbb{P}(|C_{n_k}-C|>\frac{\epsilon}{2})=1$ asymptotically, as both C_{n_k} and C are constant (event $\{|C_{n_k}-C|>\frac{\epsilon}{2}\}$ is either Ω or $\emptyset$). However, $C-C_{n_k}\leq\frac{\epsilon}{2}$ for infinitely many k. Thus, it must be true that $\mathbb{P}(A_n)\to 0$ as $n\to\infty$, and $\log w_n\xrightarrow{P}C$. □

Proof of Proposition 2 is more straightforward. We first restate the statement.

Proposition 2. *With the same condition as Corollary 1:*

$$\log w_n\xrightarrow{\mathcal{L}^2}\log c\ \Rightarrow\ \lim_{n\to\infty}\text{Var}(\log w_n)=0$$

Proof. The variance of $\log w_n$ can be bounded by triangular inequality:

$$\text{Var}(\log w_n)\leq\mathbb{E}[(\log w_n-\log c)^2]+(\log c-\mathbb{E}[\log w_n])^2$$

The first terms goes to zero by convergence in $\mathcal{L}^2$. Note that $|\mathbb{E}[\log w_n]-\log c|\leq\mathbb{E}[|\log w_n-\log c|]\overset{n\to\infty}{\longrightarrow}0$. So the second term also converges to zero. □

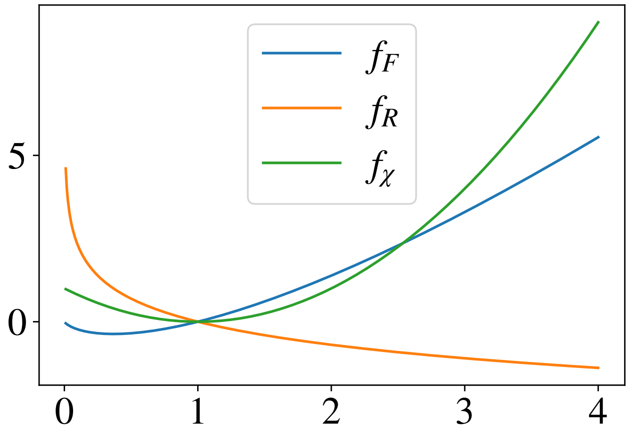


Figure 6. The characteristic convex functions for different f-divergences. f_F and f_χ are larger when the x-axis value is large, whereas f_R penalizes more when the x-axis value is close to zero.

Finally, we show that $D_{\chi^2}(p||q)$ is an upper bound on $D_{\text{KL}}(p||q)$. Generally, let Q and P be probability measures such that $Q \gg P$ (P is absolutely continuous wrt Q). Since for $x > 0$, $\log x \leq x - 1$

$$\begin{aligned} D_{\text{KL}}(P||Q) &= \int_\Omega \log\left(\frac{dP}{dQ}\right) dP \\ &\leq \int_\Omega \left(\frac{dP}{dQ} - 1\right) dP \\ &= \int_\Omega \left(\frac{dP}{dQ}\right)^2 dQ - 1 = D_{\chi^2}(P||Q) \end{aligned}$$

The different characteristic convex functions of the f-divergences are plotted in Figure 6 for reference.

D. Additional Experimental Details

Parameterization of the hierarchical proposals In all of our experiments, the conditionals $q_j(\boldsymbol{z}_j|\boldsymbol{z}_0)$ are all normal densities parameterized by a multilayer perceptron (MLP); i.e. $\mathcal{N}(\boldsymbol{z}_j; \mu(\boldsymbol{z}_0), \sigma^2(\boldsymbol{z}_0))$, where

$$\mu(\boldsymbol{z}_0) = (\boldsymbol{W}_\mu \boldsymbol{h} + \boldsymbol{b}_\mu) + \boldsymbol{W}_s \boldsymbol{z}_0 \qquad \sigma(\boldsymbol{z}_0) = (\boldsymbol{W}_\sigma \boldsymbol{h} + \boldsymbol{b}_\sigma)^+ \qquad \boldsymbol{h} = g(\boldsymbol{W}_h \boldsymbol{z}_0 + \boldsymbol{b}_h)$$

where $(\cdot)^+$ denotes the softplus nonlinearity and $g(\cdot)$ is the ELU activation (Clevert et al., 2016). We share the hidden units $\boldsymbol{h}$ ($\boldsymbol{h} \in \mathbb{R}^{1920}$ for the MNIST and OMNIGLOT experiments) for different $q_j(\boldsymbol{z}_j|\boldsymbol{z}_0)$'s.

The conditional $r(\boldsymbol{z}_0|\boldsymbol{z}_j)$ is also a normal density, with a similar parameterization. But when arithmetic averaging ($\alpha = 0$) is used, we learn the embedding of each j and conditionally scale and shift the weight norm parameters of the hidden units. We also apply the conditional weight norm to condition on the input data $\boldsymbol{x}$ in the amortized inference set-up.

Hyperparameter search of the Caltech101 experiment We perform grid search on the power set of the following hyperparameters for the Caltech101 experiments:

1. learning rate: $[0.0003, 0.0001, 0.00005, 0.0003, 0.00001]$
2. free bits (Kingma et al., 2016): $[0.00, 0.01]$
3. polyak averaging: $[0.95, 0.99]$
4. dimensionality of $\boldsymbol{h}$: $[500, 1024]$

E. Additional Experimental Results

Aside from the power heuristics in Section 3, we also explore the possibility to parameterize the weighting function $\pi_j(z)$, using an MLP with a softmax output. We visualize the learned π_j in Figure 7.

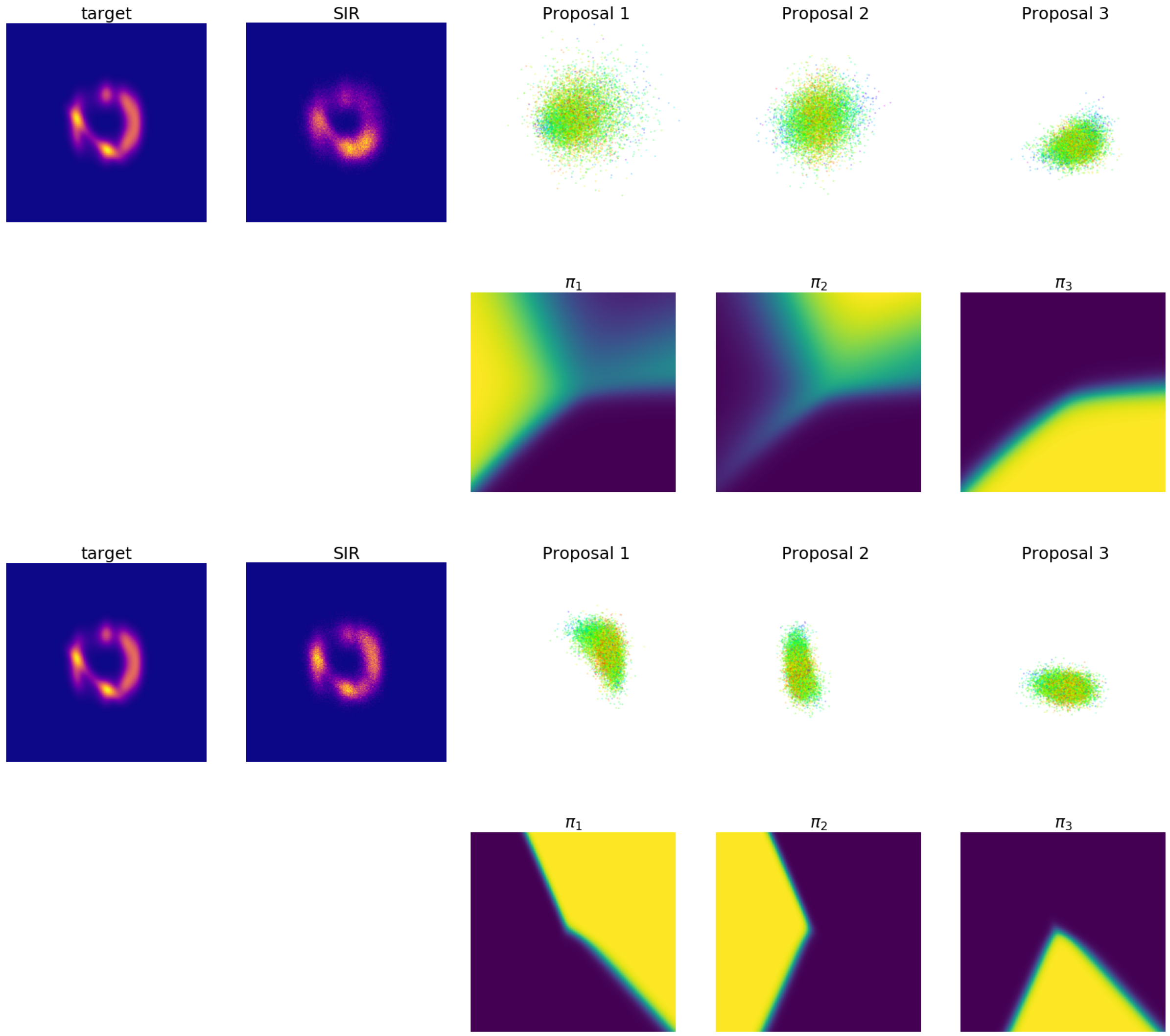


Figure 7. Learned hierarchical importance sampling proposals with a learned weighting function $\pi_j(z)$ parameterized by a one-hidden-layer MLP. The one on top is only trained for 50 iterations with SGD. The one below is trained till convergence (5,000 iterations).

Table 4. K=1 on MNIST.

mode	IWAE	H-IWAE	H-IWAE	H-IWAE
α	-	0	1	3
$\tilde{\mathcal{L}}_{tr}$	83.26 ±0.10	82.92 ±0.17	**82.63** ±0.07	82.67 ±0.15
$\tilde{\mathcal{L}}_{va}$	86.57 ±0.11	85.75 ±0.08	**85.68** ±0.06	85.72 ±0.06
$\tilde{\mathcal{L}}_{te}$	86.36 ±0.15	85.50 ±0.08	**85.43** ±0.05	85.49 ±0.07
NLL$_{va}$	82.64 ±0.11	82.24 ±0.05	**82.16** ±0.03	82.20 ±0.04
NLL$_{te}$	82.37 ±0.12	81.96 ±0.04	**81.89** ±0.04	81.93 ±0.03

Table 5. K=2 on MNIST.

mode	IWAE	H-IWAE	H-IWAE	H-IWAE
α	-	0	1	3
$\tilde{\mathcal{L}}_{tr}$	82.36 ±0.20	82.19 ±0.55	**82.15** ±0.60	82.43 ±0.59
$\tilde{\mathcal{L}}_{va}$	85.40 ±0.05	85.05 ±0.64	**85.03** ±0.65	85.18 ±0.55
$\tilde{\mathcal{L}}_{te}$	85.16 ±0.03	84.79 ±0.63	**84.76** ±0.64	84.90 ±0.56
NLL$_{va}$	82.03 ±0.04	81.93 ±0.42	**81.88** ±0.35	81.96 ±0.27
NLL$_{te}$	81.77 ±0.04	81.65 ±0.42	**81.60** ±0.35	81.66 ±0.27

Table 6. K=5 on MNIST.

mode	IWAE	H-IWAE	H-IWAE	H-IWAE
α	-	0	1	3
$\tilde{\mathcal{L}}_{tr}$	81.48 ±0.17	82.25 ±0.23	**81.28** ±0.14	81.32 ±0.18
$\tilde{\mathcal{L}}_{va}$	84.45 ±0.06	85.05 ±0.39	**84.04** ±0.16	84.12 ±0.16
$\tilde{\mathcal{L}}_{te}$	84.25 ±0.08	84.85 ±0.41	**83.79** ±0.14	83.90 ±0.12
NLL$_{va}$	81.63 ±0.04	82.19 ±0.27	81.42 ±0.07	**81.39** ±0.09
NLL$_{te}$	81.37 ±0.04	81.93 ±0.28	81.16 ±0.05	**81.13** ±0.09

Table 7. K=10 on MNIST.

mode	IWAE	H-IWAE	H-IWAE	H-IWAE
α	-	0	1	3
$\tilde{\mathcal{L}}_{tr}$	**80.85** ±0.16	82.30 ±0.93	80.89 ±0.13	81.17 ±0.24
$\tilde{\mathcal{L}}_{va}$	83.84 ±0.06	85.11 ±0.87	**83.77** ±0.23	83.95 ±0.32
$\tilde{\mathcal{L}}_{te}$	83.62 ±0.03	84.86 ±0.86	**83.56** ±0.22	83.72 ±0.34
NLL$_{va}$	81.37 ±0.05	82.42 ±0.62	**81.28** ±0.08	81.33 ±0.13
NLL$_{te}$	81.13 ±0.02	82.17 ±0.61	**81.04** ±0.09	81.08 ±0.15

Table 8. K=1 on OMNIGLOT.

mode	IWAE	H-IWAE	H-IWAE	H-IWAE
α	-	0	1	3
$\tilde{\mathcal{L}}_{tr}$	106.40 ±1.75	104.57 ±0.98	104.27 ±0.70	**103.77** ±0.49
$\tilde{\mathcal{L}}_{va}$	109.05 ±1.28	106.48 ±0.48	106.26 ±0.25	**106.13** ±0.39
$\tilde{\mathcal{L}}_{te}$	109.90 ±1.21	107.36 ±0.52	107.18 ±0.13	**106.97** ±0.36
NLL$_{va}$	102.98 ±1.56	100.36 ±0.50	100.13 ±0.39	**99.90** ±0.36
NLL$_{te}$	103.28 ±1.53	100.79 ±0.52	100.57 ±0.24	**100.50** ±0.31

Table 9. K=2 on OMNIGLOTt.

mode	IWAE	H-IWAE	H-IWAE	H-IWAE
α	-	0	1	3
$\tilde{\mathcal{L}}_{tr}$	**102.66** ±0.59	103.24 ±1.06	102.88 ±0.89	103.17 ±1.22
$\tilde{\mathcal{L}}_{va}$	105.85 ±0.19	105.27 ±0.53	**105.18** ±0.69	105.56 ±0.88
$\tilde{\mathcal{L}}_{te}$	106.70 ±0.12	106.12 ±0.39	**105.93** ±0.73	106.34 ±0.93
NLL$_{va}$	100.14 ±0.24	99.82 ±0.60	**99.68** ±0.40	99.87 ±0.63
NLL$_{te}$	100.48 ±0.34	100.16 ±0.50	**100.00** ±0.49	100.18 ±0.66

Table 10. K=5 on OMNIGLOT.

mode	IWAE	H-IWAE	H-IWAE	H-IWAE
α	-	0	1	3
$\tilde{\mathcal{L}}_{tr}$	102.01 ±1.05	**101.35** ±1.48	101.37 ±1.72	102.78 ±0.87
$\tilde{\mathcal{L}}_{va}$	104.48 ±0.58	103.76 ±0.73	**103.72** ±0.94	104.70 ±0.46
$\tilde{\mathcal{L}}_{te}$	105.37 ±0.52	104.68 ±0.67	**104.62** ±0.80	105.56 ±0.49
NLL$_{va}$	99.43 ±0.64	**98.75** ±0.79	98.85 ±0.92	99.38 ±0.36
NLL$_{te}$	99.90 ±0.67	99.23 ±0.76	**99.21** ±0.95	99.80 ±0.35

Table 11. K=10 on OMNIGLOT.

mode	IWAE	H-IWAE	H-IWAE	H-IWAE
α	-	0	1	3
$\tilde{\mathcal{L}}_{tr}$	**99.71** ±0.95	99.75 ±1.31	100.81 ±1.24	101.29 ±0.51
$\tilde{\mathcal{L}}_{va}$	**102.67** ±0.39	103.21 ±0.77	103.49 ±0.63	103.66 ±0.18
$\tilde{\mathcal{L}}_{te}$	**103.53** ±0.23	104.11 ±0.85	104.29 ±0.65	104.45 ±0.17
NLL$_{va}$	**97.97** ±0.47	98.48 ±0.67	98.58 ±0.50	98.78 ±0.30
NLL$_{te}$	**98.48** ±0.34	98.80 ±0.64	99.04 ±0.55	99.22 ±0.25