Multiple Importance Sampling ELBO and Deep Ensembles of Variational Approximations

Oskar Kviman[1,2] **Harald Melin**[1,2] **Hazal Koptagel**[1,2] **Víctor Elvira**[3] **Jens Lagergren**[1,2]
[1]KTH Royal Institute of Technology [2]Science for Life Laboratory [3]University of Edinburgh

Abstract

In variational inference (VI), the marginal log-likelihood is estimated using the standard evidence lower bound (ELBO), or improved versions as the importance weighted ELBO (IWELBO). We propose the multiple importance sampling ELBO (MISELBO), a *versatile* yet *simple* framework. MISELBO is applicable in both amortized and classical VI, and it uses ensembles, e.g., deep ensembles, of independently inferred variational approximations. As far as we are aware, the concept of deep ensembles in amortized VI has not previously been established. We prove that MISELBO provides a tighter bound than the average of standard ELBOs, and demonstrate empirically that it gives tighter bounds than the average of IWELBOs. MISELBO is evaluated in density-estimation experiments that include MNIST and several real-data phylogenetic tree inference problems. First, on the MNIST dataset, MISELBO boosts the density-estimation performances of a state-of-the-art model, nouveau VAE. Second, in the phylogenetic tree inference setting, our framework enhances a state-of-the-art VI algorithm that uses normalizing flows. On top of the technical benefits of MISELBO, it allows to unveil connections between VI and recent advances in the importance sampling literature, paving the way for further methodological advances. We provide our code at `https://github.com/Lagergren-Lab/MISELBO`.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

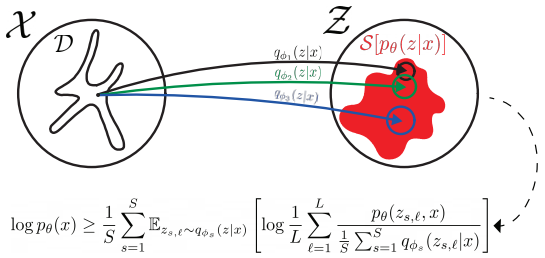


Figure 1: Our proposed framework. First, a set of variational approximations are independently obtained in the latent space above. Second, the marginal log-likelihood is estimated using MISELBO and the ensemble of variational approximations.

1 Introduction

Variational inference (VI; Jordan et al. (1999); Blei et al. (2017)) is an optimization-based approach to probability density estimation. An intractable posterior distribution, $p_\theta(z|x)$, is approximated by a variational approximation, $q_\phi(z|x)$, via maximizing of an objective function, typically the standard ELBO,

$$\mathcal{L} = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right], \tag{1}$$

where ϕ and θ are the variational and generative model parameters, respectively. In classical VI, ϕ is uniquely inferred for every data point x in the dataset, $\mathcal{D}$, rendering it computationally expensive (Zhang et al., 2019). In contrast, amortized VI is based on learning a *mapping*, $f_\phi(x)$, from the data $\mathcal{D}$ to the parameters of the approximate posterior, which is applied across all data points. This makes amortized VI efficient and preferable for large-scale problems. Typically, $f_\phi(x)$ is a neural network (NN), and its weights, ϕ, are learned via stochastic gradient descent (SGD) updates.

The variational auto-encoder (VAE; Kingma and Welling (2013); Rezende et al. (2014)) is an important class of amortized VI algorithms. During training, the

variational and model parameters, ϕ and θ, are jointly learned via SGD optimization of the objective function.

Recently, there has been a surge of research regarding alternative objective functions (Higgins et al., 2016; Kim and Mnih, 2018; Sinha and Dieng, 2021). Some of these results are based on divergence measures (Li and Turner, 2016; Dieng et al., 2017; Wang et al., 2018; Tran et al., 2021), and while others have proposed tighter lower bounds (Masrani et al., 2019). Especially, Burda et al. (2015) proposed the importance weighted ELBO (IWELBO)

$$\mathcal{L}_L = \mathbb{E}_{z_1,\dots,z_L\sim q_\phi(z|x)}\left[\log \frac{1}{L}\sum_{\ell=1}^{L}\frac{p_\theta(z_\ell,x)}{q_\phi(z_\ell|x)}\right], \quad (2)$$

which has been extensively used as an objective function (Burda et al., 2015; Sønderby et al., 2016; Aitchison, 2019; Lopez et al., 2020) and, importantly, as a metric for estimating the marginal log-likelihood, $\log p_\theta(x)$, in VAEs (e.g., Tomczak and Welling (2018); Bauer and Mnih (2019); Vahdat and Kautz (2020)) and in VI in general (Domke and Sheldon, 2018; Zhang, 2020).

There are some limiting properties imposed upon $q_\phi(z|x)$: it should be easy to sample from (preferably through reparameterization), and its likelihood must be tractable. These properties often constrain $q_\phi(z|x)$ to a unimodal family of distributions, making it a simplistic approximation of the intractable posterior. While there have been successful attempts to learn multimodal and other expressive variational approximations, e.g., via the introduction of auxiliary variables in Maaløe et al. (2016) or normalizing flows (NF), as in Rezende and Mohamed (2015); Kingma et al. (2016), these approaches can usually not be straightforwardly be applied to existing methods or require practitioners to expertise in certain methodologies.

In this paper, we propose a flexible framework for obtaining an ensemble of independently inferred variational approximations, enabling the practitioner to employ their preferred VI algorithm and still obtain a multimodal variational approximation. We derive the alternative multiple importance sampling ELBO (MISELBO), which uses the ensemble, as

$$\mathcal{L}^L_{\text{MIS}} = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log \frac{1}{L}\sum_{\ell=1}^{L}\frac{p_\theta(z_{s,\ell},x)}{\frac{1}{S}\sum_{s=1}^{S} q_{\phi_s}(z_{s,\ell}|x)}\right], \quad (3)$$

where $z_{s,\ell} \sim q_{\phi_s}(z|x)$. Our framework is visualized in Figure 1.

The MISELBO is motivated by recent advances in the importance sampling (IS) community, and more particularly in the context of multiple IS (MIS), where multiple proposals/approximations are available, as it is the case here. Recently, in Elvira et al. (2019), it was shown that, using the mixture weights in Eq. (3) (also called balance heuristic (Veach and Guibas, 1995) or deterministic mixture (Owen and Zhou, 2000)) always provides better estimators in terms of variance than using the standard weights in Eq. (2). In the VI framework, where the ELBO computes an expectation of the log transformation, the improvement of the mixture weights is translated into a tighter bound than when the standard ones are used. These connections have promising implications for future research.

Our framework can be easily applied in the context of deep learning since, as in importance sampling (IS) (see (Elvira and Martino, 2021)), the set of $q_{\phi_s}(z|x)$, proposal distributions, can be independently inferred/trained despite sharing the same $p_\theta(z,x)$, target distribution. Consequently, we can exploit deep ensemble diversity (Lakshminarayanan et al., 2016; Fort et al., 2019), which is a novel insight, as far we are aware.

In short, our contributions are:

- We prove that MISELBO is tighter than the average of standard ELBOs, $\overline{\mathcal{L}}$, and show through experiments that it is also tighter than the IWELBO (Section 3.1).
- We establish the concept of deep ensembles of variational approximations (Section 3.2).
- We propose a framework which utilizes ensembles of variational approximations in order to improve marginal log-likelihood estimates for existing algorithms, e.g., NVAE (Vahdat and Kautz (2020); Section 3 and 5).
- We show that the multimodal posterior distribution in the NN weight space (Wilson and Izmailov, 2020), known to be induced by deep ensemble diversity (Fort et al., 2019), translates into a multimodal ensemble of variational approximations in the latent space (Section 5.2).

2 Background

Deep ensembles (Lakshminarayanan et al., 2016), are known to boost prediction performance and uncertainty quantification, while being easy to train. The practitioner simply trains S models independently in parallel, with randomly initialized network weights. Indeed, random initialization provides many opportunities. For instance, Fort et al. (2019) found that

it forces the deep NNs to explore different modes in the NN weight space, making deep ensembles *diverse*. Outside the neural network setting, Yao et al. (2018) showed that diverse variational approximations may be learned in the classical mean-field setting. This was achieved via initialization of the approximations in different parts of the parameter space, and optimizing them using stochastic gradient ascent. In this work, we leverage these findings to obtain diverse ensembles of independently trained variational approximations to boost performance.

In information theory, the Jensen-Shannon divergence (JSD) is a non-negative, symmetric divergence measure for a set of probability distributions, $\mathcal{Q}_S = \{q_{\phi_s}(z|x)\}_{s=1}^S$. Assuming the distributions are equally weighted, the JSD is formulated as

$$\mathrm{JSD}(\mathcal{Q}_S) = \mathbb{H}\left[\frac{1}{S}\sum_{s=1}^{S} q_{\phi_s}(z|x)\right] - \frac{1}{S}\sum_{s=1}^{S}\mathbb{H}\left[q_{\phi_s}(z|x)\right], \tag{4}$$

where $\mathbb{H}[\cdot]$ is the entropy function. Furthermore, the JSD is upper- and lower-bounded. In fact $\mathrm{JSD}(\mathcal{Q}_S) \in [0, \log S]$. As we show in Section 5, the JSD is suitable for measuring the diversity of an ensemble of variational approximations. The benefits of obtaining diverse ensembles in Eq. (3) is motivated by recent work in the MIS literature. Namely, Elvira et al. (2019) showed that the improvement of the MIS weights are more effective (versus the standard weights) when each approximation $q_{\phi_s}(z|x)$ is significantly different with respect to the mixture (high JSD), which aims at mimicking the target distribution. Meanwhile the MIS weights do not provide any improvement when all $q_{\phi_s}(z|x)$ are identical (JSD = 0). This motivates using the JSD to measure the effectiveness of the ensemble.

We use the *average of ELBOs*

$$\overline{\mathcal{L}} = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log\frac{p_\theta(z_s,x)}{q_{\phi_s}(z_s|x)}\right], \tag{5}$$

and the *average of IWELBOs*

$$\overline{\mathcal{L}}_L = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log\frac{1}{L}\sum_{\ell=1}^{L}\frac{p_\theta(z_{s,\ell},x)}{q_{\phi_s}(z_{s,\ell}|x)}\right], \tag{6}$$

as proxies for the ELBO and IWELBO, respectively. This observation makes it simpler for us to compare MISELBO with the other two lower bounds. Indeed, in the setting of deep ensembles, we found that for a set of IWELBOs the standard deviation is small, especially for big L (e.g., see Figure 5).

3 Multiple Importance Sampling Evidence Lower Bound

Algorithm 1 Pseudocode for estimating the MISELBO in Eq. (3)

Inputs: $p_\theta(z,x), q_{\phi_1}(z|x), ..., q_{\phi_S}(z|x)$
Initialize $\tilde{\mathcal{L}} = 0$
for $s = 1, ..., S$ **do**
  $\{z_{s,l}\}_{l=1}^L \sim q_{\phi_s}(z|x)$
  $\tilde{\mathcal{L}} \leftarrow \tilde{\mathcal{L}} + \frac{1}{S}\log\frac{1}{L}\sum_{\ell=1}^{L}\frac{p_\theta(z_{s,\ell},x)}{\frac{1}{S}\sum_{j=1}^{S} q_{\phi_j}(z_{s,\ell}|x)}$
end for

Let us, using $\mathcal{S}$ to denote the support of a distribution, define a set of of variational approximations such that

$$\mathcal{Q}_S = \{q_{\phi_s}(z|x) : \mathcal{S}\left[q_{\phi_s}(z|x)\right] \subseteq \mathcal{S}\left[p_\theta(z|x)\right]\}_{s=1}^S. \tag{7}$$

Importantly, note that the constraint in Eq. (7) occurs naturally when obtaining $q_{\phi_s}(z|x)$ via minimization of $\mathrm{KL}(q_{\phi_s}(z|x)\|p_\theta(z|x))$. The constraint means that all approximations must be absolutely continuous with respect to *the same* target distribution, $p_\theta(z|x)$. For example in a VAE setting, the final training iteration for all encoder networks must be obtained using the same decoder network. Throughout the paper, all comparisons between $\overline{\mathcal{L}}_L$ and $\mathcal{L}_{\mathrm{MIS}}^L$ are done using the same set $\mathcal{Q}_S$, unless mentioned otherwise.

When using our framework, the practitioner simulates a set of $S \times L$ samples (as in the case of the average IWELBOs), and then produces the estimator of the marginal log-likelihood using MISELBO in Eq. (3). Although we in this work sample L times from each $q_{\phi_s}(z|x)$, and weight the samples and approximations equally, there are many possible ways of combining sampling and weighting schemes in the MIS literature (Elvira et al., 2019; Sbert and Elvira, 2022). Sophisticated choices of these schemes can reduce the variance of the estimator, make the ensemble focus on high-probability regions, sample more economically, and more. Also, efficient schemes to find the optimal tradeoff between performance and complexity can be explored (e.g., in the lines of Elvira et al. (2015a)). Hence, MISELBO paves the way for numerous interesting research directions by bridging the gap between VI and MIS.

3.1 Tightness of MISELBO

In the following, we prove that MISELBO provides a tighter bound than the average $\overline{\mathcal{L}}(\mathcal{Q}_S)$. First, let

$$\Delta_L = \mathcal{L}_{\mathrm{MIS}}^L(\mathcal{Q}_S) - \overline{\mathcal{L}}_L(\mathcal{Q}_S) \tag{8}$$

denote the difference between MISELBO and the average of IWELBOs, which with $L = 1$ turns into the average of ELBOs and is denoted Δ_1.

Theorem 1. *For a given set $\mathcal{Q}_S$, the following in-*

equality holds

$$\mathcal{L}_{MIS}(\mathcal{Q}_S) \geq \overline{\mathcal{L}}(\mathcal{Q}_S),$$

and, since $\overline{\mathcal{L}}_1 \equiv \overline{\mathcal{L}}$*, the difference,* $\Delta_1 = \mathcal{L}_{MIS}(\mathcal{Q}_S) - \overline{\mathcal{L}}(\mathcal{Q}_S)$*, satisfies both the upper and lower bounds,*

$$\log S \geq \Delta_1 \geq 0.$$

Proof. We evaluate the difference directly

$$\begin{aligned}
\Delta_1 &= \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{q_{\phi_s}(z|x)}\left[\log \frac{p_\theta(x,z)}{\frac{1}{S}\sum_{s=1}^{S} q_{\phi_s}(z|x)}\right] \\
&\quad - \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{q_{\phi_s}(z|x)}\left[\log \frac{p_\theta(x,z)}{q_{\phi_s}(z|x)}\right] \\
&= -\frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{q_{\phi_s}(z|x)}\left[\log \frac{1}{S}\sum_{s=1}^{S} q_{\phi_s}(z|x)\right] \\
&\quad + \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{q_{\phi_s}(z|x)}\left[\log q_{\phi_s}(z|x)\right] \\
&= \mathbb{H}\left[\frac{1}{S}\sum_{s=1}^{S} q_{\phi_s}(z|x)\right] - \frac{1}{S}\sum_{s=1}^{S} \mathbb{H}\left[q_{\phi_s}(z|x)\right] \\
&= \mathrm{JSD}(\mathcal{Q}_S) \geq 0.
\end{aligned}$$

From Section 2, we know that $\mathrm{JSD}(\mathcal{Q}_S) \in [0, \log S]$, and so the proof is complete. □

Corollary 1.1. *The inequality is strict when* $\bigcap_{s=1}^{S} \mathcal{S}\left[q_{\phi_s}(z|x)\right] = \emptyset$*,* $\mathcal{L}_{MIS}(\mathcal{Q}_S) > \overline{\mathcal{L}}(\mathcal{Q}_S)$*.*

Proof. See supplementary material. □

Corollary 1.2. *There is equality when all distributions in* $\mathcal{Q}_S$ *are the same,* $\mathcal{L}_{MIS}(\mathcal{Q}_S) = \overline{\mathcal{L}}(\mathcal{Q}_S)$*.*

Proof. See supplementary material. □

Moreover, as Burda et al. (2015) does not restrict the variational approximation to any certain family of distributions in their results, the following also holds for MISELBO (where the variational approximation is an equally weighted ensemble)

$$\mathcal{L}^L_{\mathrm{MIS}} \geq \mathcal{L}^{L-1}_{\mathrm{MIS}} \geq \mathcal{L}^1_{\mathrm{MIS}}, \tag{9}$$

and

$$\log p_\theta(x) \geq \mathcal{L}^L_{\mathrm{MIS}}. \tag{10}$$

Although we have not proven that Δ_L is strictly non-negative, this is what our experiments, presented in Section 5, consistently indicate.

Algorithm 2 Pseudocode for deep ensembles of variational approximations. Here, $\mathcal{H}(\cdot,\cdot)$ is the practitioners choice of objective function

Inputs: $f_{\phi_1}(x), f_\theta(z), \mathcal{D}$
Initialize $\phi_2, ..., \phi_S$ randomly
for $s = 2, ..., S$ **do**
  Train $f_{\phi_s}(x)$ on $\mathcal{D}$ via $\partial_{\phi_s}\mathcal{H}\left(f_{\phi_s}(x), f_\theta(z)\right)$ in parallel
end for

3.2 Deep Ensembles in Amortized Variational Inference

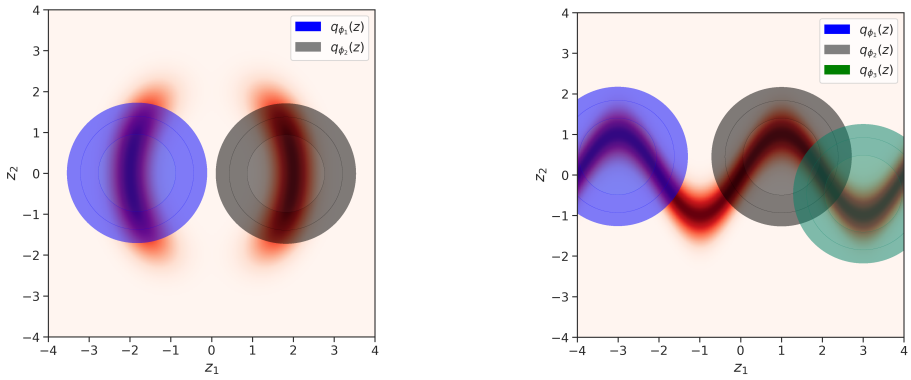
In the previous section, we showed that Δ_1 increases with the $\mathrm{JSD}(\mathcal{Q}_S)$. Although diversity is a vague concept, we hypothesize that the diversity of an ensemble of mappings $\{f_{\phi_s}(x)\}_{s=1}^S$ can be measured in the latent space using $\mathrm{JSD}(\mathcal{Q}_S)$. Consequently, Δ_1 grows as the ensemble of variational approximations becomes more diverse. Reversely, if we obtain a larger $\mathrm{JSD}(\mathcal{Q}_S)$ simply by using a deep ensemble of independently trained variational approximations, then diversity in deep ensembles would promote multimodal posterior distributions, not only over the NN weights Wilson and Izmailov (2020), but also in the latent space. Measuring deep ensemble diversity in the latent space, appears to be a novel idea.

As mentioned in Section 1, VAEs are important instances of amortized VI, and so we apply them to deep ensembles of variational approximations in Section 5.2. In order for deep ensembles of variational approximations to be applicable, however, there needs to be a single mapping from the latent space to the likelihood parameters, $f_\theta(z)$. We solve this by, first, training a VAE, obtaining $\{f_{\phi_1}(x), f_\theta(z)\}$, and then fixing $f_\theta(z)$ — no gradient updates with respect to θ are computed at this point. We now have a generative model, $p_\theta(x,z)$, to use to independently train the $S-1$ other $f_{\phi_s}(x)$, or, equivalently, for inferring $\{q_{\phi_s}(z|x)\}_{s=2}^S$. Algorithm 2 displays the simplicity of the framework.

We refer to a deep ensemble of independently trained mappings $\{f_{\phi_s}(x)\}_{s=1}^S$ as a deep ensemble of variational approximations, $\mathcal{Q}_S$, and their mappings are learned using the same decoder, satisfying $\mathcal{S}[q_{\phi_s}(z|x)] \subseteq \mathcal{S}[p_\theta(z|x)]$ for all $q_{\phi_s}(z|x) \in \mathcal{Q}_S$.

4 Related Work

In recent years, some examples of ensembles of variational approximations and multimodal variational approximations have been proposed. The most important work relating to ours is that of Lopez et al. (2020), as their framework can result in an ensemble of encoder networks with a single decoder. The ensem-



(a) $p_1(z)$ (b) $p_2(z)$

Figure 2: The two true distributions, in (a) and (b) are represented as heat maps. The variational approximations are the contour distributions.

ble is then used for density estimation in interesting decision-making settings. To obtain their variational approximations and the decoder, they follow a three-step procedure which, briefly summarized, involves using different objective functions and selecting which $p_\theta(x, z)$ to use among a set of learned generative models. This is a different approach to ours, and their is not easily applied to the experiments conducted here. Our framework, deep ensembles of variational approximations, is flexible, principled and a generalization of their framework. Finally, they do not show how to use the ensemble to compute the evidence lower bound. We do this using the MISELBO.

In Hernández-Lobato et al. (2016) and Daxberger and Hernández-Lobato (2019) they work with ensembles of variational autoencoders, viewed as a Bayesian deep learning setup by obtaining distributions over the network parameters. They either (i) obtain both $q(\phi)$ *and* $q(\theta)$, or (ii) learn a single encoder (point mass ϕ) and $q(\theta)$. Clearly, (i) is the scheme most similar to our framework, albeit substantially different. They learn pairs of jointly trained encoders and decoders, making $q_{\phi_j}(z|x)$, for any $j \neq s$, not necessarily absolutely continuous with respect to $p_{\theta_s}(z|x)$. This violates the condition from Sec. 3, leaving MISELBO undefined in (i) and (ii).

Recently, Thin et al. (2021) combine annealed importance sampling and sequential Monte Carlo methods to obtain better estimates of the marginal likelihood in the VAE setting. This is indeed an interesting direction which relates to our work by also being on the front line of importance sampling research. Since it is not clear how to apply this method to hierarchical VAEs or outside the VAE setting, and the resulting estimate of the marginal log-likelihood is only proven to empirically be tighter than the IWELBO, we do not compare with it here.

Table 1: The two multimodal distributions used in Section 4.

$\log p_1(z) \propto -\frac{1}{2}\left(\frac{\|z\|-2}{0.4}\right)^2 - \left(e^{-\left(\frac{z_1-2}{1.2}\right)} + e^{-\left(\frac{z_1+2}{1.2}\right)}\right)$
$\log p_2(z) \propto -\frac{1}{2}\frac{(z_2-w(z))^2}{0.4}$, with $w(z) = \sin\left(\frac{2\pi z_1}{4}\right)$

Guo et al. (2016) introduced boosting VI, where ensemble components from a simple parametric base model are iteratively combined to create a more complex posterior. Also, discrete particle variational inference (Saeedi et al., 2017) utilizes an ensemble of variational approximations. However, in both methods the components are jointly trained, and neither of them offer obvious extensions to the deep learning setting.

For deep latent variable models, the prior of the generative model has been the predominant target of multimodal endeavours (Bauer and Mnih, 2019; Bozkurt et al., 2021; Tran et al., 2021). For VAEs, Tomczak and Welling (2018) introduced a mixture distribution prior by leveraging an aggregate posterior over pseudo-inputs, and Jiang et al. (2016) a GMM prior by adding a categorical latent variable. Since these approaches relate to the prior distribution, they are not competing approaches, albeit well-worth mentioning due to their multimodal latent space assumptions.

A GMM variational posterior is provided in the deep latent Gaussian Mixture model of Nalisnick et al. (2016) and in the Varitiational Information Bottleneck of Uğur et al. (2020); both of these approaches rely on jointly optimized components of the GMMs.

In Shi et al. (2019) they propose mixture of experts multimodal VAE (MMVAE), which spans multiple *data modalities*, such as vision and language. Namely,

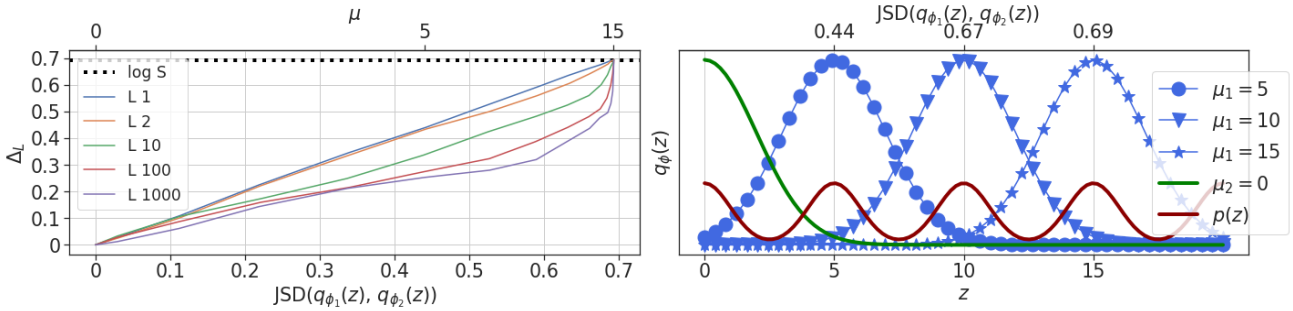


Figure 3: **Left**: Visualization of how Δ_L increases with the JSD($\mathcal{Q}_S$). Note that, regardless of L, Δ_L converges to $\log S$ ($\log 2 \approx 0.69$) as the two Gaussians are separated, or, equivalently, the JSD($\mathcal{Q}_S$) grows. Three examples of μ for $q_{\phi_1}(z)$ are shown in the top labels. **Right**: Examples of how shifting $q_{\phi_1}(z)$ (blue with $\mu_1 = 5, 10, 15$) away from $q_{\phi_2}(z)$ (the green with $\mu_2 = 0$) affects the JSD($\mathcal{Q}_S$). The three corresponding JSD values are shown in the top labels.

Table 2: Unnormalized KL divergences and JSD.

	KL_{MIS}	$\overline{\text{KL}}$	JSD
$p_1(z)$	-0.03	0.61	0.64
$p_2(z)$	0.15	1.05	0.90

the latent space is decomposed into multiple modes, each representing the corresponding data type. Conversely, the multimodality in our framework is w.r.t. the latent space as one component. Finally, MMVAE applies stratified sampling instead of important sampling.

The insights regarding deep ensembles in Lakshminarayanan et al. (2016) are essential for out amortized VI variant. The NNs are independently trained, but concern discriminative networks.

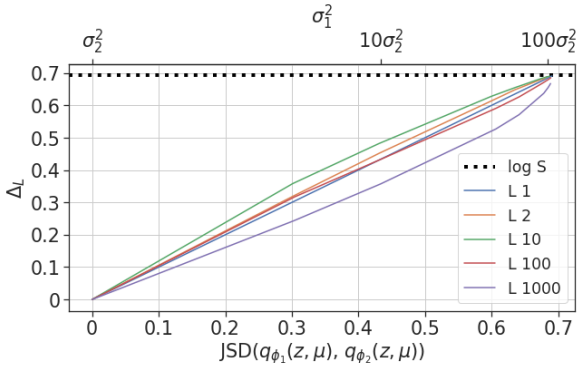


Figure 4: Visualization of how Δ_L increases with the JSD($\mathcal{Q}_S$) in a hierarchical model example (see Section 5.1.2 for details). Here we make the ensemble of variational approximations more diverse by increasing the variance in one of approximations, σ_1^2. Three examples of σ_1^2 for $q_{\phi_2}(\mu)$ are shown in the top labels.

5 Experiments

We consider four density-estimation tasks. The first two experiments concern one- or two-dimensional distributions, in order to easier visualize our framework's strengths. Two large-scale experiments are then considered, displaying how MISELBO enables the use of deep ensembles in amortized VI, as well as its applicability to modern VI techniques such as VAEs and NF.

5.1 Representative Power of the Proposed Framework

To more easily display the power of our proposed method, we here consider two low-dimensional density estimation tasks.

5.1.1 Ensembling Variational Approximations

We use two of the unnormalized and multimodal/periodic distributions described in Rezende and Mohamed (2015) and defined in Table 1. These distributions are known to be too complicated to fit using any standard approximate distribution. This experiment allows us to show (i) that, inspired by Yao et al. (2018), variational approximations with different initializations may ultimately cover different parts of $p(z)$, and (ii) how the resulting ensemble diversity can be leveraged.

The obtained fit is quantified in terms of the KL divergence. Specifically, we compare the following two

	$L=1$	$L=2$	$L=50$	$L=500$	$L=1000$	$\text{JSD}(\mathcal{Q}_S)$
$\mathcal{L}^L_{\text{MIS}}$	$\mathbf{79.10} \pm 0.2$	$\mathbf{79.00} \pm 0.2$	$\mathbf{78.12} \pm 0.2$	$\mathbf{77.81} \pm 0.2$	$\mathbf{77.77} \pm 0.2$	
$\overline{\mathcal{L}_L}$	79.54 ± 0.2	79.43 ± 0.2	78.56 ± 0.1	78.25 ± 0.1	78.21 ± 0.1	
Δ_L	0.42 ± 0.24	0.43 ± 0.25	0.44 ± 0.25	0.44 ± 0.25	0.44 ± 0.25	0.44 ± 0.25
Best $\mathcal{L}_L$	79.86 ± 0.2	79.29 ± 0.2	78.34 ± 0.1	78.20 ± 0.1	78.19 ± 0.1	

Table 3: NLL scores and Δ_L for NVAEs on the MNIST dataset when $S = 2$. The results were averaged using five different random seeds. In this experiments we observed that $\Delta_L \geq \text{JSD}(\mathcal{Q}_S)$ ($\text{JSD}(\mathcal{Q}_S)$ is constant in L). Furthermore, comparing the red entries, we note that the deep ensemble of NVAEs+MISELBO requires 90% less importance samples to outperform NVAE using IWELBOs.

quantities

$$\text{KL}_{\text{MIS}} = \text{KL}\left(\frac{1}{S}\sum_{s=1}^{S} q_{\phi_s}(z) \middle\| p(z)\right), \tag{11}$$

$$\overline{\text{KL}} = \frac{1}{S}\sum_{s=1}^{S} \text{KL}(q_{\phi_s}(z)\|p(z)). \tag{12}$$

We let $q_{\phi_s}(z) = \mathcal{N}(\mu_s, 0.8)$, and minimize the $\text{KL}(q_{\phi_s}(z)\|p(z))$ with respect to the variational parameter μ_s using the Adam optimizer (Kingma and Ba, 2014). The variational approximations are obtained independently, and the parameters are initialized in separate parts of z-space. The resulting scores are presented in Table 2 where we observe high JSDs (close to $\log S$) in both settings. This indicates that diverse ensembles were obtained via different parameter initializations. Also, we see a clear improvement when using the diverse ensembles of variational approximations over averaging the S solutions.

5.1.2 JSD as a Diversity Metric and Visualizing Δ_L

Here we visualize, using an ensemble of two variational approximations, how Δ_L is affected by the $\text{JSD}(\mathcal{Q}_S)$. We do this for two cases. First, we let the variational approximations be unimodal Gaussians with variable means, i.e. $\mathcal{Q}_S = \{q_{\phi_1}(z), q_{\phi_2}(z)\}$, where $q_{\phi_s}(z) = \mathcal{N}(z|\mu_s, 1)$. Second, we consider a hierarchical model by introducing a prior on the mean parameter, and let $\mathcal{Q}_S = \{q_{\phi_1}(z|\mu)q_{\phi_1}(\mu), q_{\phi_2}(z|\mu)q_{\phi_2}(\mu)\}$. In the latter setting, $q_{\phi_s}(\mu) = \mathcal{N}(\mu|10, \sigma_s^2)$ has variable variance.

In both cases, we quantify the diversity of the two approximations using the $\text{JSD}(\mathcal{Q}_S)$. In the first case, we control the diversity by shifting $q_{\phi_1}(z)$ away from $q_{\phi_2}(z)$. In the second case, we gradually increase the variance of $q_{\phi_1}(\mu)$, starting from $\sigma_1^2 = \sigma_2^2$. By doing this, we indirectly parameterize Δ_L by the ensemble diversities.

In Figure 3, we present the results of the first case, where we let the true model, $p(z)$, have six modes. In the supplementary material we include the results of the same experiment with less number of modes for $p(z)$ (the results are similar). In the right plot of Figure 3, it can be observed how shifting the two variational approximations apart increases the $\text{JSD}(\mathcal{Q}_S)$, and hence the ensemble diversity. Recall from Equation (4) that $\text{JSD}(\mathcal{Q}_S)$ is independent of both L and the true model. The results corresponding to the second case are displayed in Figure 4.

Table 4: Negative log-likelihood (NLL) results on the MNIST dataset when $L = 1000$ and $S = 2$. We compare the marginal log-likelihood estimates from NVAE when using our MISELBO and when using the IWELBO.

Model w. lower bound	**NLL**
NVAE w. IWELBO	78.21 ± 0.1
NVAE w. MISELBO	$\mathbf{77.77 \pm 0.2}$

For both cases we confirm that, firstly, Δ_1 follows Theorem 1, i.e. $\Delta_1 = \text{JSD}(\mathcal{Q}_S)$, and, secondly, Corollaries 1.1 and 1.2 hold in practice. Especially, we take note that the performance gain of using MISELBO over the averaged lower bounds increases as the ensemble becomes more diverse, irrespective of L.

Most importantly, in all of the experiments Δ_L is strictly non-negative when $\text{JSD}(\mathcal{Q}_S) > 0$. This inequality is recurring in all of our experiments in this paper, and indicates that the inequality might be true in general. However, it remains to be proven.

5.2 MNIST

We now move to large-scale experiments, starting with the benchmark dataset MNIST (LeCun, 1998). We experiment with deep ensembles of variational approximations using the state-of-the-art Nouveau VAE (NVAE; Vahdat and Kautz (2020)) without NF. In order to obtain $f_{\phi_1}(x)$ and $f_\theta(z)$, we used the authors' exemplarily well-documented code, available on `https://github.com/NVlabs/NVAE`, training with the standard-ELBO-based objective as described in

Table 5: NLL results on six phylogenetic tree inference benchmark datasets for VBPI-NF with RealNVP, $S = 5$, $L = 1000$, $K = 10$. The improvements obtained from using our framework (right column) are substantial in the field of phylogenetics (cf. Table 1 in Zhang and Matsen IV (2018b) or Table 1 in Zhang (2020)).

Dataset	Reference	Taxa	Sites	$\overline{\mathcal{L}}_L$	$\mathcal{L}^L_{\text{MIS}}$
DS1	(Hedges et al., 1990)	27	1949	7108.42	**7108.10**
DS2	(Garey et al., 1996)	29	2520	26367.74	**26367.37**
DS3	(Yang and Yoder, 2003)	36	1812	33735.15	**33734.89**
DS4	(Henk et al., 2003)	41	1137	13329.97	**13329.58**
DS5	(Lakner et al., 2008)	50	378	8214.70	**8214.06**
DS8	(Rossman et al., 2001)	64	1008	8650.73	**8650.32**

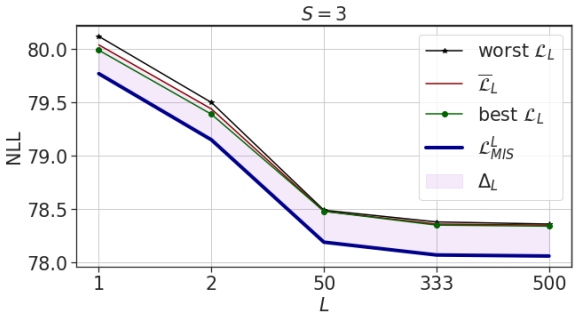


Figure 5: NLL curves and Δ_L for NVAEs and a single random seed on the MNIST dataset when $S = 3$ and varying L. As L increases the IWELBOs converge to their average, while Δ_L is more or less unchanged.

their work, by executing the commands provided there. Unfortunately we were not able to reproduce their results based on their descriptions[4], however it is not critical for us to reproduce their results. Rather, we wish to compare MISELBO with IWELBO and ELBO for a state-of-the-art VAE.

After training $\{f_{\phi_1}(x), f_\theta(z)\}$, we followed the algorithmic description in Algorithm 2 to get the ensemble of deep variational approximations. This means that we froze the NN weights in the decoder, $f_\theta(z)$, while initializing the NN weights in $\{f_{\phi_s}(x) : s \neq 1\}$ randomly. Apart from these minor changes, the original code was not modified for our training, i.e. the architecture and training procedure was identical to what the authors had made available. This illustrates the simplicity and versatility of our framework.

Next, we estimated the marginal log-likelihood using MISELBO (according to Algorithm 1), and reported the results in Table 3, 4 and in Figure 5. Impressively, MISELBO consistently enhances the performance of the NVAE, i.e. Δ_L is strictly non-negative. In this experiment setting, we constrained ourselves to $S \in \{2, 3\}$, as our ablation study, performed on subsets of the data, showed that the JSD($\mathcal{Q}_S$) decreased for $S = 4$ (see supplementary material). From Section 5.1.2 we have inferred that Δ_L might also decrease in this case, while $\overline{\mathcal{L}}_L$ does not appear to gain from increasing S.

In their paper, Vahdat and Kautz (2020) use $L = 1000$ importance samples to estimate the marginal log-likelihood. Indeed, in the MISELBO framework a total of $S \times L$ samples are drawn. However, when using MISELBO, far less importance samples are required in order to outperform NVAE using IWELBO. In Table 3 we demonstrate that NVAE with $L = 50$ and $S = 2$ still outperforms the IWELBO-based benchmarks we were able to obtain for NVAE with $L = 1000$. The improvement in wall-clock time when comparing $\mathcal{L}^{50}_{\text{MIS}}$ ($S = 2$) versus $\mathcal{L}_{1000}$ (note, not the average but a single NVAE) was remarkable. Averaged over four runs, $\mathcal{L}^{50}_{\text{MIS}}$ took 1605±3 seconds, while $\mathcal{L}_{1000}$ needed 8054±3 seconds. Hence, MISELBO used 90% less importance samples, was more than five times faster in computation time, and it was still superior to the IWELBO.

Ultimately, we highlight a novel insight: since non-zero JSD($\mathcal{Q}_S$) is obtained solely from random initialization of the NN weights in $\{f_{\phi_s}(x)\}$, this experiment also demonstrates that we can leverage deep ensemble diversity to translate the multimodal posterior distribution in the NN weight space (Wilson and Izmailov, 2020), into a multimodal ensemble of variational approximations in the latent space.

5.3 Phylogenetic Tree Inference

In contrast to the VAE considered above, the variational Bayesian phylogenetic inference (VBPI; Zhang and Matsen IV (2018b)) framework does not train a generative model. Instead, $p_\theta(x|\lambda, \tau)$ is a likelihood function commonly used in phylogeny, which can be computed using the pruning algorithm presented in Felsenstein (2004). We use the prior distributions over branch lengths λ, and tree topologies, τ, described in Zhang (2020); given the priors, the generative model has no free parameters, i.e., there is no need to infer parameters (see supplementary material for details).

[4]We tried different seeds and more training epochs. In the paper the authors reported an NLL score of 78.01.

We train an $S = 5$ ensemble of variational approximations — the tree topology encoders are subsplit Bayesian networks (SBNs; Zhang and Matsen IV (2018a)) — using $p_\theta(x, \lambda, \tau)$ for six real datasets, and experiment with MISELBO. The results are presented in table 5 from evaluating the NLL scores of VBPI using NF when using MISELBO compared to the average of IWELBOs. The latter lower bound was the reported metric in the original paper, where they used $K = 10$ flows of RealNVP Dinh et al. (2016).

Utilizing the MISELBO framework consistently improves the marginal log-likelihood estimates. Note that $\overline{\mathcal{L}}_L$ and $\mathcal{L}^L_{\text{MIS}}$ use the same number of importance samples throughout the experiments. We stress that in the context of phylogenetics, these improvements are substantial (cf. Table 1 in Zhang and Matsen IV (2018b) or Table 1 in Zhang (2020)). The estimated JSDs in these experiments were all approximately 0.1. Based on our reasoning in this work, this implies that the corresponding $\mathcal{Q}_S$ were not diverse (or at least far from the upper bound on $\text{JSD}(\mathcal{Q}_S)$, $\log 5 \approx 1.61$). This was not necessarily expected as the SBNs are not deep NNs.

VBPI-NF employs a strongly non-Gaussian variational approximation in the tree topology space, $q_\phi(\tau|x)$, while the base-distribution over branch lengths for the NF, $q^0_\phi(\lambda|\tau, x)$, is a log-Normal distribution. The resulting $q^K_\phi(\lambda|\tau, x)$, i.e. after K RealNVP flows, is highly expressive. Therefore, this experiment does not only display that MISELBO improves density-estimation performances even when the variational approximations are non-Gaussian and/or expressive, it also shows that the framework is versatile enough to be applicable to modern, complex VI methods.

6 Conclusion

We have established the concept of deep ensembles of variational approximations and exciting connections between recent advances in IS and VI (MISELBO). These two contributions are the major components of our proposed framework, visualized in Figure 1. We have shown that the framework is versatile, simple to apply for VI methods, and powerful, which we demonstrate by improving the density-estimation performances for two state-of-the-art models, NVAE and VBPI.

Moreover, we have provided a novel proof showing that MISELBO is tighter than the average of ELBOs. In practice, this result appears to generalize to the average of IWELBOs. Also, we are the first to measure the diversity of deep ensembles in the latent space of a latent variable model. This is a useful tool when, for instance, researching the role of the prior distribution in VAEs, a topic currently attracting much attention. The role of the prior distribution on the diversity of the ensemble is furthermore an interesting research path.

Finally, this work paves the way for new research in IS-based deep learning. In particular, we have incorporated recent advances in the IS literature, and we believe that many new connections and developments are ahead. For instance, adaptive IS (AIS) methods (Bugallo et al., 2017), including gradient-based algorithms (Elvira et al., 2015b; Elvira and Chouzenoux, 2019), could be employed or developed ad-hoc for updating/refining the variational approximations and, thereby, also the ensemble weights.

7 Acknowledgements

First, we acknowledge the insightful comments provided by the reviewers which have helped improve our work. This project was made possible through funding from the Swedish Foundation for Strategic Research grant BD15-0043, and from the Swedish Research Council grant 2018-05417_VR. Some of the computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

References

Aitchison, L. (2019). Tensor monte carlo: particle methods for the gpu era. *Advances in Neural Information Processing Systems*, 32.

Bauer, M. and Mnih, A. (2019). Resampled priors for variational autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 66–75. PMLR.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bozkurt, A., Esmaeili, B., Tristan, J.-B., Brooks, D., Dy, J., and Meent, J.-W. (2021). Rate-regularization and generalization in variational autoencoders. In *International Conference on Artificial Intelligence and Statistics*, pages 3880–3888. PMLR.

Bugallo, M. F., Elvira, V., Martino, L., Luengo, D., Míguez, J., and Djuric, P. M. (2017). Adaptive importance sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015).

Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Daxberger, E. and Hernández-Lobato, J. M. (2019). Bayesian variational autoencoders for unsupervised out-of-distribution detection. *arXiv preprint arXiv:1912.05651*.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. M. (2017). Variational inference via x upper bound minimization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 2729–2738.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*.

Domke, J. and Sheldon, D. (2018). Importance weighting and variational inference. *arXiv preprint arXiv:1808.09034*.

Elvira, V. and Chouzenoux, E. (2019). Langevin-based strategy for efficient proposal adaptation in population monte carlo. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5077–5081. IEEE.

Elvira, V. and Martino, L. (2021). Advances in importance sampling. *Wiley StatsRef: Statistics Reference Online*, pages 1–22.

Elvira, V., Martino, L., Luengo, D., and Bugallo, M. F. (2015a). Efficient multiple importance sampling estimators. *IEEE Signal Processing Letters*, 22(10):1757–1761.

Elvira, V., Martino, L., Luengo, D., and Bugallo, M. F. (2019). Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155.

Elvira, V., Martino, L., Luengo, L., and Corander, J. (2015b). A gradient adaptive population importance sampler. In *IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4075–4079.

Felsenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Garey, J. R., Near, T. J., Nonnemacher, M. R., and Nadler, S. A. (1996). Molecular evidence for acanthocephala as a subtaxon of rotifera. *Journal of Molecular Evolution*, 43(3):287–292.

Guo, F., Wang, X., Fan, K., Broderick, T., and Dunson, D. B. (2016). Boosting variational inference. *arXiv preprint arXiv:1611.05559*.

Hedges, S. B., Moberg, K. D., and Maxson, L. R. (1990). Tetrapod phylogeny inferred from 18s and 28s ribosomal rna sequences and a review of the evidence for amniote relationships. *Molecular Biology and Evolution*, 7(6):607–633.

Henk, D. A., Weir, A., and Blackwell, M. (2003). Laboulbeniopsis termitarius, an ectoparasite of termites newly recognized as a member of the laboulbeniomycetes. *Mycologia*, 95(4):561–564.

Hernández-Lobato, D., Bui, T. D., Li, Y., Hernández-Lobato, J. M., and Turner, R. E. (2016). Importance weighted autoencoders with random neural network parameters. In *Workshop on Bayesian Deep Learning, NIPS*, volume 2016.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework.

Jiang, Z., Zheng, Y., Tan, H., Tang, B., and Zhou, H. (2016). Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kim, H. and Mnih, A. (2018). Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29:4743–4751.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lakner, C., Van Der Mark, P., Huelsenbeck, J. P., Larget, B., and Ronquist, F. (2008). Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. *Systematic biology*, 57(1):86–103.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2016). Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Li, Y. and Turner, R. E. (2016). R\'enyi divergence variational inference. *arXiv preprint arXiv:1602.02311*.

Lopez, R., Boyeau, P., Yosef, N., Jordan, M. I., and Regier, J. (2020). Decision-making with auto-encoding variational bayes. *arXiv preprint arXiv:2002.07217*.

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. (2016). Auxiliary deep generative models. In *International conference on machine learning*, pages 1445–1453. PMLR.

Masrani, V., Le, T. A., and Wood, F. (2019). The thermodynamic variational objective. *Advances in Neural Information Processing Systems*, 32.

Nalisnick, E., Hertel, L., and Smyth, P. (2016). Approximate inference for deep latent gaussian mixtures. In *NIPS Workshop on Bayesian Deep Learning*, volume 2, page 131.

Owen, A. and Zhou, Y. (2000). Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2, page 2. Citeseer.

Rossman, A. Y., McKemy, J. M., Pardo-Schultheiss, R. A., and Schroers, H.-J. (2001). Molecular studies of the bionectriaceae using large subunit rdna sequences. *Mycologia*, 93(1):100–110.

Saeedi, A., Kulkarni, T. D., Mansinghka, V. K., and Gershman, S. J. (2017). Variational particle approximations. *The Journal of Machine Learning Research*, 18(1):2328–2356.

Sbert, M. and Elvira, V. (2022). Generalizing the balance heuristic estimator in multiple importance sampling. *Entropy*, 24(2):191.

Shi, Y., Siddharth, N., Paige, B., and Torr, P. H. (2019). Variational mixture-of-experts autoencoders for multi-modal deep generative models. *arXiv preprint arXiv:1911.03393*.

Sinha, S. and Dieng, A. B. (2021). Consistency regularization for variational auto-encoders. *arXiv preprint arXiv:2105.14859*.

Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. (2016). Ladder variational autoencoders. *Advances in neural information processing systems*, 29:3738–3746.

Thin, A., Kotelevskii, N., Doucet, A., Durmus, A., Moulines, E., and Panov, M. (2021). Monte carlo variational auto-encoders. In *International Conference on Machine Learning*, pages 10247–10257. PMLR.

Tomczak, J. and Welling, M. (2018). Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR.

Tran, L., Pantic, M., and Deisenroth, M. P. (2021). Cauchy-schwarz regularized autoencoder. *arXiv preprint arXiv:2101.02149*.

Uğur, Y., Arvanitakis, G., and Zaidi, A. (2020). Variational information bottleneck for unsupervised clustering: Deep gaussian mixture embedding. *Entropy*, 22(2):213.

Vahdat, A. and Kautz, J. (2020). Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*.

Veach, E. and Guibas, L. (1995). Optimally combining sampling techniques for Monte Carlo rendering. *In SIGGRAPH 1995 Proceedings*, pages 419–428.

Wang, D., Liu, H., and Liu, Q. (2018). Variational inference with tail-adaptive f-divergence. *arXiv preprint arXiv:1810.11943*.

Wilson, A. G. and Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*.

Yang, Z. and Yoder, A. D. (2003). Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic biology*, 52(5):705–716.

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, 13(3):917 – 1007.

Zhang, C. (2020). Improved variational bayesian phylogenetic inference with normalizing flows. *arXiv preprint arXiv:2012.00459*.

Zhang, C., Bütepage, J., Kjellström, H., and Mandt, S. (2019). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):2008–2026.

Zhang, C. and Matsen IV, F. A. (2018a). Generalizing tree probability estimation via bayesian networks. *arXiv preprint arXiv:1805.07834*.

Zhang, C. and Matsen IV, F. A. (2018b). Variational bayesian phylogenetic inference. In *International Conference on Learning Representations*.

Supplementary Material: Multiple Importance Sampling ELBO and Deep Ensembles of Variational Approximations

Here we provide supplementary proofs and experimental details. We provide code for the experiments on GitHub: `https://github.com/Lagergren-Lab/MISELBO`.

A Proofs

Here we provide the proofs for Corollaries 1.1 and 1.2. Recall that, for both Corollaries, we consider $L = 1$.

A.1 Proof of Corollary 1.1

A condition for this Corollary to hold is that the supports of the distributions in $\mathcal{Q}_S$ are mutually disjoint. The assumption implies that, when $z \sim q_{\phi_s}(z|x)$

$$\sum_{s'=1}^{S} q_{\phi_{s'}}(z|x) = q_{\phi_s}(z|x). \tag{13}$$

We start by reformulating the JSD as follows

$$\begin{aligned}
\mathrm{JSD}(\mathcal{Q}_S) &= \mathbb{H}\left[\frac{1}{S}\sum_{s=1}^{S} q_{\phi_s}(z|x)\right] - \frac{1}{S}\sum_{s=1}^{S}\mathbb{H}\left[q_{\phi_s}(z|x)\right] && (14)\\
&= -\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log\frac{1}{S}\sum_{s=1}^{S} q_{\phi_s}(z|x)\right] + \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log q_{\phi_s}(z|x)\right] && (15)\\
&= \log S + \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log\frac{q_{\phi_s}(z|x)}{\sum_{s'=1}^{S} q_{\phi_{s'}}(z|x)}\right]. && (16)
\end{aligned}$$

Using the assumption we, get that

$$\frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log\frac{q_{\phi_s}(z|x)}{\sum_{s'=1}^{S} q_{\phi_{s'}}(z|x)}\right] = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log\frac{q_{\phi_s}(z|x)}{q_{\phi_s}(z|x)}\right] = 0, \tag{17}$$

and so we can complete the proof:

$$\begin{aligned}
\Delta_1 = \mathrm{JSD}(\mathcal{Q}_S) &= \log S + \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log\frac{q_{\phi_s}(z|x)}{\sum_{s'=1}^{S} q_{\phi_{s'}}(z|x)}\right] && (18)\\
&= \log S + \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(z|x)}\left[\log\frac{q_{\phi_s}(z|x)}{q_{\phi_s}(z|x)}\right] = \log S > 0, && (19)
\end{aligned}$$

when $S > 1$.

A.2 Proof of Corollary 1.2

For this Corollary, we instead assume that all variational approximations are identical, implying that

$$\sum_{s'=1}^{S} q_{\phi_{s'}}(z|x) = S q_{\phi_s}(z|x), \tag{20}$$

when $z \sim q_{\phi_s}(z|x)$.

Using the reformulation of the JSD in Eq. 14 and Eq. 20, we have

$$\Delta_1 = \text{JSD}(\mathcal{Q}_S) = \log S + \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{q_{\phi_s}(z|x)} \left[\log \frac{q_{\phi_s}(z|x)}{\sum_{s'=1}^{S} q_{\phi_{s'}}(z|x)} \right] \tag{21}$$

$$= \log S + \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{q_{\phi_s}(z|x)} \left[\log \frac{q_{\phi_s}(z|x)}{S q_{\phi_s}(z|x)} \right] \tag{22}$$

$$= \log S + \frac{1}{S}\sum_{s=1}^{S} \mathbb{E}_{q_{\phi_s}(z|x)} \left[\log \frac{q_{\phi_s}(z|x)}{q_{\phi_s}(z|x)} - \log S \right] = \log S - \log S = 0, \tag{23}$$

and so the equality in Corollary 1.2 holds.

B Experiment 5.1.1 Details

For the experiment using $p_1(z)$ as target distribution, we initialized μ_1 to $(-3, 0)$ and μ_2 to $(3, 0)$. For the experiment using $p_2(z)$ as target distribution, we initialized μ_1 to $(-3, 0)$, μ_2 to $(0, 0)$ and μ_3 to $(3, 0)$. The co-variance matrix of each variational distribution was fixed to $\sigma^2 I$ where $\sigma = 0.8$ and I is the identity matrix of size 2. We trained each variational distribution for 10000 iterations, sampling 1000 z's in each iteration. We used a learning rate of 0.001 for the Adam optimizer. The training seed was set to 0.

We evaluated our models on 10000 samples using seed = 1.

C Experiment 5.1.2 Details

C.1 Non-Hierarchical Case

We consider three variants of the true distribution, $p(z)$.

Setting (i): we let $p(z)$ be a uniform mixture of six Gaussians with $\mu \in \{-5, 0, 5, 10, 15, 20\}$ and $\sigma = 0.5$ (for all components). This setting is included in the main text and visualized in Figure 3.

Setting (ii): we let $p(z)$ be a uniform mixture of three Gaussians with $\mu \in \{0, 10, 20\}$ and $\sigma = 1.1$ (for all components). We visualize this setting here, in Figure 6.

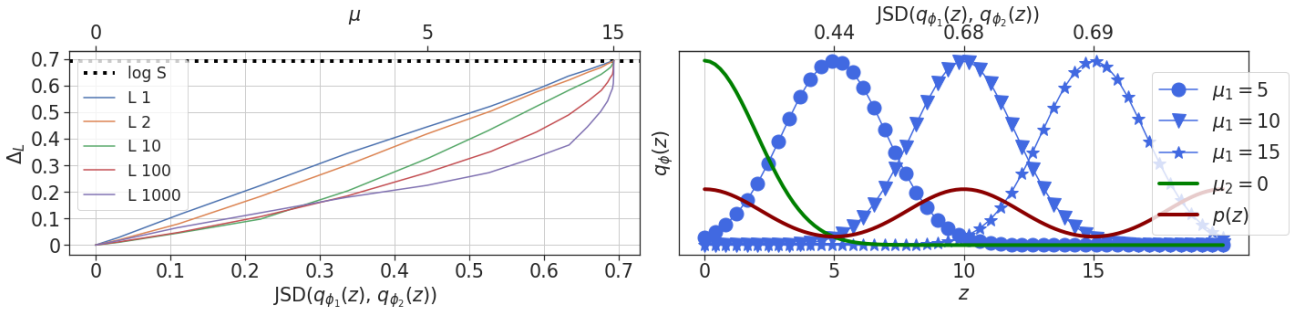


Figure 6: Setting (ii): trimodal Gaussian distribution and the corresponding JSD($\mathcal{Q}_S$).

Setting (iii): we let $p(z)$ be a uniform mixture of two Gaussians,

$$p(z) = \frac{1}{2}\mathcal{N}(z|0, 4) + \frac{1}{2}\mathcal{N}(z|10, 16). \tag{24}$$

We visualize this setting here, in Figure 7.

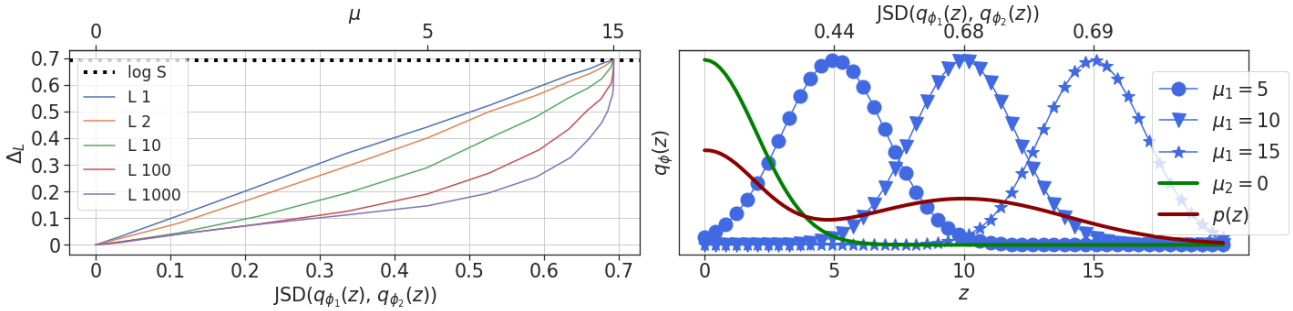


Figure 7: Setting (iii): bimodal Gaussian distribution and the corresponding JSD($\mathcal{Q}_S$).

C.2 Hierarchical Case

We also considered a hierarchical model, which is visualized in the main text (see Figure 4). We let $p(z, \mu) = p(z)p(\mu)$, where $p(\mu) = \mathcal{N}(10, 9)$ and $p(z)$ is the same as in setting (i) above.

D MNIST Experiment (5.2) Details

When using the NVAE (Vahdat and Kautz, 2020) model, we first trained $f_{\phi_1}(x)$ and $f_\theta(z)$ with the same hyperparameters and code as in the original paper (Vahdat and Kautz, 2020) for five different seeds: $1, 2, 3, 4$ and 5 (1 is used for the results in the original NVAE paper). The only exemption was related to the hardware: we used a single 32-GB Tesla V100 GPU, instead of two 16-GB Tesla V100 GPUs.

A complete description of the hyperparameters can be found in Table 6 of Appendix A in Vahdat and Kautz (2020), and in the MNIST experiment README at `https://github.com/NVlabs/NVAE`.

Training seeds - For the VAEs $\{f_{\phi_1}(x), f_\theta(z)\}$ trained with seeds $2, 3, 4, 5$, we used seed $= 0$ when training each corresponding $f_{\phi_2}(x)$. For $\{f_{\phi_1}(x), f_\theta(z)\}$ trained with seed 1, we used seed $= 2$ for $f_{\phi_2}(x)$; Where appropriate, we trained $f_{\phi_3}(x)$ with seed $= 3$.

Evaluation seed - We used seed $= 0$ during evaluation.

D.1 Ablation Study

In our ablation study, we used the same three models as in section 5.2. We then trained an additional encoder (using seed $= 4$) following the same scheme as for $f_{\phi_2}(x)$ and $f_{\phi_3}(x)$, making a total of $S = 4$. These were all trained using the entire MNIST training set. Due to computational restrictions, we then performed evaluations on eight subsets of the test data, 100 samples a time. The samples in the subsets were indexed by $\{1000 - 1099, 1100 - 1199, ..., 1700 - 1799\}$.

Finally, we calculated $\mathcal{L}^L_{\text{MIS}}$, JSD($\mathcal{Q}_S$) and $\overline{\mathcal{L}}_L$ for each subset, reporting the mean and standard deviation over the subsets for each quantity. The results are presented in Table 6. The entries are the means, and the standard deviations are in parentheses. We set seed $= 0$ for all evaluations in the study.

Observing the values in the table, we note that the JSD($\mathcal{Q}_S$) decreases as we go from $S = 3$ to $S = 4$. As we found in general that Δ_L decreases with the JSD($\mathcal{Q}_S$), we decided, based on our ablation study, not to perform the full experiment with $S = 4$ for computational reasons. Indeed, the largest Δ_L and best (mean) $\mathcal{L}^L_{\text{MIS}}$ were achieved when $S = 3, L = 50$ (bold entry in Table 6). Additionally, note that the JSD($\mathcal{Q}_S$) is estimated using $L \times S$ importance samples, and averaged over the data samples. This is the explanation for varying JSDs when they should, in theory, be independent of L.

Table 6: Mean and standard deviation of $\mathcal{L}_{\text{MIS}}^{L}$, JSD($\mathcal{Q}_S$) and $\overline{\mathcal{L}}_L$. Results from our ablation study on eight subsets of size 100 MNIST data using NVAE.

S	L	$\mathcal{L}_{\text{MIS}}^{L}$	JSD($\mathcal{Q}_S$)	$\overline{\mathcal{L}}_L$
2	1	79.46(2.055)	0.22(0.021)	79.67(2.059)
2	2	78.79(2.057)	0.26(0.007)	79.07(2.055)
2	3	78.55(2.086)	0.22(0.012)	78.82(2.079)
2	5	78.28(2.050)	0.25(0.005)	78.56(2.051)
2	10	78.02(2.065)	0.25(0.006)	78.30(2.065)
2	50	77.73(2.046)	0.25(0.003)	78.01(2.054)
3	1	79.42(2.040)	0.29(0.028)	79.70(2.045)
3	2	78.76(2.084)	0.29(0.013)	79.07(2.091)
3	3	78.50(2.065)	0.28(0.011)	78.81(2.058)
3	5	78.24(2.075)	0.28(0.006)	78.54(2.074)
3	10	78.00(2.045)	0.29(0.005)	78.31(2.050)
3	50	**77.71**(2.047)	0.28(0.002)	78.01(2.054)
4	1	79.41(2.066)	0.26(0.020)	79.67(2.065)
4	2	78.77(2.066)	0.27(0.012)	79.04(2.068)
4	3	78.53(2.058)	0.27(0.009)	78.82(2.054)
4	5	78.26(2.070)	0.27(0.009)	78.54(2.074)
4	10	78.03(2.042)	0.27(0.005)	78.31(2.045)
4	50	77.75(2.040)	0.27(0.004)	78.02(2.044)

E VBPI-NF Experiment (5.3) Details

VBPI-NF Details:

- Version: Repository cloned at 7 May 2021. `https://github.com/zcrabbit/vbpi-nf`
- Flow Type (flow_type): realnvp
- Number of layers for permutation invariant flow (lnf): 10
- Step size for branch length parameters (stepszBranch): 0.0001 (we consulted with the the author, 14 May 2021)
- Rest of the parameters are used with their default settings.
- We modified the code so that we can fix the seed to reproduce the results.

Example script:

```
python main.py –dataset data_name –flow_type realnvp –Lnf 10 –stepszBranch 0.0001 –vbpi_seed 1
```

Vbpi-Nf requires bootstrap trees to construct CPTs. The bootstrap trees for DS[1-4] are available in `https://github.com/zcrabbit/vbpi-nf`. For DS5, DS6 and DS8, we used UFBoot to create the bootstrap trees.

- IQ-TREE Version: 1.6.12 `http://www.iqtree.org/`
- Number of independent runs: 10
- Model (m): JC69
- Number of bootstrap replicates (bb): 10,000

Example script:

```
iqtree -s dataset_name -bb 10000 -wbt -m JC69 -redo
```

Lower bound details:

- For each dataset, VBPI-NF is run with 5 different seeds ($S = 5$) independently (used seeds values are $\{4, 15, 23, 42, 108\}$).
- For each trained model s, we sampled $L = 1000$ trees ($\tau_{1:L}$) and base branch lengths ($\lambda^{(0)}_{1:L}$).
- We used the same tree and branch length samples to compute the IWELBOs and MISELBO. Note that, in normalizing flows, one *samples* from the base distribution, whereas the final variables are obtained via a deterministic series of transformations. Here, this means that we sample base branch lengths, $\lambda^{(0)}_s$, from $q^{(0)}_{\phi_s}$. The final branch lengths, $\lambda^{(K+1)}_{s'}$, are not sampled, but obtained via the s'th model's normalizing flows.

Next we provide the expressions for the average IWELBOs and the MISELBO for this experiment. They are useful in order to understand how to apply normalizing flows in our framework.

IWELBO for VBPI-NF

$$\overline{\mathcal{L}}_L = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(\tau_{1:L},\lambda^{(0)}_{1:L}|x)}\left[\log\frac{1}{L}\sum_{\ell=1}^{L}\frac{p_\theta\big(x|\tau_{s,\ell},\lambda^{(K+1)}_{s,\ell}\big)p_\theta\big(\tau_{s,\ell},\lambda^{(K+1)}_{s,\ell}\big)}{q_{\phi_s}(\tau_{s,\ell})q^{(0)}_{\phi_s}\big(\lambda^{(0)}_{s,\ell}|\tau_{s,\ell}\big)\prod_{k=0}^{K}\big|\det\frac{\partial(\lambda^{(k+1)}_{s,\ell})}{\partial(\lambda^{(k)}_{s,\ell})}\big|^{-1}}\right]. \tag{25}$$

MISELBO for VBPI-NF

$$\mathcal{L}^L_{\text{MIS}} = \frac{1}{S}\sum_{s=1}^{S}\mathbb{E}_{q_{\phi_s}(\tau_{1:L},\lambda^{(0)}_{1:L}|x)}\left[\log\frac{1}{L}\sum_{\ell=1}^{L}\frac{p_\theta\big(x|\tau_{s,\ell},\lambda^{(K+1)}_{s,\ell}\big)p_\theta\big(\tau_{s,\ell},\lambda^{(K+1)}_{s,\ell}\big)}{\frac{1}{S}\sum_{s'=1}^{S}q_{\phi_{s'}}(\tau_{s,\ell})q^{(0)}_{\phi_{s'}}\big(\lambda^{(0)}_{s,\ell}|\tau_{s,\ell}\big)\prod_{k=0}^{K}\big|\det\frac{\partial(\lambda^{(k+1)}_{s',\ell})}{\partial(\lambda^{(k)}_{s',\ell})}\big|^{-1}}\right]. \tag{26}$$

Comment on the generative model: In the associated section in the main text, we state that there are no free (read *learnable*) parameters in the generative model, $p_\theta(z, \lambda, \tau)$. Meanwhile, we parameterize p by θ in order to emphasize that there are indeed model assumptions: The likelihood function assumes an evolutionary substitution model (JC69). The prior on branch lengths, λ, is assumed to be an exponential distribution, $p_\theta(\lambda) = \text{Exp}(10)$. The prior on the tree topologies, τ, is assumed to be a uniform distribution over the space of unrooted binary trees, $p_\theta(\tau) = \left(\frac{(2n-5)!}{2^{n-3}(n-3)!}\right)^{-1}$, where $n \geq 3$ are the number of taxa. When $n < 3$ there exists only a single topology.

All the above assumptions are the same as in Zhang (2020).