Doubly Semi-Implicit Variational Inference

Dmitry Molchanov[1,2,*]
dmolch111@gmail.com

Valery Kharitonov[2,*]
kharvd@gmail.com

Artem Sobolev[1]
asobolev@bayesgroup.ru

Dmitry Vetrov[1,2]
vetrovd@yandex.ru

[1] Samsung AI Center Moscow
[2] Samsung-HSE Laboratory, National Research University Higher School of Economics
[*] Equal contribution

Abstract

We extend the existing framework of semi-implicit variational inference (SIVI) and introduce doubly semi-implicit variational inference (DSIVI), a way to perform variational inference and learning when both the approximate posterior and the prior distribution are semi-implicit. In other words, DSIVI performs inference in models where the prior and the posterior can be expressed as an intractable infinite mixture of some analytic density with a highly flexible implicit mixing distribution. We provide a sandwich bound on the evidence lower bound (ELBO) objective that can be made arbitrarily tight. Unlike discriminator-based and kernel-based approaches to implicit variational inference, DSIVI optimizes a proper lower bound on ELBO that is asymptotically exact. We evaluate DSIVI on a set of problems that benefit from implicit priors. In particular, we show that DSIVI gives rise to a simple modification of VampPrior, the current state-of-the-art prior for variational autoencoders, which improves its performance.

1 INTRODUCTION

Bayesian inference is an important tool in machine learning. It provides a principled way to reason about uncertainty in parameters or hidden representations. In recent years, there has been great progress in scalable Bayesian methods, which made it possible to perform approximate inference for large-scale datasets and deep learning models.

One of such methods is variational inference (VI) [3], which is an optimization-based approach. Given a probabilistic model $p(x, z) = p(x \mid z)p(z)$, where x are observed data and z are latent variables, VI seeks to maximize the evidence lower bound (ELBO).

$$\mathcal{L}(\phi) = \mathbb{E}_{q_\phi(z)}[\log p(x \mid z)] - \mathrm{KL}(q_\phi(z) \, \| \, p(z)), \quad (1)$$

where $q_\phi(z)$ approximates the intractable true posterior distribution $p(z \mid x)$. The parametric approximation family for q_ϕ is chosen in such a way, that we can efficiently estimate $\mathcal{L}(\phi)$ and its gradients w.r.t. ϕ.

Such approximations to the true posterior are often too simplistic. There exists a variety of ways to extend the variational family to mitigate this. They can be divided roughly into two main groups: those that require the probability density function of the approximate posterior to be analytically tractable (which we will call *explicit* models) [30, 12, 9, 5, 36, 38, 6, 27, 20] and those that do not (*implicit* models) [11, 22, 37, 17, 21, 31, 40]. For latter, we only assume that it is possible to sample from such distributions, whereas the density may be inaccessible.

Not only approximate posteriors but also priors in such models are often chosen to be very simple to make computations tractable. This can lead to overregularization and poor hidden representations in generative models such as variational autoencoders (VAE, [15]) [10, 35, 1]. In Bayesian neural networks, a standard normal prior is the default choice, but together with the mean field posterior, it can lead to overpruning and consequently underfitting [39]. To overcome such problem in practice, one usually scales the KL divergence term in the expression for ELBO or truncates the variances of the approximate posterior [24, 19, 18].

Another way to overcome this problem is to consider more complicated prior distributions, e.g. implicit priors. For example, hierarchical priors usually impose an implicit marginal prior when hyperparameters are integrated out. To perform inference in such models, one often resorts to joint inference over both param-

eters and hyperparameters, even though we are only interested in the marginal posterior over parameters of the model. Another example of implicit prior distributions is the optimal prior for variational autoencoders. It can be shown that the aggregated posterior distribution is the optimal prior for VAE [10], and it can be regarded as an implicit distribution. The VampPrior model [35] approximates this implicit prior using an explicit discrete mixture of Gaussian posteriors. However, this model can be further improved if we consider an arbitrary trainable semi-implicit prior.

In this paper, we extend the recently proposed framework of semi-implicit variational inference (SIVI) [40] and consider priors and posteriors that are defined as semi-implicit distributions. By "semi-implicit" we mean distributions that do not have a tractable PDF (i.e. implicit), but that can be represented as a mixture of some analytically tractable density with a flexible mixing distribution, either explicit or implicit.

Our contributions can be summarized as follows. Firstly, we prove that the SIVI objective is actually a lower bound on the true ELBO, which allows us to sandwich the ELBO between an upper bound and a lower bound which are both asymptotically exact. Secondly, we propose *doubly semi-implicit variational inference* (DSIVI), a general-purpose framework for variational inference and variational learning in the case when both the posterior and the prior are semi-implicit. We construct a SIVI-inspired asymptotically exact lower bound on the ELBO for this case, and use the variational representation of the KL divergence to obtain the upper bound. Finally, we consider a wide range of applications where semi-implicit distributions naturally arise, and show how the use of DSIVI in these settings is beneficial.

2 PRELIMINARIES

Consider a probabilistic model defined by its joint distribution $p(x, z) = p(x \mid z)p(z)$, where variables x are observed, and z are the latent variables. Variational inference is a family of methods that approximate the intractable posterior distribution $p(z \mid x)$ with a tractable parametric distribution $q_\phi(z)$. To do so, VI methods maximize the evidence lower bound (ELBO):

$$\log p(x) \geq \mathcal{L}(\phi) = \mathbb{E}_{q_\phi(z)} \log \frac{p(x \mid z)p(z)}{q_\phi(z)} \to \max_\phi. \tag{2}$$

The maximum of the evidence lower bound corresponds to the minimum of the KL-divergence $\mathrm{KL}(q_\phi(z) \,\|\, p(z \mid x))$ between the variational distribution $q_\phi(z)$ and the exact posterior $p(z \mid x)$. In the more general *variational learning* setting, the prior distribution may also be a parametric distribution $p_\theta(z)$ [11].

In this case, one would optimize the ELBO w.r.t. both the variational parameters ϕ and the prior parameters θ, thus performing approximate maximization of the marginal likelihood $p(x \mid \theta)$.

The common way to estimate the gradient of this objective is to use the reparameterization trick [15]. The reparameterization trick recasts the sampling from the parametric distribution $q_\phi(z)$ as the sampling of non-parametric noise $\varepsilon \sim p(\varepsilon)$, followed by a deterministic parametric transformation $z = f(\varepsilon, \phi)$. Still, such gradient estimator requires log-densities of both the prior distribution $p(z)$ and the approximate posterior $q_\phi(z)$ in closed form. Several methods have been proposed to overcome this limitation [27, 21, 31]. However, such methods usually provide a biased estimate of the evidence lower bound with no practical way of estimating the introduced bias.

The reparameterizable distributions with no closed-form densities are usually referred to as *implicit distributions*. In this paper we consider the so-called semi-implicit distributions that are defined as an implicit mixture of explicit conditional distributions:

$$q_\phi(z) = \int q_\phi(z \mid \psi) q_\phi(\psi) \, d\psi. \tag{3}$$

Here, the conditional distribution $q_\phi(z \mid \psi)$ is explicit. However, when its condition ψ follows an implicit distribution $q_\phi(\psi)$, the resulting marginal distribution $q_\phi(z)$ is implicit. We will refer to $q_\phi(\psi)$ as the mixing distribution, and to ψ as the mixing variables.

Note that we may easily sample from semi-implicit distributions: in order to sample z from $q_\phi(z)$, we need to first sample the mixing variable $\psi \sim q_\phi(\psi)$, and then sample z from the conditional $q_\phi(z \mid \psi)$. Further in the text, we will assume this sampling scheme when using expectations $\mathbb{E}_{z \sim q_\phi(z)}$ over semi-implicit distributions. Also note that an arbitrary implicit distribution can be represented in a semi-implicit form: $q_\phi(z) = \int \delta(z - z') q_\phi(z') \, dz'$.

3 RELATED WORK

There are several approaches to inference and learning in models with implicit distributions.

One approach is commonly referred to as hierarchical variational inference or auxiliary variable models. It allows for inference with implicit approximate posteriors $q_\phi(z)$ that can be represented as a marginal distribution of an explicit joint distribution $q_\phi(z) = \int q_\phi(z, \psi) \, d\psi$. The ELBO is then bounded from below using a reverse variational model $r_\omega(\psi \mid z) \approx q_\phi(\psi \mid z)$ [27, 29, 19]. This method does not allow for implicit prior distributions, requires access to the explicit joint

density $q_\phi(z, \psi)$ and has no way to estimate the increased inference gap, introduced by the imperfect reverse model. However, recently proposed deep weight prior [2] provides a new lower bound, suitable for learning hierarchical priors in a similar fashion.

Another family of models uses an optimal discriminator to estimate the ratio of implicit densities $r(z) = \frac{q_\phi(z)}{p_\theta(z)}$ [21, 22, 11]. This is the most general approach to inference and learning with implicit distributions, but it also optimizes a biased surrogate ELBO, and the induced bias cannot be estimated. Also, different authors report that the performance of this approach is poor if the dimensionality of the implicit densities is high [32, 37]. This is the only approach that allows to perform variational learning (learning the parameters θ of the prior distribution $p_\theta(z)$). However, it is non-trivial and requires differentiation through a series of SGD updates. This approach has not been validated in practice yet and has only been proposed as a theoretical concept [11]. On the contrary, DSIVI provides a lower bound that can be directly optimized w.r.t. both the variational parameters ϕ and the prior parameters θ, naturally enabling variational learning.

Kernel implicit variational inference (KIVI) [31] is another approach that uses kernelized ridge regression to approximate the density ratio. It is reported to be more stable than the discriminator-based approaches, as the proposed density ratio estimator can be computed in closed form. Still, this procedure introduces a bias that is not addressed. Also, KIVI relies on adaptive contrast that does not allow for implicit prior distributions [21, 31].

There are also alternative formulations of variational inference that are based on different divergences. One example is operator variational inference [26] that uses the Langevin-Stein operator to design a new variational objective. Although it allows for arbitrary implicit posterior approximations, the prior distribution has to be explicit.

4 DOUBLY SEMI-IMPLICIT VARIATIONAL INFERENCE

In this section, we will describe semi-implicit variational inference, study its properties, and then extend it for the case of semi-implicit prior distributions.

4.1 Semi-Implicit Variational Inference

Semi-implicit variational inference [40] considers models with an explicit joint distribution $p(x, z)$ and a semi-implicit approximate posterior $q_\phi(z)$, as defined in Eq. (3). The basic idea of semi-implicit variational

inference is to approximate the semi-implicit approximate posterior with a finite mixture:

$$
\begin{aligned}
q_\phi(z) = \int q_\phi(z \mid \psi) q_\phi(\psi) \, d\psi \approx \\
\approx \frac{1}{K} \sum_{k=1}^{K} q_\phi(z \mid \psi^k), \quad \psi^k \sim q_\phi(\psi).
\end{aligned} \tag{4}
$$

SIVI provides an upper bound $\overline{\mathcal{L}}_K^q \geq \overline{\mathcal{L}}_{K+1}^q \geq \mathcal{L}$, and a surrogate objective $\underline{\mathcal{L}}_K^q$ that both converge to ELBO as K goes to infinity ($\underline{\mathcal{L}}_\infty^q = \overline{\mathcal{L}}_\infty^q = \mathcal{L}$):

$$
\overline{\mathcal{L}}_K^q = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) p(z) - \tag{5}
$$
$$
- \mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)} \mathbb{E}_{z \sim q_\phi(z \mid \psi^0)} \log \frac{1}{K} \sum_{k=1}^{K} q_\phi(z \mid \psi^k),
$$

$$
\underline{\mathcal{L}}_K^q = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) p(z) - \tag{6}
$$
$$
- \mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)} \mathbb{E}_{z \sim q_\phi(z \mid \psi^0)} \log \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z \mid \psi^k).
$$

The surrogate objective $\underline{\mathcal{L}}_K^q$ is then used for optimization.

4.2 SIVI Lower Bound

Although it was shown that $\underline{\mathcal{L}}_0^q$ is a lower bound for ELBO, it has not been clear whether this holds for arbitrary K, and whether maximizing $\underline{\mathcal{L}}_K^q$ leads to a correct procedure. Here, we show that $\underline{\mathcal{L}}_K^q$ is indeed a lower bound on ELBO $\mathcal{L}$.

Theorem 1. *Consider $\mathcal{L}$ and $\underline{\mathcal{L}}_K^q$ defined as in Eq. (2) and (6). Then $\underline{\mathcal{L}}_K^q$ converges to $\mathcal{L}$ from below as $K \to \infty$, satisfying $\underline{\mathcal{L}}_K^q \leq \underline{\mathcal{L}}_{K+1}^q \leq \mathcal{L}$, and*

$$
\underline{\mathcal{L}}_K^q = \mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)} \mathbb{E}_{q_\phi^K(z \mid \psi^{0..K})} \log \frac{p(x \mid z) p(z)}{q_\phi^K(z \mid \psi^{0..K})}, \tag{7}
$$

where $q_\phi^K(z \mid \psi^{0..K}) = \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z \mid \psi^k).$ \tag{8}

The proof can be found in Appendix A.

It can be seen from Eq. (7) that the surrogate objective $\underline{\mathcal{L}}_K^q$ proposed by [40] is actually the ELBO for a finite mixture approximation $q_\phi^K(z \mid \psi^0, \ldots, \psi^K)$, that is averaged over all such mixtures (averaged over samples of $\psi^0, \ldots, \psi^K \sim q_\phi(\psi)$).

4.3 Semi-Implicit Priors

Inspired by the derivation of the SIVI upper bound, we can derive the lower bound $\underline{\mathcal{L}}_K^p$ for the case of semi-implicit prior distributions. Right now, for simplicity,

assume an explicit approximate posterior $q_\phi(z)$, and a semi-implicit prior $p_\theta(z) = \int p_\theta(z|\zeta)p_\theta(\zeta)\,d\zeta$

$$\underline{\mathcal{L}}_K^p = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) -$$
$$- \mathbb{E}_{\zeta^{1..K} \sim p_\theta(\zeta)} \mathbb{E}_{q_\phi(z)} \log \frac{q_\phi(z)}{\frac{1}{K}\sum_{k=1}^{K} p_\theta(z \mid \zeta^k)}, \quad (9)$$

$$\underline{\mathcal{L}}_K^p \leq \underline{\mathcal{L}}_{K+1}^p \leq \underline{\mathcal{L}}_\infty^p = \mathcal{L}. \quad (10)$$

This bound has the same properties: it is non-decreasing in K and is asymptotically exact. To see why $\underline{\mathcal{L}}_K^p \leq \mathcal{L}$, one just needs to apply the Jensen's inequality for the logarithm:

$$\mathbb{E}_{\zeta^{1..K} \sim p_\theta(\zeta)} \mathbb{E}_{q_\phi(z)} \log \frac{1}{K} \sum_{k=1}^{K} p_\theta(z \mid \zeta^k) \leq$$
$$\leq \mathbb{E}_{q_\phi(z)} \log \mathbb{E}_{\zeta^{1..K} \sim p_\theta(\zeta)} \frac{1}{K} \sum_{k=1}^{K} p_\theta(z \mid \zeta^k) = \quad (11)$$
$$= \mathbb{E}_{q_\phi(z)} \log p_\theta(z).$$

To show that this bound is non-decreasing in K, one can refer to the proof of proposition 3 in the SIVI paper [40, Appendix A].

Note that it is no longer possible to use the same trick to obtain the upper bound. Still, we can obtain an upper bound using the variational representation of the KL-divergence [25]:

$$\mathrm{KL}(q_\phi(z) \parallel p_\theta(z)) =$$
$$= 1 + \sup_{g:\mathrm{dom}\,z \to \mathbb{R}} \left\{ \mathbb{E}_{q_\phi(z)} g(z) - \mathbb{E}_{p_\theta(z)} e^{g(z)} \right\} \geq$$
$$\geq 1 + \sup_{\eta} \left\{ \mathbb{E}_{q_\phi(z)} g(z, \eta) - \mathbb{E}_{p_\theta(z)} e^{g(z,\eta)} \right\}, \quad (12)$$

$$\overline{\mathcal{L}}_\eta^p = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) -$$
$$- 1 - \mathbb{E}_{q_\phi(z)} g(z, \eta) + \mathbb{E}_{p_\theta(z)} e^{g(z,\eta)}. \quad (13)$$

Here we substitute the maximization over all functions with a single parametric function. In order to obtain a tighter bound, we can minimize this bound w.r.t. the parameters η of function $g(z, \eta)$.

Note that in order to find the optimal value for η, one does not need to estimate the entropy term or the likelihood term of the objective:

$$\eta^* = \arg\min_\eta \left[-\mathbb{E}_{q_\phi(z)} g(z, \eta) + \mathbb{E}_{p_\theta(z)} e^{g(z,\eta)} \right]. \quad (14)$$

This allows us to obtain a lower bound on the KL-divergence between two arbitrary (semi-)implicit distributions, and, consequently, results in an upper bound on the ELBO.

4.4 Final Objective

We can combine the bounds for the semi-implicit posterior and the semi-implicit prior to obtain the final

lower bound

$$\underline{\underline{\mathcal{L}}}_{K_1, K_2}^{q,p} = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) -$$
$$- \mathbb{E}_{\psi^{0..K_1} \sim q_\phi(\psi)} \mathbb{E}_{q_\phi(z \mid \psi^0)} \log \frac{1}{K_1 + 1} \sum_{k=0}^{K_1} q_\phi(z \mid \psi^k) +$$
$$+ \mathbb{E}_{\zeta^{1..K_2} \sim p_\theta(\zeta)} \mathbb{E}_{q_\phi(z)} \log \frac{1}{K_2} \sum_{k=1}^{K_2} p_\theta(z \mid \zeta^k), \quad (15)$$

and the upper bound

$$\overline{\mathcal{L}}_\eta^{q,p} = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) -$$
$$- 1 - \mathbb{E}_{q_\phi(z)} g(z, \eta) + \mathbb{E}_{p_\theta(z)} e^{g(z,\eta)}. \quad (16)$$

The lower bound $\underline{\underline{\mathcal{L}}}_{K_1, K_2}^{q,p}$ is non-decreasing in both K_1 and K_2, and is asymptotically exact:

$$\underline{\underline{\mathcal{L}}}_{K_1, K_2}^{q,p} \leq \underline{\underline{\mathcal{L}}}_{K_1+1, K_2}^{q,p}, \qquad \underline{\underline{\mathcal{L}}}_{K_1, K_2}^{q,p} \leq \underline{\underline{\mathcal{L}}}_{K_1, K_2+1}^{q,p}, \quad (17)$$
$$\lim_{K_1, K_2 \to \infty} \underline{\underline{\mathcal{L}}}_{K_1, K_2}^{q,p} = \mathcal{L}. \quad (18)$$

We use the lower bound for optimization, whereas the upper bound may be used to estimate the gap between the lower bound and the true ELBO. The final algorithm for DSIVI is presented in Algorithm 1. Unless stated otherwise, we use 1 MC sample to estimate the gradients of the lower bound (see Algorithm 1 for more details). In the case where the prior distribution is explicit, one may resort to the upper bound $\overline{\mathcal{L}}_K^q$, proposed in SIVI [40].

5 APPLICATIONS

In this section we describe several settings that can benefit from semi-implicit prior distributions.

5.1 VAE with Semi-Implicit Priors

The default choice of the prior distribution $p(z)$ for the VAE model is the standard Gaussian distribution. However, such choice is known to over-regularize the model [35, 8].

It can be shown that the so-called aggregated posterior distribution is the optimal prior distribution for a VAE in terms of the value of ELBO [10, 35]:

$$p^*(z) = \frac{1}{N} \sum_{n=1}^{N} q_\phi(z \mid x_n), \quad (19)$$

where the summation is over all training samples x_n, $n = 1, \ldots, N$. However, this extreme case leads to overfitting [10, 35], and is highly computationally inefficient. A possible middle ground is to consider

---

Algorithm 1 Doubly semi-implicit VI (and learning)

Require: SI posterior $q_\phi(z) = \int q_\phi(z \mid \psi) q_\phi(\psi) \, d\psi$
Require: SI prior $p_\theta(z) = \int p_\theta(z \mid \zeta) p_\theta(\zeta) \, d\zeta$
Require: explicit log-likelihood $\log p(x \mid z)$
   Variational inference (find ϕ) and learning (find θ)
   for $t \leftarrow 1$ to T **do**
      $\psi^0, \ldots, \psi^{K_1} \sim q_\phi(\psi)$ ▷ Reparameterization
      $\zeta^1, \ldots, \zeta^{K_2} \sim p_\theta(\zeta)$ ▷ Reparameterization
      $z \sim q_\phi(z \mid \psi^0)$ ▷ Reparameterization
      Estimate $L_{LH} \simeq \log p(x \mid z)$
      $L_E \leftarrow -\log \frac{1}{K_1+1} \sum_{k=0}^{K_1} q_\phi(z \mid \psi^k)$
      $L_{CE} \leftarrow -\log \frac{1}{K_2} \sum_{k=1}^{K_2} p_\theta(z \mid \zeta^k)$
      $\hat{\underline{\underline{\mathcal{L}}}}_{K_1,K_2}^{q,p} \leftarrow L_{LH} + L_E - L_{CE}$
      Use $\nabla_\phi \hat{\underline{\underline{\mathcal{L}}}}_{K_1,K_2}^{q,p}$ to update ϕ
      if Variational learning **then**
         Use $\nabla_\theta \hat{\underline{\underline{\mathcal{L}}}}_{K_1,K_2}^{q,p}$ to update θ
      end if
   end for
   Upper bound
   for $t \leftarrow 1$ to T **do**
      $z \sim q_\phi(z)$ ▷ Reparameterization
      $z' \sim p_\theta(z)$ ▷ Reparameterization
      $L \leftarrow -g(z, \eta) + e^{g(z', \eta)}$
      Use $-\nabla_\eta L$ to update η
   end for
   Estimate $\underline{\underline{\mathcal{L}}}_{K_1,K_2}^{q,p}$ and $\overline{\mathcal{L}}_\eta^{q,p}$ using Eq. (15) and (16)
   return $\phi, \theta, \eta, \underline{\underline{\mathcal{L}}}_{K_1,K_2}^{q,p}, \overline{\mathcal{L}}_\eta^{q,p}$

---

the *variational mixture of posteriors* prior distribution (the VampPrior) [35]:

$$p^{Vamp}(z) = \frac{1}{K} \sum_{k=1}^{K} q_\phi(z \mid u_k). \qquad (20)$$

The VampPrior is defined as a mixture of K variational posteriors $q_\phi(z \mid u_k)$ for a set of inducing points $\{u_k\}_{k=1}^{K}$. These inducing points may be learnable (an ordinary VampPrior) or fixed at a random subset of the training data (VampPrior-data). The VampPrior battles over-regularization by considering a flexible empirical prior distribution, being a mixture of fully-factorized Gaussians, and by coupling the parameters of the prior distribution and the variational posteriors.

There are two ways to improve this technique by using DSIVI. We can regard the aggregated posterior $p^*(z)$ as a semi-implicit distribution:

$$p^*(z) = \frac{1}{N} \sum_{n=1}^{N} q_\phi(z|x_n) = \int q_\phi(z|x) p_{data}(x) dx. \quad (21)$$

Next, we can use it as a semi-implicit prior and exploit

the lower bound, presented in Section 4.3:

$$\underline{\mathcal{L}}_K^p = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{q_\phi(z \mid x_n)} \left[\log \frac{p(x_n \mid z)}{q_\phi(z \mid x_n)} + \right.$$
$$\left. + \mathbb{E}_{u_{1..K} \sim p_{data}(x)} \log \frac{1}{K} \sum_{k=1}^{K} q_\phi(z \mid u_k) \right]. \qquad (22)$$

Note that the only difference from the training objective of VampPrior-data is that the inducing points u_k are not *fixed*, but are *resampled* at each estimation of the lower bound. As we show in the experiments, such reformulation of VampPrior-data drastically improves its test log-likelihood.

We can also consider an arbitrary semi-implicit prior distribution:

$$p_\theta^{SI}(z) = \int p_\theta(z \mid \zeta) p_\theta(\zeta) \, d\zeta. \qquad (23)$$

For example, we consider a fully-factorized Gaussian conditional prior $p_\theta(z \mid \zeta) = \mathcal{N}(z \mid \zeta, \text{diag}(\sigma^2))$ with mean ζ and trainable variances σ_j^2. The implicit generator $p_\theta(\zeta)$ can be parameterized by an arbitrary neural network with weights θ that transforms a standard Gaussian noise ε to mixing parameters ζ. As we show in the experiments, such semi-implicit posterior outperforms VampPrior even though it does not couple the parameters of the prior and the variational posteriors.

We can also apply the importance-weighted lower bound [4] similarly to the importance weighted SIVAE [40], and obtain IW-DSIVAE, a lower bound on the IWAE objective for a variational autoencoder with a semi-implicit prior and a semi-implicit posterior. The exact expression for this lower bound is presented in Appendix B.

5.2 Variational Inference with Hierarchical Priors

A lot of probabilistic models use hierarchical prior distributions: instead of a non-parametric prior $p(w)$ they use a parametric conditional prior $p(w \mid \alpha)$ with hyperparameters α, and a hyperprior over these parameters $p(\alpha)$. A discriminative model with such hierarchical prior may be defined as follows [23, 33, 34, 7, 31, 18]:

$$p(t, w, \alpha \mid x) = p(t \mid x, w) p(w \mid \alpha) p(\alpha). \qquad (24)$$

A common way to perform inference in such models is to approximate the joint posterior $q_\phi(w, \alpha) \approx p(w, \alpha \mid X_{tr}, T_{tr})$ given the training data (X_{tr}, T_{tr}) [7, 31, 18]. Then the marginal approximate posterior $q_\phi(w) = \int q_\phi(w, \alpha) \, d\alpha$ is used to approximate the pre-

dictive distribution on unseen data $p(t \,|\, x, X_{tr}, T_{tr})$:

$$
\begin{aligned}
p(t \,|\, x, X_{tr}, T_{tr}) &= \int p(t \,|\, x, w) p(w \,|\, X_{tr}, T_{tr}) \, dw = \\
&= \int p(t \,|\, x, w) \int p(w, \alpha \,|\, X_{tr}, T_{tr}) \, d\alpha \, dw \approx \\
&\approx \int p(t \,|\, x, w) \int q_\phi(w, \alpha) d\alpha \, dw = \\
&= \int p(t \,|\, x, w) q_\phi(w) \, dw.
\end{aligned}
\tag{25}
$$

The inference is performed by maximization of the following variational lower bound:

$$
\mathcal{L}^{joint}(\phi) = \mathbb{E}_{q_\phi(w,\alpha)} \log \frac{p(t \,|\, x, w) p(w \,|\, \alpha) p(\alpha)}{q_\phi(w, \alpha)}. \tag{26}
$$

We actually are not interested in the joint posterior $q_\phi(w, \alpha)$, and we only need it to obtain the marginal posterior $q_\phi(w)$. In this case we can reformulate the problem as variational inference with a semi-implicit prior $p(w) = \int p(w \,|\, \alpha) p(\alpha) \, d\alpha$ and a semi-implicit posterior $q_\phi(w) = \int q_\phi(w \,|\, \alpha) q_\phi(\alpha) \, d\alpha$:

$$
\mathcal{L}^{marginal}(\phi) = \mathbb{E}_{q_\phi(w)} \log \frac{p(t \,|\, x, w) p(w)}{q_\phi(w)}. \tag{27}
$$

Then it can be shown that optimization of the second objective results in a better fit of the marginal posterior:

Theorem 2. *Let ϕ_j and ϕ_m maximize $\mathcal{L}^{joint}$ and $\mathcal{L}^{marginal}$ correspondingly. Then*

$$
\begin{aligned}
\mathrm{KL}(q_{\phi_m}(w) \,\|\, p(w \,|\, X_{tr}, T_{tr})) &\leq \\
\mathrm{KL}(q_{\phi_j}(w) &\,\|\, p(w \,|\, X_{tr}, T_{tr})). \tag{28}
\end{aligned}
$$

The proof can be found in Appendix C.

It means that if the likelihood function does not depend on the hyperparameters α, it is beneficial to consider the semi-implicit formulation instead of the joint formulation of variational inference even if the approximation family stays exactly the same. In the experiments, we show that the proposed DSIVI procedure matches the performance of direct optimization of $\mathcal{L}^{marginal}$, whereas joint VI performs much worse.

6 EXPERIMENTS

6.1 Variational Inference with Hierarchical Priors

We consider a Bayesian neural network with a fully-factorized hierarchical prior distribution with a Gaussian conditional $p(w_{ij} \,|\, \alpha_{ij}) = \mathcal{N}(w_{ij} \,|\, 0, \alpha_{ij}^{-1})$ and a Gamma hyperprior over the inverse variances $p(\alpha_{ij}) =$

Gamma$(\alpha_{ij} \,|\, 0.5, 2)$. Such hierarchical prior induces a fully-factorized Student's t-distribution with one degree of freedom as the marginal prior $p(w_{ij}) = t(w_{ij} \,|\, \nu = 1)$. Note that in this case, we can estimate the marginal evidence lower bound directly. We consider a fully-connected neural network with two hidden layers of 300 and 100 neurons on the MNIST dataset [16]. We train all methods with the same hyperparameters: we use batch size 200, use Adam optimizer [13] with default parameters, starting with learning rate 10^{-3}, and train for 200 epochs, using linear learning rate decay.

We consider three different ways to perform inference in this model, the marginal inference, the joint inference, and DSIVI, as described in Section 5.2. For joint inference, we consider a fully-factorized joint approximate posterior $q_\phi(w, \alpha) = q_\phi(w) q_\phi(\alpha)$, with $q_\phi(w)$ being a fully-factorized Gaussian, and $q_\phi(\alpha)$ being a fully-factorized Log-Normal distribution. Such joint approximate posterior induces a fully-factorized Gaussian marginal posterior $q_\phi(w)$. Therefore, we use a fully-factorized Gaussian posterior for the marginal inference and DSIVI. Note that in this case, only the prior distribution is semi-implicit. All models have been trained with the local reparameterization trick [14].

We perform inference using these three different variational objectives, and then compare the true evidence lower bound $\mathcal{L}^{marginal}$ on the training set. As the marginal variational approximation is the same in all four cases, the training ELBO can act as a proxy metric for the KL-divergence between the marginal approximate posterior and the true marginal posterior. The results are presented in Figure 1. DSIVI with as low as $K = 10$ samples during training exactly matches the performance of the true marginal variational inference, whereas other approximations fall far behind. All three methods achieve $97.7-98.0\%$ test set accuracy, and the test log-likelihood is approximately the same for all methods, ranging from -830 to -855. However, the difference in the marginal ELBO is high. The final values of the ELBO, its decomposition into train log-likelihood and the KL term, and the test log-likelihood are presented in Table 2 in Appendix C.

6.2 Comparison to Alternatives

We compare DSIVI to other methods for implicit VI on a toy problem of approximating a centered standard Student's t-distribution $p(z)$ with 1 degrees of freedom with a Laplace distribution $q_\phi(z)$ by representing them as scale mixtures of Gaussians. Namely, we represent $p(z) = \int \mathcal{N}(z \,|\, 0, \alpha^{-1}) \text{Gamma}(\alpha \,|\, 0.5, 2) \, d\alpha$, and $q_\phi(z) = \int \mathcal{N}(z \,|\, \mu, \tau) \text{Exp}(\tau \,|\, \lambda) \, d\tau$. We train all methods by minimizing the corresponding approxima-
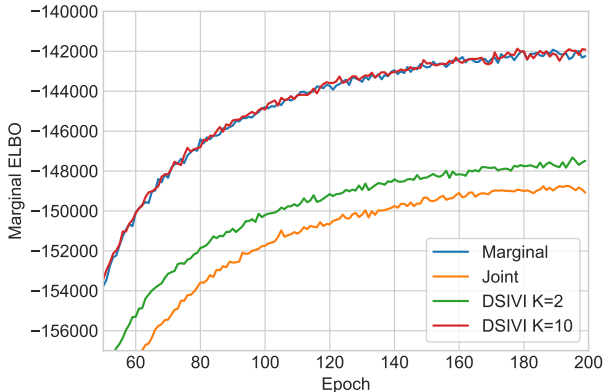
Figure 1: Variational inference with a hierarchical prior. Models are trained using different variational objectives. The estimates of the marginal evidence lower bound are presented in this plot.
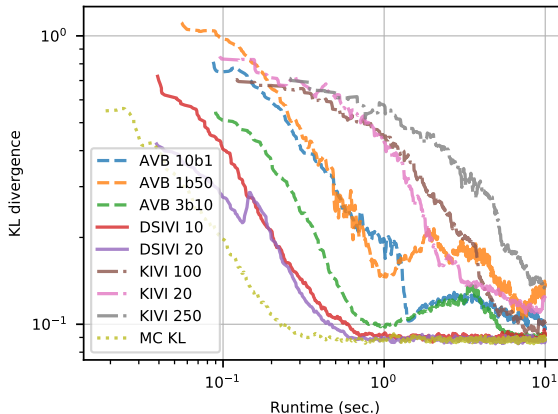

Figure 2: Comparison of different techniques for (semi-)implicit VI. "KIVI K" corresponds to KIVI, K being the number of MC samples, used to approximate KL divergence; "DSIVI K" corresponds to DSIVI with K=K; "AVB MbK" corresponds to AVB with M updates of discriminator per one update of ϕ and K MC samples to estimate the discriminator's gradients. "MC KL" corresponds to direct stochastic minimization of the KL divergence.

tions to the KL-divergence $\mathrm{KL}(q_\phi(z) \,\|\, p(z))$ w.r.t. the parameters μ and λ of the approximation $q_\phi(z)$.

As baselines, we use prior methods for implicit VI: Adversarial Variational Bayes (AVB) [21], which is a discriminator-based method, and Kernel Implicit Variational Inference (KIVI) [31]. For AVB we fix architecture of the "discriminator" neural network to have 2 hidden layers with 3 and 4 hidden units with LeakyReLU ($\alpha = 0.2$) activation, and for KIVI we use fixed $\lambda = 0.001$ with varying number of samples. For AVB we tried different numbers of training samples and optimization steps to optimize the discriminator at each step of optimizing over ϕ. We used Adam optimizer with learning rate 10^{-2} and one MC sample to estimate gradients w.r.t. ϕ.

We report the KL-divergence $KL(q_\phi(z) \,\|\, p(z))$, estimated using 10000 MC samples averaged over 10 runs. The results are presented in Figure 2. DSIVI converges faster, is more stable, and only has one hyperparameter, the number of samples K in the DSIVI objective.

6.3 Sequential Approximation

We illustrate the expressiveness of DSIVI with implicit prior and posterior distributions on the following toy problem. Consider an explicit distribution $p(z)$. We would like to learn a semi-implicit distribution $q_{\phi_1}(z) = \mathbb{E}_{q_{\phi_1}(\psi)}[q_{\phi_1}(z \,|\, \psi)]$ to match $p(z)$. During the first step, we apply DSIVI to tune the parameters ϕ_1 so as to minimize $\mathrm{KL}(q_{\phi_1}(z) \,\|\, p(z))$. Then, we take the trained semi-implicit $q_{\phi_1}(z)$ as a new target for z and tune ϕ_2 minimizing $\mathrm{KL}(q_{\phi_2}(z) \,\|\, q_{\phi_1}(z))$. After we repeat the iterative process k times, $q_{\phi_k}(z)$ obtained through minimization of $\mathrm{KL}(q_{\phi_k}(z) \,\|\, q_{\phi_{k-1}}(z))$ should still match $p(z)$.

In our experiments, we follow [40] and model $q_{\phi_i}(\psi)$ by a multi-layer perceptron (MLP) with layer widths [30,60,30] with ReLU activations and a ten-dimensional standard normal noise as its input. We also fix all conditionals $q_{\phi_i}(z \,|\, \psi) = \mathcal{N}(z \,|\, \psi, \sigma^2 I)$, $\sigma^2 = 0.1$. We choose $p(z)$ to be either a one-dimensional mixture of Gaussians or a two-dimensional "banana" distribution. In Figure 3 we plot values of $\mathrm{KL}(q_{\phi_i}(z) \,\|\, p(z))$, $i = 1, \ldots, 9$ for different values of $K_1 = K_2 = K$ (see Algorithm 1) when $p(z)$ is a one-dimensional mixture of Gaussians (see Appendix D for a detailed description and additional plots). In Figure 4 we plot the approximate PDF of $q_{\phi_k}(z)$ after 9 steps for different values of K. As we can see, even though both "prior" and "posterior" distributions are semi-implicit, the algorithm can still accurately learn the original target distribution after several iterations.

6.4 VAE with Semi-Implicit Optimal Prior

We follow the same experimental setup and use the same hyperparameters, as suggested for VampPrior [35]. We consider two architectures, the VAE and the HVAE (hierarchical VAE, [35]), applied to the MNIST dataset with dynamic binarization [28]. In both cases, all distributions (except the prior) have been modeled by fully-factorized neural networks with two hidden layers of 300 hidden units each. We used 40-dimensional latent vectors z (40-dimensional z_1 and z_2 for HVAE) and Bernoulli likelihood with dynamic binarization for the MNIST dataset. As suggested in the VampPrior paper, we used 500 pseudo-inputs for VampPrior-based models in all cases (higher num-
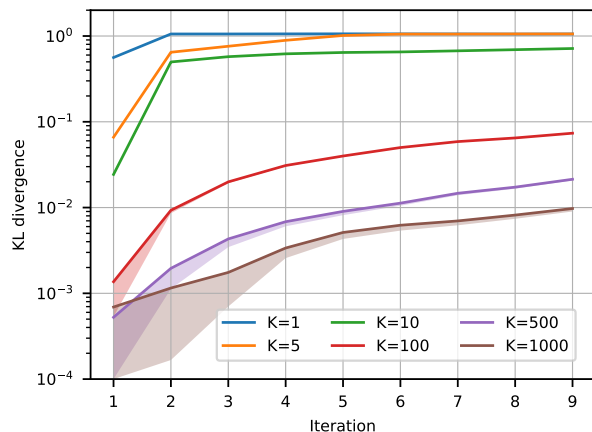
Figure 3: Sequential approximation. Area is shaded between lower and upper bounds of $\mathrm{KL}(q_{\phi_i}(z) \,\|\, p(z))$ for different *training* values of $K_1 = K_2 = K$, and the solid lines represent the corresponding upper bounds. During *evaluation*, $K = 10^4$ is used. Here $p(z)$ is a one dimensional Gaussian mixture (see Appendix D for details.) Lower is better.

Table 1: We compare VampPrior with its semi-implicit modifications, DSIVI-agg and DSIVI-prior. We report the the IWAE objective $\mathcal{L}^S$ for VampPrior-data, and the corresponding lower bound $\underline{\mathcal{L}}_K^{p,S}$ for DSIVI-based methods (see Appendix B). Only the prior distribution is semi-implicit.

Method	LL
VAE+VampPrior-data	-85.05
VAE+VampPrior	-82.38
VAE+DSIVI-prior (K=2000)	≥ -82.27
VAE+DSIVI-agg (K=500)	≥ -83.02
VAE+DSIVI-agg (K=5000)	$\geq \mathbf{-82.16}$
HVAE+VampPrior-data	-81.71
HVAE+VampPrior	-81.24
HVAE+DSIVI-agg (K=5000)	$\geq \mathbf{-81.09}$

ber of pseudo-inputs led to overfitting). To measure the performance of all models, we bound the test log-likelihood with the IWAE objective [4] with 5000 samples for the VampPrior-based methods, and estimate the corresponding IW-DSIVAE lower bound with $K = 20000$ for the DSIVI-based methods (see Appendix B for more details).

We consider two formulations, described in Section 5.1: DSIVI-agg stands for the semi-implicit formulation of the aggregated posterior (21), and DSIVI-prior stands for a general semi-implicit prior (23). For the DSIVI-prior we have used a fully-factorized Gaussian conditional $p(z \,|\, \zeta) = \mathcal{N}(z \,|\, \zeta, \mathrm{diag}(\sigma^2))$, where the mixing parameters ζ are the output of a fully-connected neural network with two hidden layers with 300 and 600 hidden units respectively, applied to a 300-dimensional standard Gaussian noise ϵ. The first and second hid-
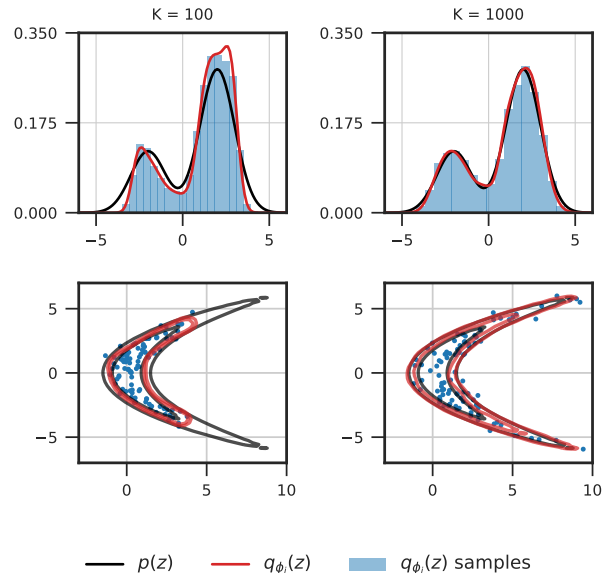

Figure 4: Learned probability distributions $q_{\phi_k}(z)$ after 9 iterations of sequential approximation for $K_1 = K_2 = 100$ and 1000 (red), and the two original priors $p(z)$ (black). During evaluation, $K = 10^4$ is used.

den layers were followed by ReLU non-linearities, and no non-linearities were applied to obtain ζ. We did not use warm-up [35] with DSIVI-prior.

The results are presented in Table 1. DSIVI-agg is a simple modification of VampPrior-data that significantly improves the test log-likelihood, and even outperforms the VampPrior with trained inducing inputs. DSIVI-prior outperforms VampPrior even without warm-up and without coupling the parameters of the prior and the variational posteriors.

7 CONCLUSION

We have presented DSIVI, a general-purpose framework that allows to perform variational inference and variational learning when both the approximate posterior distribution and the prior distribution are semi-implicit. DSIVI provides an asymptotically exact lower bound on the ELBO, and also an upper bound that can be made arbitrarily tight. It allows us to estimate the ELBO in any model with semi-implicit distributions, which was not the case for other methods. We have shown the effectiveness of DSIVI applied to a range of problems, e.g. models with hierarchical priors and variational autoencoders with semi-implicit empirical priors. In particular, we show how DSIVI-based treatment improves the performance of VampPrior, the current state-of-the-art prior distribution for VAE.

ACKNOWLEDGMENTS

References

[1] A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168, 2018.

[2] A. Atanov, A. Ashukha, K. Struminsky, D. Vetrov, and M. Welling. The deep weight prior. In *International Conference on Learning Representations*, 2019.

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

[4] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[5] R. J. Giordano, T. Broderick, and M. I. Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449, 2015.

[6] S. Han, X. Liao, D. Dunson, and L. Carin. Variational gaussian copula inference. In *Artificial Intelligence and Statistics*, pages 829–838, 2016.

[7] J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

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

[9] M. D. Hoffman and D. M. Blei. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, 2015.

[10] M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, 2016.

[11] F. Huszár. Variational inference using implicit distributions. *arXiv preprint arXiv:1702.08235*, 2017.

[12] T. S. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. In *Learning in graphical models*, pages 163–173. Springer, 1998.

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

[14] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583, 2015.

[15] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[17] Y. Li and R. E. Turner. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

[18] C. Louizos, K. Ullrich, and M. Welling. Bayesian compression for deep learning. In *Advances in Neural Information Processing Systems*, pages 3288–3298, 2017.

[19] C. Louizos and M. Welling. Multiplicative normalizing flows for variational bayesian neural networks. *arXiv preprint arXiv:1703.01961*, 2017.

[20] L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther. Auxiliary deep generative models. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1445–1453, New York, New York, USA, 20–22 Jun 2016. PMLR.

[21] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *ICML*, 2017.

[22] S. Mohamed and B. Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[23] R. M. Neal. *Bayesian learning for neural networks*, volume 118. 1995.

[24] K. Neklyudov, D. Molchanov, A. Ashukha, and D. P. Vetrov. Structured bayesian pruning via log-normal multiplicative noise. In *Advances in Neural Information Processing Systems*, pages 6775–6784, 2017.

[25] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[26] R. Ranganath, D. Tran, J. Altosaar, and D. Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.

[27] R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016.

[28] R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pages 872–879. ACM, 2008.

[29] T. Salimans, D. Kingma, and M. Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226, 2015.

[30] L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In *Advances in neural information processing systems*, pages 486–492, 1996.

[31] J. Shi, S. Sun, and J. Zhu. Kernel implicit variational inference. *arXiv preprint arXiv:1705.10119*, 2017.

[32] M. Sugiyama, T. Suzuki, and T. Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[33] M. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. 1:211–244, 2000.

[34] M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *International Conference on Machine Learning*, pages 1971–1979, 2014.

[35] J. M. Tomczak and M. Welling. Vae with a vampprior. *arXiv preprint arXiv:1705.07120*, 2017.

[36] D. Tran, D. Blei, and E. M. Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems*, pages 3564–3572, 2015.

[37] D. Tran, R. Ranganath, and D. Blei. Hierarchical implicit models and likelihood-free variational inference. In *Advances in Neural Information Processing Systems*, pages 5523–5533, 2017.

[38] D. Tran, R. Ranganath, and D. M. Blei. The variational gaussian process. *ICLR*, 2016.

[39] B. Trippe and R. Turner. Overpruning in variational bayesian neural networks. *arXiv preprint arXiv:1801.06230*, 2018.

[40] M. Yin and M. Zhou. Semi-implicit variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 5660–5669. PMLR, 2018.

A Proof of the SIVI Lower Bound for Semi-Implicit Posteriors

Theorem 1. *Consider* $\mathcal{L}$ *and* $\underline{\mathcal{L}}_K^q$ *defined as in Eq. (2) and (6). Then* $\underline{\mathcal{L}}_K^q$ *converges to* $\mathcal{L}$ *from below as* $K \to \infty$, *satisfying* $\underline{\mathcal{L}}_K^q \leq \underline{\mathcal{L}}_{K+1}^q \leq \mathcal{L}$, *and*

$$\underline{\mathcal{L}}_K^q = \mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)} \mathbb{E}_{q_\phi^K(z \,|\, \psi^{0..K})} \log \frac{p(x \,|\, z)p(z)}{q_\phi^K(z \,|\, \psi^{0..K})}, \tag{29}$$

$$where \quad q_\phi^K(z \,|\, \psi^{0..K}) = \frac{1}{K+1} \sum_{k=0}^K q_\phi(z \,|\, \psi^k). \tag{30}$$

Proof. For brevity, we denote $\mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)}$ as $\mathbb{E}_{\psi^{0..K}}$ and $\mathbb{E}_{z \sim q_\phi^K(z \,|\, \psi^{0..K})}$ as $\mathbb{E}_{z \,|\, \psi^{0..K}}$. First, notice that due to the symmetry in the indices, the regularized lower bound $\underline{\mathcal{L}}_K^q$ does not depend on the index in the conditional $q_\phi(z \,|\, \psi^i)$:

$$\underline{\mathcal{L}}_K^q = \mathbb{E}_{\psi^{0..K}} \mathbb{E}_{z \,|\, \psi^0} \log \frac{p(x, z)}{q_\phi^K(z \,|\, \psi^{0..K})} = \tag{31}$$

$$= \mathbb{E}_{\psi^{0..K}} \mathbb{E}_{z \,|\, \psi^i} \log \frac{p(x, z)}{q_\phi^K(z \,|\, \psi^{0..K})}. \tag{32}$$

Therefore, we can rewrite $\underline{\mathcal{L}}_K^q$ as follows:

$$\underline{\mathcal{L}}_K^q = \frac{1}{K+1} \sum_{i=0}^K \underline{\mathcal{L}}_K^q = \tag{33}$$

$$= \frac{1}{K+1} \sum_{i=0}^K \mathbb{E}_{\psi^{0..K}} \mathbb{E}_{z \,|\, \psi^i} \log \frac{p(x, z)}{q_\phi^K(z \,|\, \psi^{0..K})} = \tag{34}$$

$$= \mathbb{E}_{\psi^{0..K}} \mathbb{E}_{z \,|\, \psi^{0..K}} \log \frac{p(x, z)}{q_\phi^K(z \,|\, \psi^{0..K})}. \tag{35}$$

Note that it is just the value of the evidence lower bound with the approximate posterior $q_\phi^K(z \,|\, \psi^{0..K})$, averaged over all values of $\psi^{0..K}$. We can also use that $\mathbb{E}_{\psi^{0..K}} q_\phi^K(z \,|\, \psi^{0..K}) = q_\phi(z)$ to rewrite the true ELBO in the same expectations:

$$\mathcal{L} = \mathbb{E}_{q_\phi(z)} \log \frac{p(x, z)}{q_\phi(z)} = \tag{36}$$

$$= \mathbb{E}_{\psi^{0..K}} \mathbb{E}_{z \,|\, \psi^{0..K}} \log \frac{p(x, z)}{q_\phi(z)}. \tag{37}$$

We want to prove that $\mathcal{L} \geq \underline{\mathcal{L}}_K^q$. Consider their difference $\mathcal{L} - \underline{\mathcal{L}}_K^q$:

$$\mathcal{L} - \underline{\mathcal{L}}_K^q = \tag{38}$$

$$= \mathbb{E}_{\psi^{0..K}} \mathbb{E}_{z \,|\, \psi^{0..K}} \log \frac{q_\phi^K(z \,|\, \psi^{0..K})}{q_\phi(z)} = \tag{39}$$

$$= \mathbb{E}_{\psi^{0..K}} \mathrm{KL}\left(q_\phi^K(z \,|\, \psi^{0..K}) \,\|\, q_\phi(z)\right) \geq 0. \tag{40}$$

We can use the same trick to prove that this bound is non-decreasing in K. First, let's use the symmetry in the indices once again, and rewrite $\underline{\mathcal{L}}_K^q$ and $\underline{\mathcal{L}}_{K+1}^q$ in the same expectations:

$$\underline{\mathcal{L}}_K^q = \mathbb{E}_{\psi^{0..K}} \mathbb{E}_{z \,|\, \psi^{0..K}} \log \frac{p(x, z)}{q_\phi^K(z \,|\, \psi^{0..K})} = \tag{41}$$

$$= \mathbb{E}_{\psi^{0..K+1}} \mathbb{E}_{z \,|\, \psi^{0..K}} \log \frac{p(x, z)}{q_\phi^K(z \,|\, \psi^{0..K})}, \tag{42}$$

$$\underline{\mathcal{L}}_{K+1}^q = \mathbb{E}_{\psi^{0..K+1}} \mathbb{E}_{z \,|\, \psi^0} \log \frac{p(x, z)}{q_\phi^{K+1}(z \,|\, \psi^{0..K+1})} = \tag{43}$$

$$= \mathbb{E}_{\psi^{0..K+1}} \mathbb{E}_{z \,|\, \psi^{0..K}} \log \frac{p(x, z)}{q_\phi^{K+1}(z \,|\, \psi^{0..K+1})}. \tag{44}$$

Then their difference would be equal to the expected KL-divergence, hence being non-negative:

$$\underline{\mathcal{L}}_{K+1}^q - \underline{\mathcal{L}}_K^q = \tag{45}$$

$$= \mathbb{E}_{\psi^{0..K+1}} \mathbb{E}_{z \,|\, \psi^{0..K}} \log \frac{q_\phi^K(z \,|\, \psi^{0..K})}{q_\phi^{K+1}(z \,|\, \psi^{0..K+1})} = \tag{46}$$

$$= \mathbb{E}_{\psi^{0..K+1}} \mathrm{KL}\left(q_\phi^K(z \,|\, \psi^{0..K}) \,\|\, q_\phi^{K+1}(z \,|\, \psi^{0..K+1}))\right)$$

$$\geq 0.$$

$\square$

B Importance Weighted Doubly Semi-Implicit VAE

The standard importance-weighted lower bound for VAE is defined as follows:

$$\log p(x) \geq \mathcal{L}^S = \mathbb{E}_{z^{1..S} \sim q_\phi(z)} \log \frac{1}{S} \sum_{i=1}^S \frac{p(x \,|\, z^i)p(z^i)}{q_\phi(z_i \,|\, x)} \tag{47}$$

We propose IW-DSIVAE, a new lower bound on the IWAE objective, that is suitable for VAEs with semi-implicit priors and posteriors:

$$\underline{\mathcal{L}}_{K_1, K_2}^{q, p, S} = \mathbb{E}_{\psi^{1..K_1} \sim q_\phi(\psi)} \mathbb{E}_{\zeta^{1..K_2} \sim p_\theta(\zeta)} \Bigg[$$

$$\mathbb{E}_{(z^1, \hat{\psi}^1), \dots, (z^S, \hat{\psi}^S) \sim q_\phi(z, \psi)} \Bigg[$$

$$\log \frac{1}{S} \sum_{i=1}^S \frac{p(x \,|\, z^i) \frac{1}{K_2} \sum_{k=1}^{K_2} p_\theta(z^i \,|\, \zeta^k)}{\frac{1}{K_1 + 1}(q_\phi(z^i \,|\, \hat{\psi}^i) + \sum_{k=1}^{K_1} q_\phi(z^i \,|\, \psi^k))} \Bigg] \Bigg]. \tag{48}$$

This objective is a lower bound on the IWAE objective ($\underline{\mathcal{L}}_{K_1, K_2}^{q, p, S} \leq \mathcal{L}^S$), is non-decreasing in both K_1 and K_2, and is asymptotically exact ($\underline{\mathcal{L}}_{\infty, \infty}^{q, p, S} = \mathcal{L}^S$).

C   Variational inference with hierarchical priors

Theorem 2. *Consider two different variational objectives* $\mathcal{L}^{joint}$ *and* $\mathcal{L}^{marginal}$. *Then*

$$\mathcal{L}^{joint}(\phi) = \mathbb{E}_{q_\phi(w,\alpha)} \log \frac{p(t\,|\,x,w)p(w\,|\,\alpha)p(\alpha)}{q_\phi(w,\alpha)} \quad (49)$$

$$\mathcal{L}^{marginal}(\phi) = \mathbb{E}_{q_\phi(w)} \log \frac{p(t\,|\,x,w)p(w)}{q_\phi(w)} \quad (50)$$

Let ϕ_j *and* ϕ_m *maximize* $\mathcal{L}^{joint}$ *and* $\mathcal{L}^{marginal}$ *correspondingly. Then* $q_{\phi_m}(w)$ *is a better fit for the marginal posterior that* $q_{\phi_j}(w)$ *in terms of the KL-divergence:*

$$\mathrm{KL}(q_{\phi_m}(w)\,\|\,p(w\,|\,X_{tr}, T_{tr})) \leq$$
$$\mathrm{KL}(\,q_{\phi_j}(w)\,\|\,p(w\,|\,X_{tr}, T_{tr})) \quad (51)$$

Proof. Note that maximizing $\mathcal{L}^{marginal}(\phi)$ directly minimizes $\mathrm{KL}(q_\phi(w)\,\|\,p(w\,|\,X_{tr}, T_{tr}))$, as $\mathcal{L}^{marginal}(\phi) + \mathrm{KL}(q_\phi(w)\,\|\,p(w\,|\,X_{tr}, T_{tr})) = const.$ The sought-for inequality (51) then immediately follows from $\mathcal{L}^{marginal}(\phi_m) \geq \mathcal{L}^{marginal}(\phi_j)$. $\quad\square$

To see the cause of this inequality more clearly, consider $\mathcal{L}^{joint}(\phi)$:

$$\mathcal{L}^{joint}(\phi) = \mathbb{E}_{q_\phi(w,\alpha)} \log \frac{p(t\,|\,x,w)p(w\,|\,\alpha)p(\alpha)}{q_\phi(w,\alpha)} = \tag{52}$$

$$= \mathbb{E}_{q_\phi(w)} \log p(t\,|\,x,w) - \mathrm{KL}(q_\phi(w,\alpha)\,\|\,p(w,\alpha)) = \tag{53}$$

$$= \mathbb{E}_{q_\phi(w)} \log p(t\,|\,x,w) - \mathrm{KL}(q_\phi(w)\,\|\,p(w)) - \tag{54}$$

$$- \mathbb{E}_{q_\phi(w)}\mathrm{KL}(q_\phi(\alpha\,|\,w)\,\|\,p(\alpha\,|\,w)) = \tag{55}$$

$$= \mathcal{L}^{marginal}(\phi) - \mathbb{E}_{q_\phi(w)}\mathrm{KL}(q_\phi(\alpha\,|\,w)\,\|\,p(\alpha\,|\,w)) \tag{56}$$

If $\mathcal{L}^{joint}$ and $\mathcal{L}^{marginal}$ coincide, the inequality (51) becomes an equality. However, $\mathcal{L}^{joint}$ and $\mathcal{L}^{marginal}$ only coincide if the reverse posterior $q_\phi(\alpha\,|\,w)$ is an exact match for the reverse prior $p(\alpha\,|\,w)$. Due to the limitations of the approximation family of the joint posterior, this is not the case in many practical applications. In many cases [7, 18] the joint approximate posterior is modeled as a factorized distribution $q_\phi(w,\alpha) = q_\phi(w)q_\phi(\alpha)$. Therefore in the case of the joint variational inference, we optimize a lower bound on the marginal ELBO and therefore obtain a sub-optimal approximation.

Table 2: The values of the marginal ELBO, the train negative log-likelihood, the KL-divergence between the marginal posterior $q_\phi(w)$ and the marginal prior $p_\phi(w)$, and the test-set accuracy and negative log-likelihood for different inference procedures for a model with a standard Student's prior. The predictive distribution during test-time was estimated using 200 samples from the marginal posterior $q_\phi(w)$

Method	Train			Test	
	ELBO	NLL	KL	Acc.	NLL
Marginal	$-\mathbf{1.42 \times 10^5}$	7.2×10^3	1.35×10^5	97.80	855
Joint	-1.48×10^5	6.7×10^3	1.42×10^5	97.74	831
DSIVI(K=2)	-1.47×10^5	7.0×10^3	1.41×10^5	97.75	846
DSIVI(K=10)	$-\mathbf{1.42 \times 10^5}$	7.2×10^3	1.35×10^5	97.76	843

D   Toy data for sequential approximation

For sequential approximation toy task, we follow [40] and use the following target distributions. For one-dimensional Gaussian mixture, $p(z) = 0.3\mathcal{N}(z\,|\,-2, 1) + 0.7\mathcal{N}(z\,|\,2, 1)$. For the "banana" distribution, $p(z_1, z_2) = \mathcal{N}(z_1\,|\,z_2^2/4, 1)\mathcal{N}(z_2\,|\,0, 4)$.

For both target distributions, we optimize the objective using Adam optimizer with initial learning rate 10^{-2} and decaying it by 0.5 every 500 steps. On each iteration of sequential approximation, we train for 5000 steps. We reinitialize all trainable parameters and optimizer statistics before each iteration. Before each update of the parameters, we average 200 Monte Carlo samples of the gradients. During evaluation, we used 10^5 Monte Carlo samples to estimate the expectations involved in the lower and upper bounds on KL divergence.
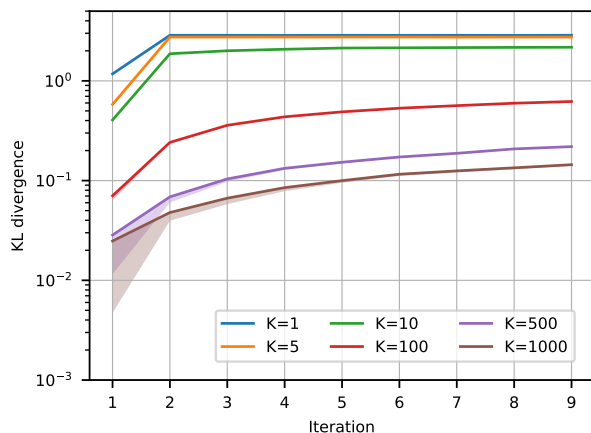

Figure 5: Sequential approximation. Area is shaded between lower and upper bounds of $\mathrm{KL}(q_{\phi_i}(z)\,\|\,p(z))$ for different *training* values of $K_1 = K_2 = K$, and the solid lines represent the corresponding upper bounds. During *evaluation*, $K = 10^4$ is used. Here $p(z)$ is a two-dimensional "banana". Lower is better.

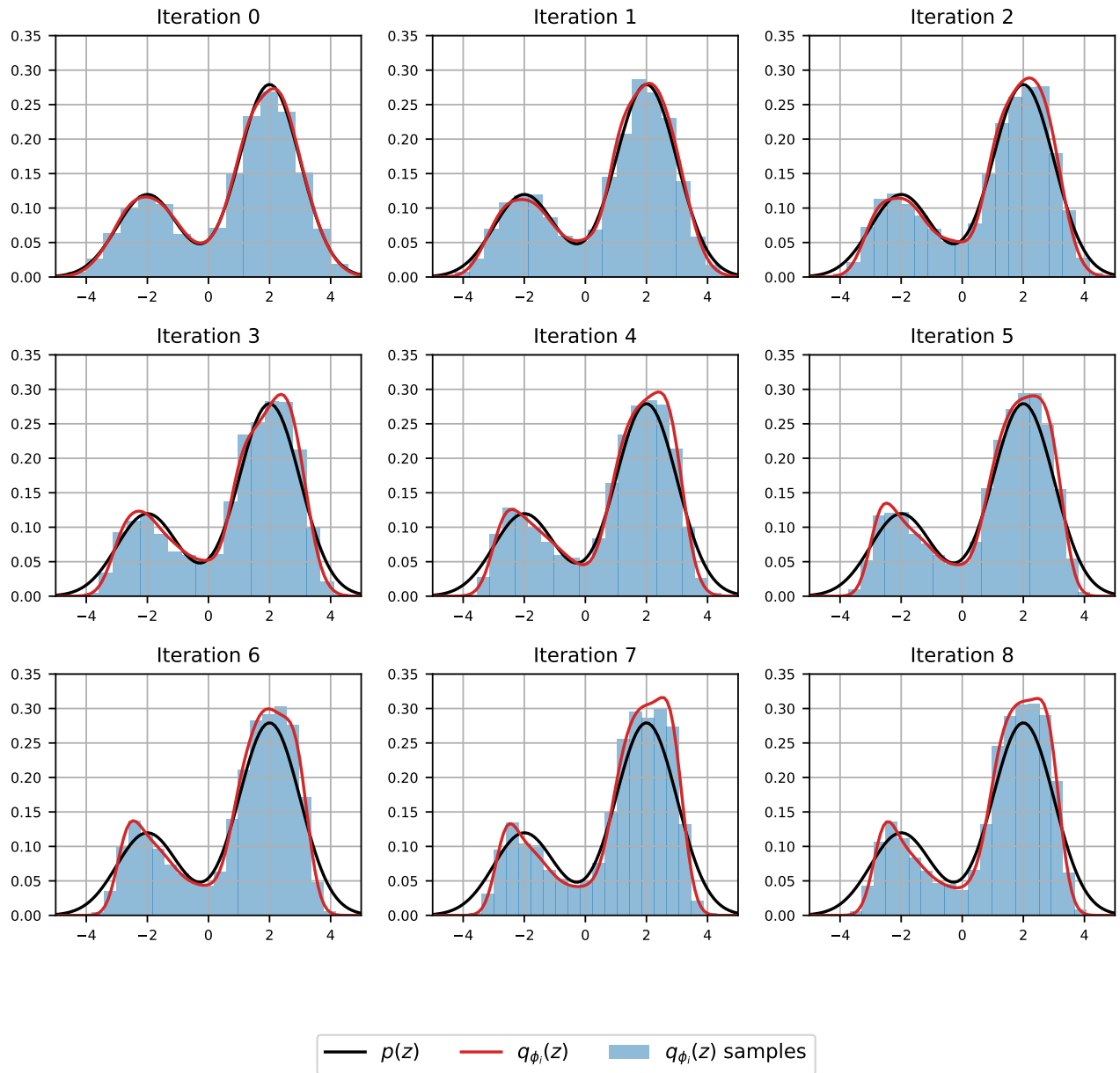Dmitry Molchanov, Valery Kharitonov, Artem Sobolev, Dmitry Vetrov

Figure 6: Learned distributions after each iteration for Gaussian mixture target distribution, $K = 100$ during training.
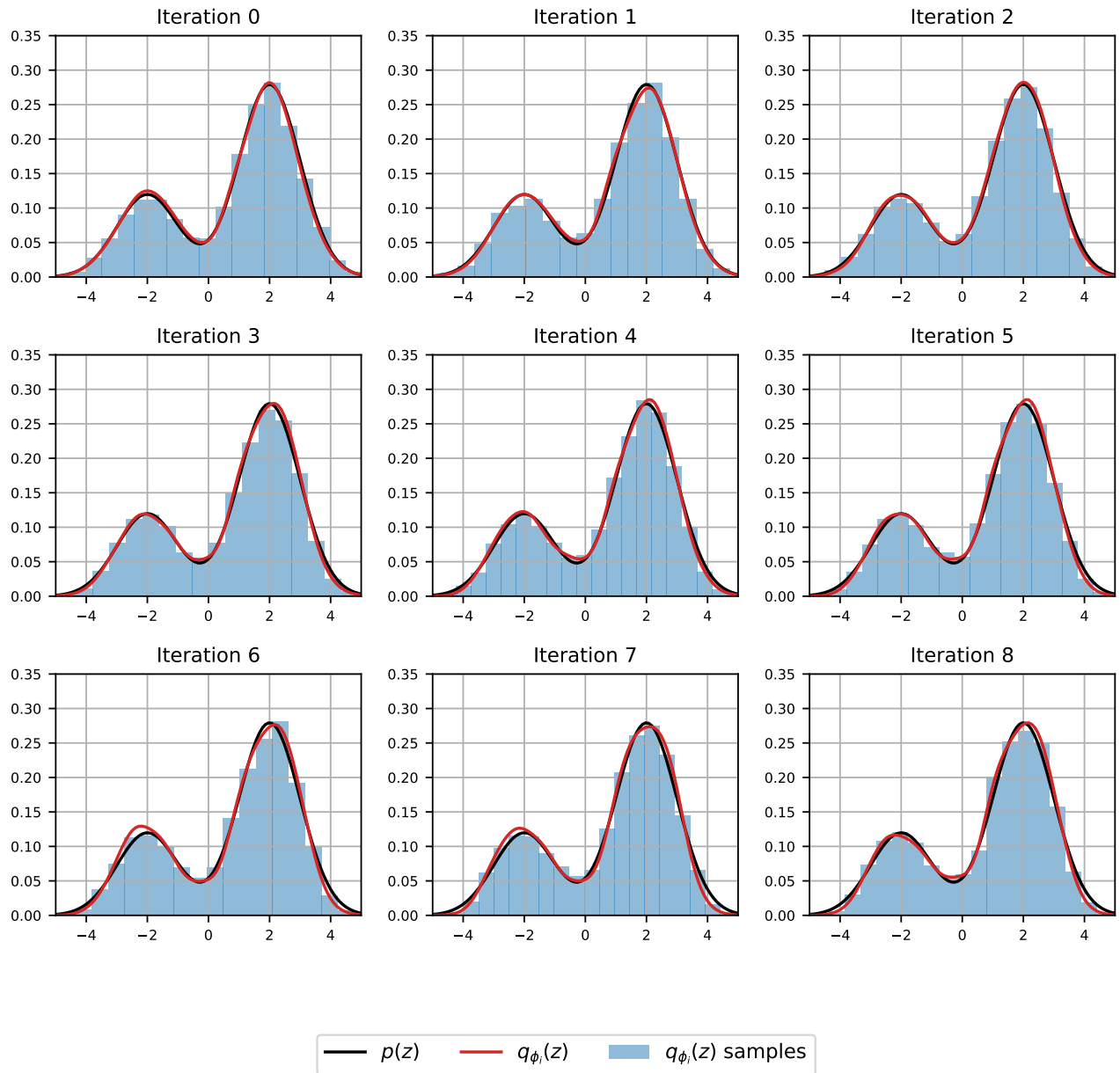
Figure 7: Learned distributions after each iteration for Gaussian mixture target distribution, $K = 1000$ during training.
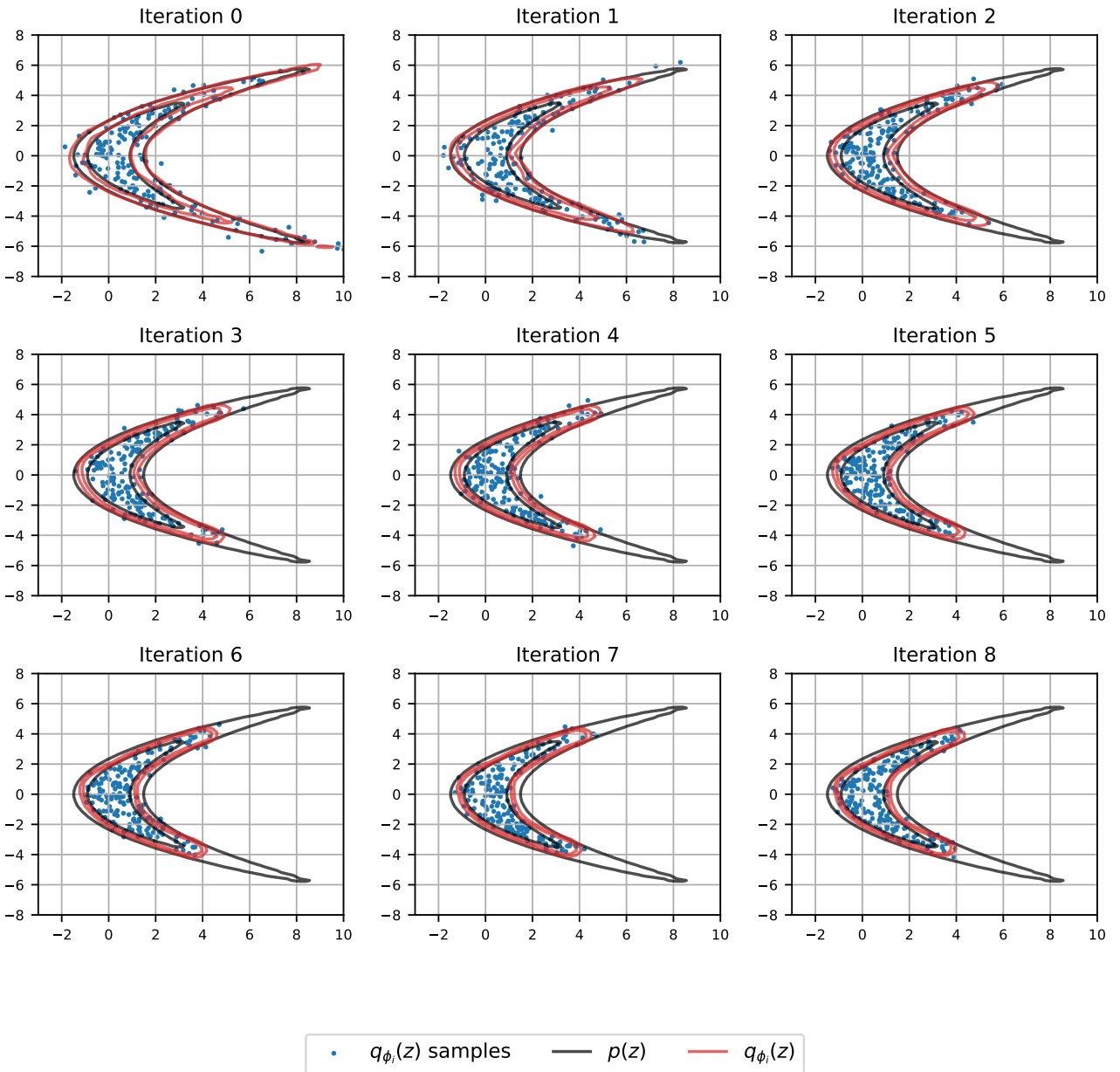
Figure 8: Learned distributions after each iteration for "banana" target distribution, $K = 100$ during training.
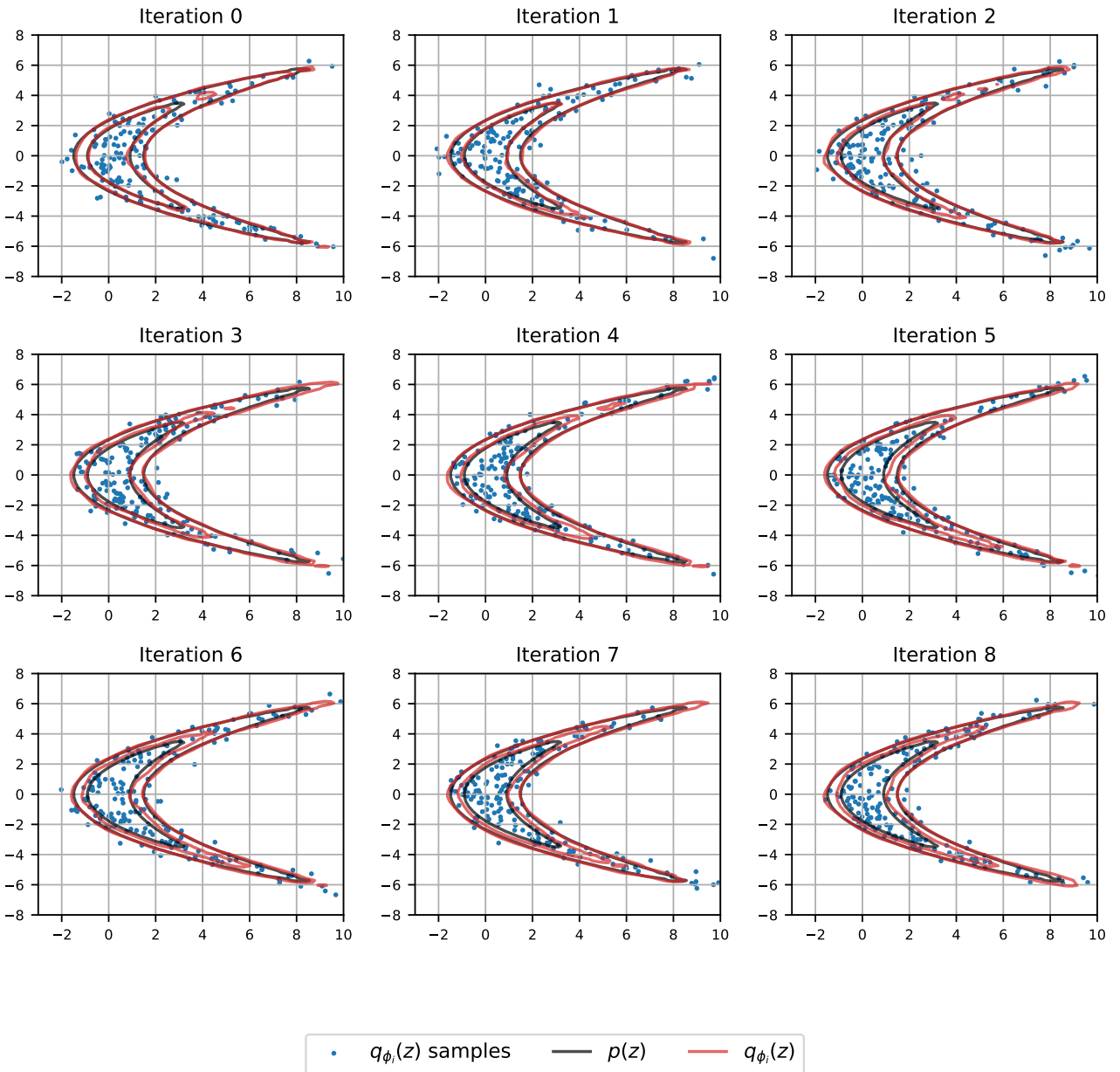
Figure 9: Learned distributions after each iteration for "banana" target distribution, $K = 1000$ during training.