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

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

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

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

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

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

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

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

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

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

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

Note that it is just the value of the evidence lower bound with the approximate posterior $q^K_\phi(z\,|\,\psi^{0..K})$, averaged over all values of $\psi^{0..K}$. We can also use that $\mathbb{E}_{\psi^{0..K}} q^K_\phi(z\,|\,\psi^{0..K}) = q_\phi(z)$ 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 $\mathcal{L} \ge \underline{\mathcal{L}}^q_K$. Consider their difference $\mathcal{L} - \underline{\mathcal{L}}^q_K$:

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

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

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

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 $\underline{\mathcal{L}}^q_K$ and $\underline{\mathcal{L}}^q_{K+1}$ in the same expectations:

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

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

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

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

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

$$\underline{\mathcal{L}}^q_{K+1} - \underline{\mathcal{L}}^q_K = \tag{45}$$

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

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

$$\ge 0.$$

□

B Importance Weighted Doubly Semi-Implicit VAE

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

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

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

$$\underline{\underline{\mathcal{L}}}^{q,p,S}_{K_1,K_2} = \mathbb{E}_{\psi^{1..K_1}\sim q_\phi(\psi)}\mathbb{E}_{\zeta^{1..K_2}\sim p_\theta(\zeta)}\Bigg[\mathbb{E}_{(z^1,\hat\psi^1),\dots,(z^S,\hat\psi^S)\sim q_\phi(z,\psi)}\Bigg[\log\frac{1}{S}\sum_{i=1}^{S}\frac{p(x\,|\,z^i)\frac{1}{K_2}\sum_{k=1}^{K_2}p_\theta(z^i\,|\,\zeta^k)}{\frac{1}{K_1+1}(q_\phi(z^i\,|\,\hat\psi^i)+\sum_{k=1}^{K_1}q_\phi(z^i\,|\,\psi^k))}\Bigg]\Bigg]. \tag{48}$$

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

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 $\mathcal{L}^{marginal}(\phi_m) \geq \mathcal{L}^{marginal}(\phi_j)$. □

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

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

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

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

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

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

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 $q_\phi(w)$

	Train			Test	
Method	**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.

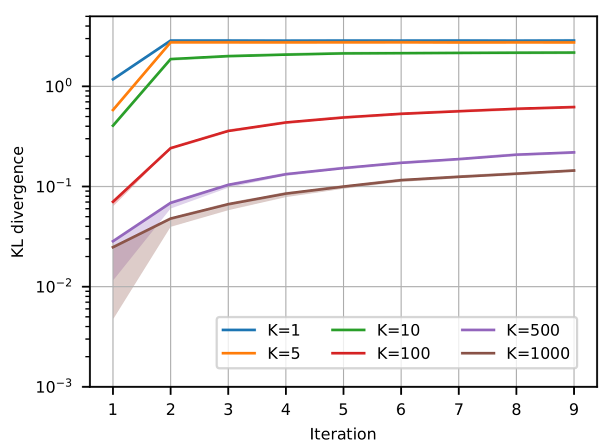

Figure 5: Sequential approximation. Area is shaded between lower and upper bounds of $\mathrm{KL}(q_{\phi_i}(z)\,\|\,p(z))$ for different *training* values of $K_1 = K_2 = K$, and the solid lines represent the corresponding upper bounds. During *evaluation*, $K = 10^4$ is used. Here $p(z)$ is a two-dimensional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

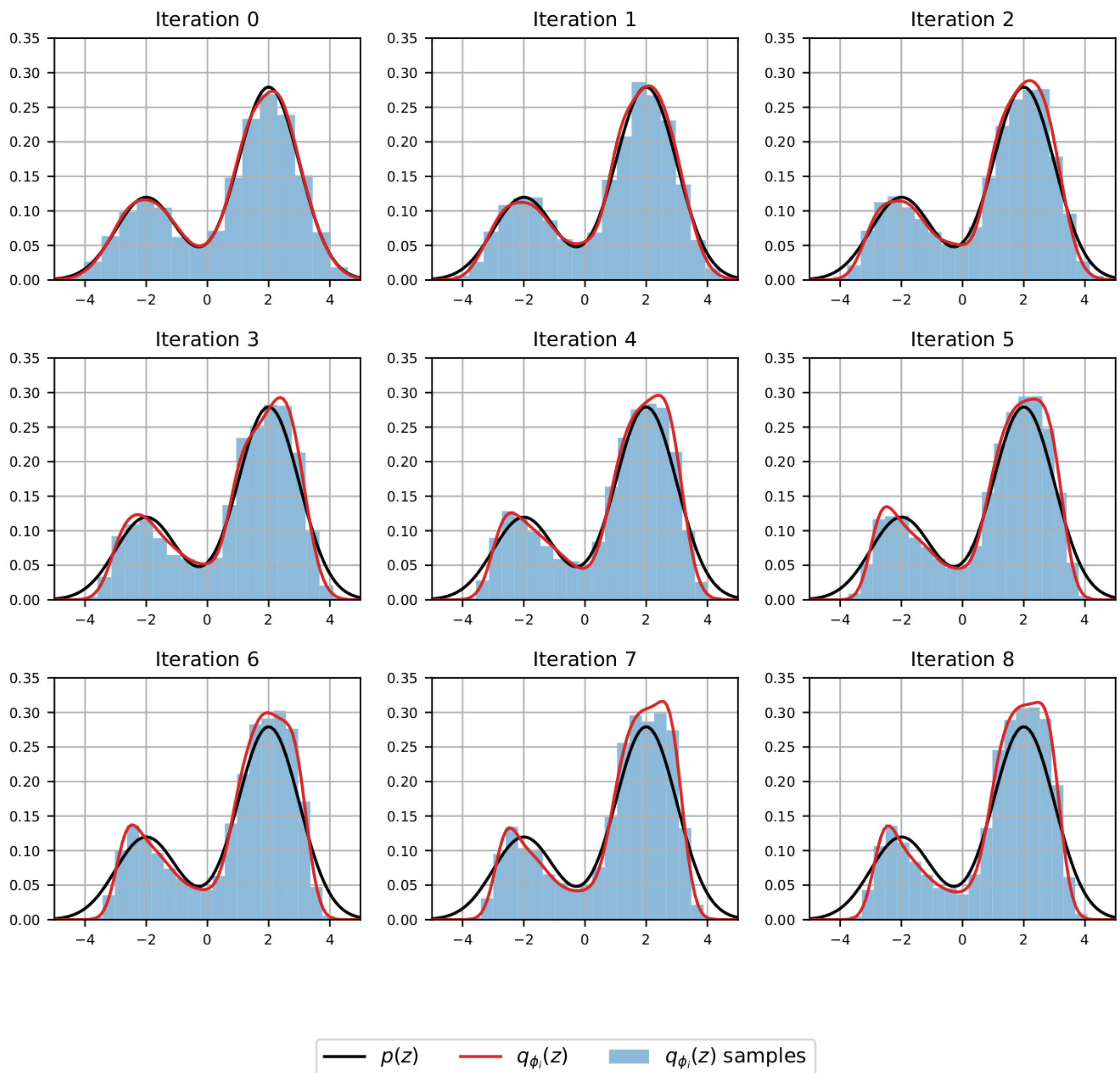

Figure 6: Learned distributions after each iteration for Gaussian mixture target distribution, $K = 100$ during training.

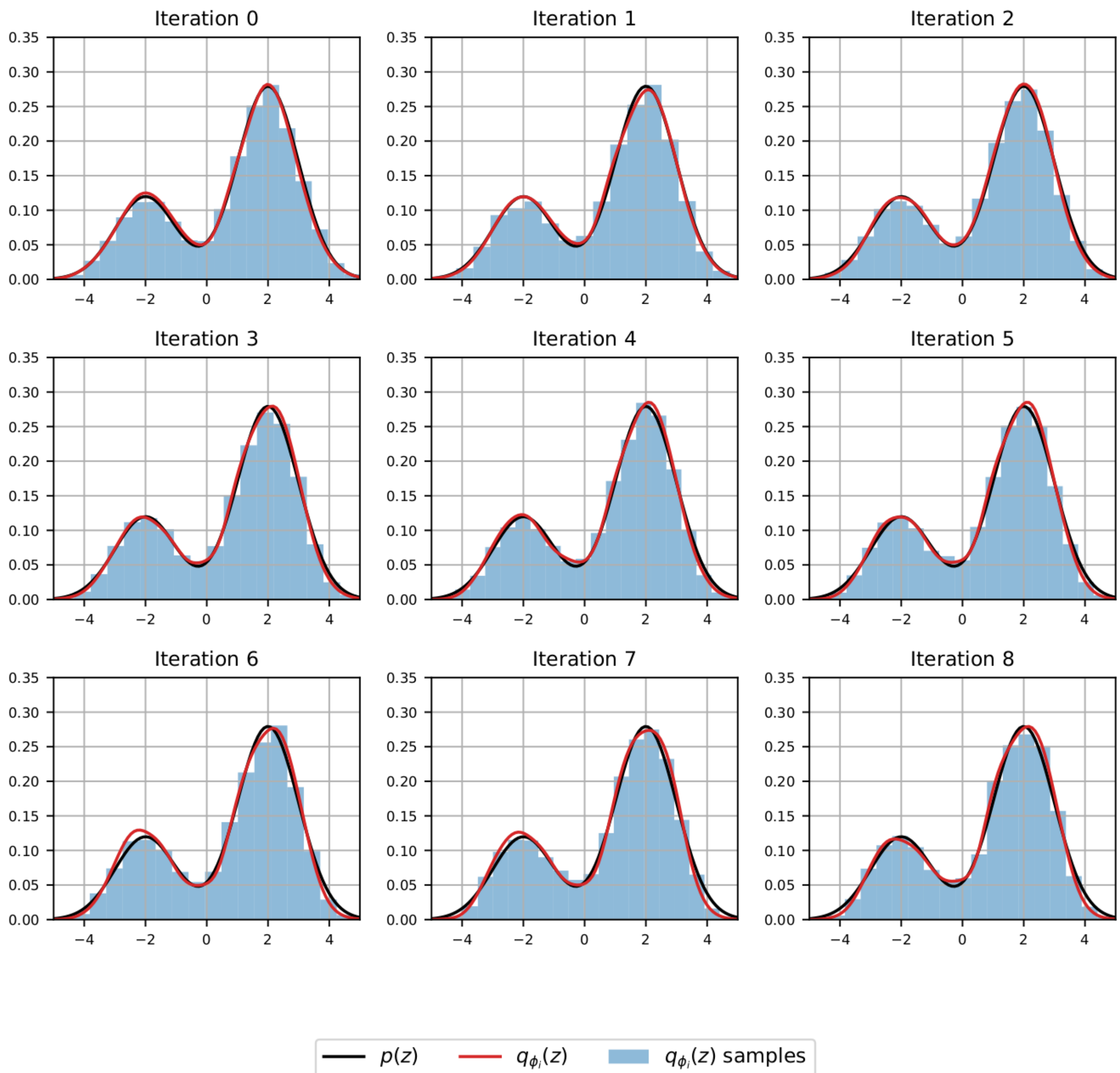

Figure 7: Learned distributions after each iteration for Gaussian mixture target distribution, $K = 1000$ during training.

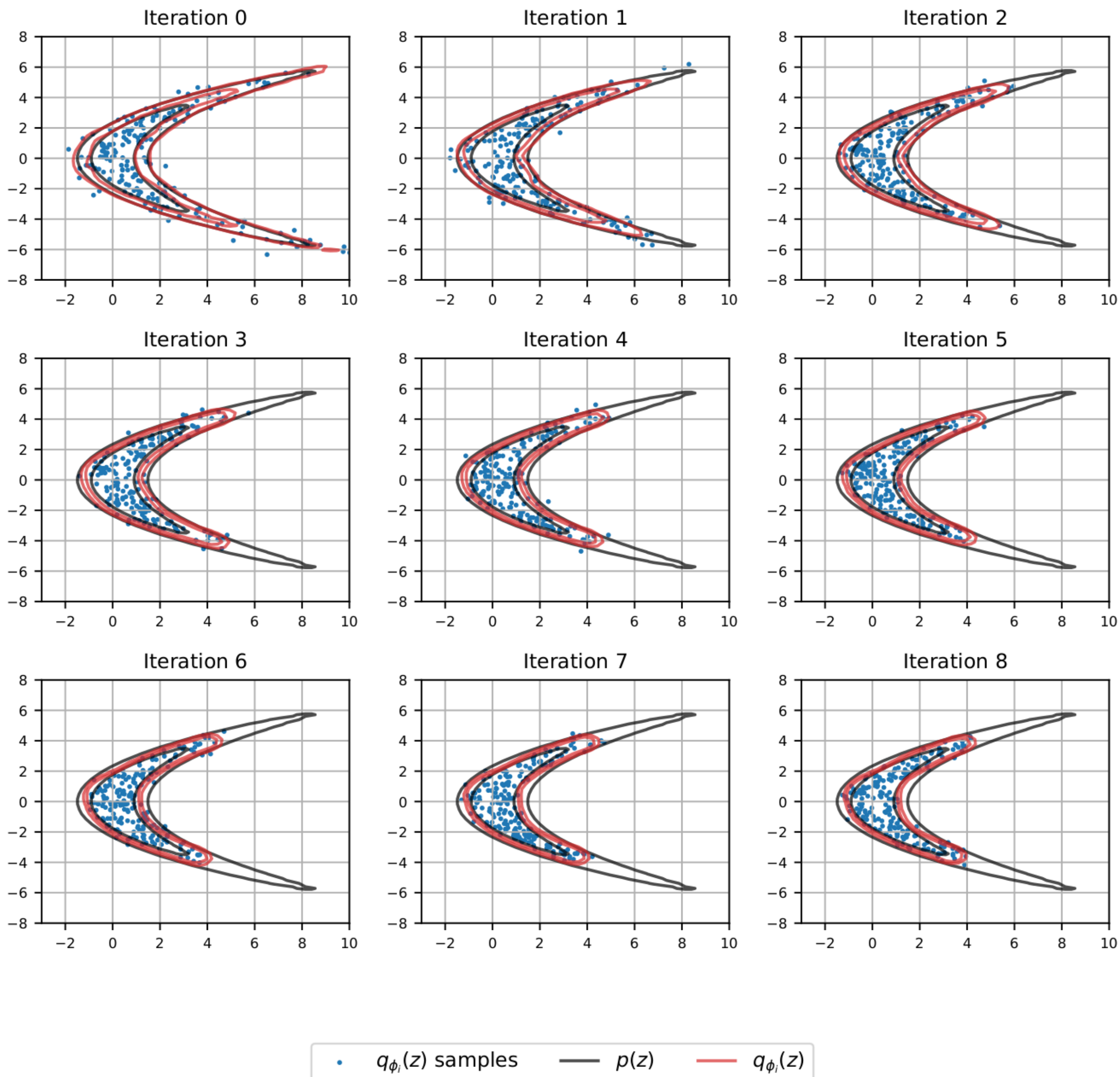

Figure 8: Learned distributions after each iteration for "banana" target distribution, $K = 100$ during training.

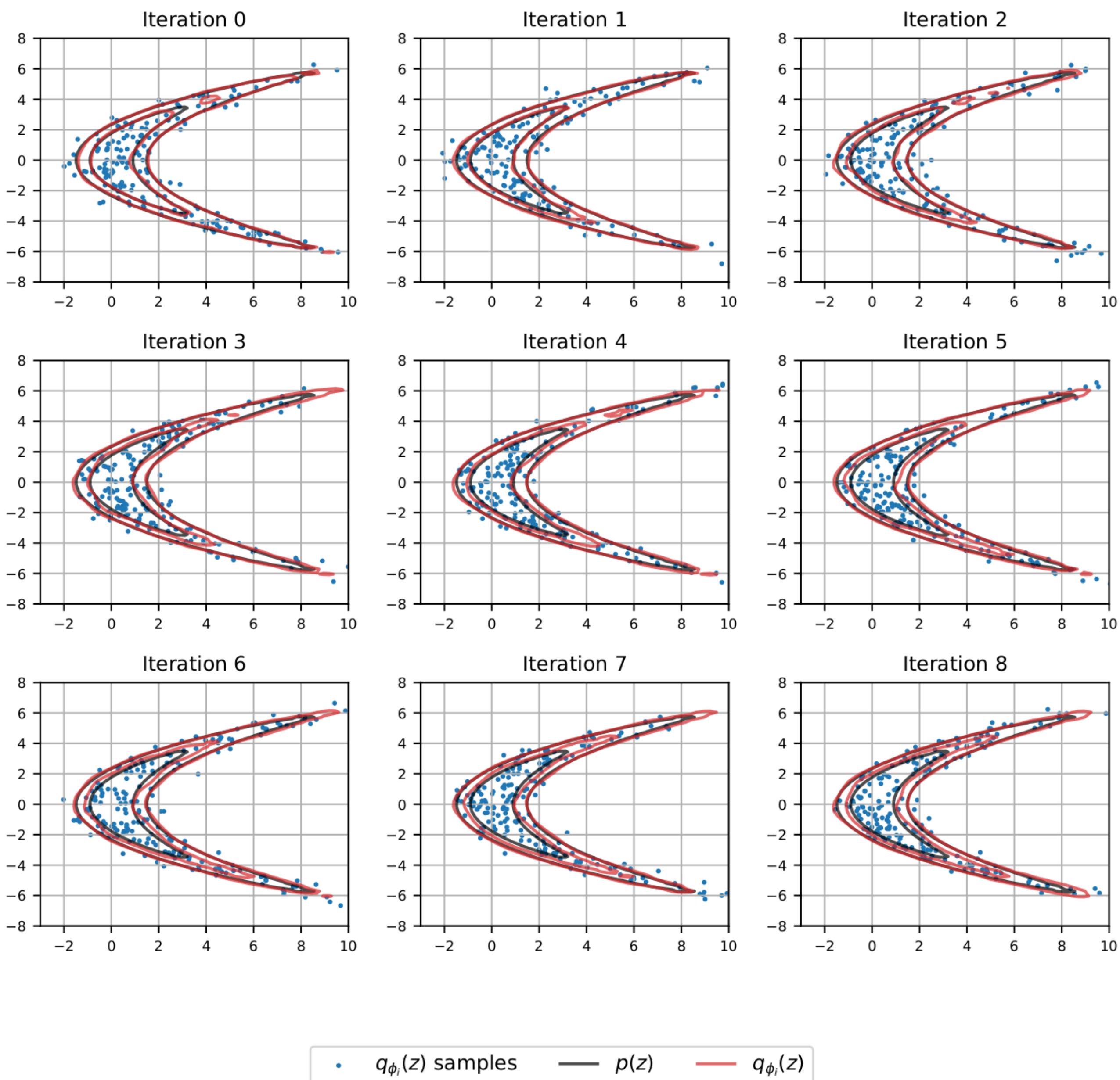

Figure 9: Learned distributions after each iteration for "banana" target distribution, $K = 1000$ during training.