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

Theorem 1. Consider $\mathcal{L}$ and $\mathcal{L}_K^q$ defined as in Eq. (2) and (3). Then $\mathcal{L}_K^q$ converges to $\mathcal{L}$ from below as $K \to \infty$, satisfying $\mathcal{L}_K^q \leq \mathcal{L}$, and

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

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

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

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

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 $\mathcal{L}_K^q$ and $\mathcal{L}_K^{q+1}$ in the same expectations:

$$
\mathcal{L}_K^q = \mathbb{E}_{\psi^{0..K}} \log p(x | z)_q^{K}(z | \psi^{0..K}),
$$

$$
\mathcal{L}_K^{q+1} = \mathbb{E}_{\psi^{0..K+1}} \log p(x | z)_q^{K+1}(z | \psi^{0..K+1}).
$$

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

$$
\mathcal{L}_K^{q+1} - \mathcal{L}_K^q = \mathbb{E}_{\psi^{0..K+1}} \log q^{K+1}_\phi(z | \psi^{0..K+1}) - q^K_\phi(z | \psi^{0..K}).
$$

B Importance Weighted Doubly Semi-Implicit VAE

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

$$
\log p(x) \geq \mathcal{L}^S = \mathbb{E}_{z^{1..S} \sim q(s)} \log \frac{1}{S} \sum_{i=1}^{S} p(x | z_i) q(z_i | x).
$$

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

$$
\mathcal{L}^{S,p,S}_{K_1,K_2} = \mathbb{E}_{\psi^{1..K_1} \sim q_\psi(\psi)} \mathbb{E}_{\zeta^{1..K_2} \sim p_\zeta(\zeta)} \log \frac{1}{\sum_{i=1}^{S} \prod_{j=1}^{K_1} q(z_i | s_j) q(z_i | \psi^j)}.
$$

This objective is a lower bound on the IWVAE objective ($\mathcal{L}^{S,p,S}_{K_1,K_2} \leq \mathcal{L}^S$), is non-decreasing in both $K_1$ and $K_2$, and is asymptotically exact ($\mathcal{L}^{S,p,S}_{\infty,\infty} = \mathcal{L}^S$).
C  Variational inference with hierarchical priors

Theorem 2. Consider two different variational objectives \( \mathcal{L}^{\text{joint}} \) and \( \mathcal{L}^{\text{marginal}} \). Then

\[
\mathcal{L}^{\text{joint}}(\phi) = \mathbb{E}_{q_\phi(w, \alpha)} \log \frac{p(t \mid x, w)p(w \mid \alpha)p(\alpha)}{q_\phi(w, \alpha)} \tag{49}
\]

\[
\mathcal{L}^{\text{marginal}}(\phi) = \mathbb{E}_{q_\phi(w)} \log \frac{p(t \mid x, w)p(w)}{q_\phi(w)} \tag{50}
\]

Let \( \phi_j \) and \( \phi_m \) maximize \( \mathcal{L}^{\text{joint}} \) and \( \mathcal{L}^{\text{marginal}} \) 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:

\[
\text{KL}(q_{\phi_m}(w) \parallel p(w \mid X_{tr}, T_{tr})) \leq \text{KL}(q_{\phi_j}(w) \parallel p(w \mid X_{tr}, T_{tr})) \tag{51}
\]

Proof. Note that maximizing \( \mathcal{L}^{\text{marginal}}(\phi) \) directly minimizes \( \text{KL}(q_\phi(w) \parallel p(w \mid X_{tr}, T_{tr})) \), and \( \mathcal{L}^{\text{marginal}}(\phi) + \text{KL}(q_\phi(w) \parallel p(w \mid X_{tr}, T_{tr})) = \text{const.} \)

The sought-for inequality \( \tag{51} \) then immediately follows from \( \mathcal{L}^{\text{marginal}}(\phi_m) \geq \mathcal{L}^{\text{marginal}}(\phi_j) \). □

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

\[
\mathcal{L}^{\text{joint}}(\phi) = \mathbb{E}_{q_\phi(w, \alpha)} \log \frac{p(t \mid x, w)p(w \mid \alpha)p(\alpha)}{q_\phi(w, \alpha)} =
\]

\[
= \mathbb{E}_{q_\phi(w)} \log p(t \mid x, w) - \text{KL}(q_\phi(w, \alpha) \parallel p(w, \alpha)) =
\]

\[
= \mathbb{E}_{q_\phi(w)} \log p(t \mid x, w) - \text{KL}(q_\phi(w) \parallel p(w)) - \mathbb{E}_{q_\phi(w)} \text{KL}(q_\phi(\alpha \mid w) \parallel p(\alpha \mid w)) =
\]

\[
= \mathcal{L}^{\text{marginal}}(\phi) - \mathbb{E}_{q_\phi(w)} \text{KL}(q_\phi(\alpha \mid w) \parallel p(\alpha \mid w))
\]

If \( \mathcal{L}^{\text{joint}} \) and \( \mathcal{L}^{\text{marginal}} \) coincide, the inequality \( \tag{51} \) becomes an equality. However, \( \mathcal{L}^{\text{joint}} \) and \( \mathcal{L}^{\text{marginal}} \) only coincide if the reverse posterior \( q_\phi(\alpha \mid w) \) is an exact match for the reverse prior \( p(\alpha \mid w) \). Due to the limitations of the approximation family of the joint posterior, this is not the case in many practical applications. In many cases \( \textit{7, 15} \) 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 suboptimal 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 \( q_\phi(w) \)

<table>
<thead>
<tr>
<th>Method</th>
<th>ELBO</th>
<th>NLL</th>
<th>KL</th>
<th>Acc. NLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marginal</td>
<td>-1.42×10^9</td>
<td>7.2×10^3</td>
<td>1.35×10^6</td>
<td>97.80</td>
</tr>
<tr>
<td>Joint</td>
<td>-1.47×10^5</td>
<td>6.7×10^3</td>
<td>1.42×10^6</td>
<td>97.74</td>
</tr>
<tr>
<td>DSIVI(K=2)</td>
<td>-1.47×10^5</td>
<td>7.0×10^3</td>
<td>1.41×10^6</td>
<td>97.75</td>
</tr>
<tr>
<td>DSIVI(K=10)</td>
<td>-1.42×10^6</td>
<td>7.2×10^3</td>
<td>1.35×10^6</td>
<td>97.76</td>
</tr>
</tbody>
</table>

D  Toy data for sequential approximation

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

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 KL(q_\phi(z) \parallel p(z)) for different training values of K. For different training values of K, the solid lines represent the corresponding upper bounds. During evaluation, K = 10^5 is used. Here p(z) is a two-dimensional "banana". Lower is better.](image-url)
Figure 6: Learned distributions after each iteration for Gaussian mixture target distribution, $K = 100$ during training.
Doubly Semi-Implicit Variational Inference

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.