IWAE vs. VAE

Tighter Bounds, Richer Latents?

Cedric Damais, Yacine Benihaddadene, Amine Mike El Maalouf, Leon Ayral

GAIDM
EPITA

January 19, 2026

Standard VAE: The Objective

Goal: Maximize marginal log-likelihood $\log p(x)$.

Since $p(x)$ is intractable, VAE ($K = 1$) maximizes the **ELBO**:

$$\log p(x) \geq \mathcal{L}_{\text{VAE}} = \mathbb{E}_{z \sim q}\left[\log \frac{p(x,z)}{q(z|x)}\right]$$

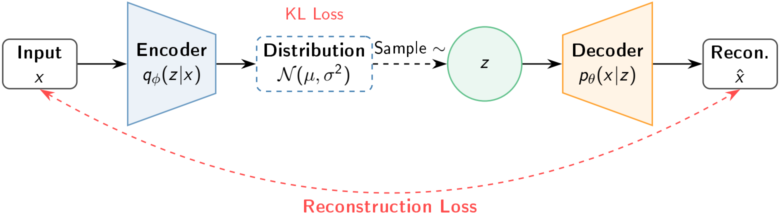
The Problem: The "Gap"

The bound is strictly lower than the evidence due to the KL divergence:

$$\log p(x) - \mathcal{L}_{\text{VAE}} = \text{KL}(q(z|x)||p(z|x))$$

- If $q(z|x)$ is too simple $\rightarrow$ **Loose Bound**.
- Loose bound $\rightarrow$ Risk of **Posterior Collapse**.

Model Architecture



The IWAE Solution: Importance Sampling

Idea

Use K samples to tighten the bound (Burda et al., 2015).

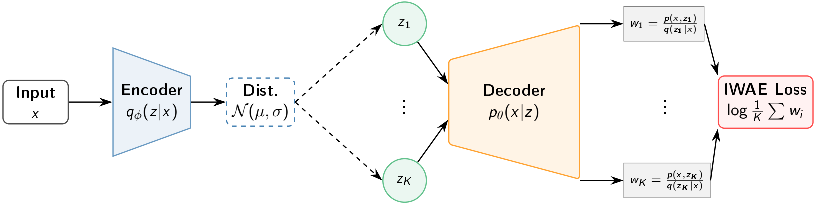
$$\mathcal{L}_K(x) = \mathbb{E}_{z_{1\dots K}\sim q}\left[\log\left(\frac{1}{K}\sum_{i=1}^{K}\frac{p(x,z_i)}{q(z_i|x)}\right)\right]$$

Key Property (Monotonicity):

$$\mathcal{L}_1 \leq \mathcal{L}_5 \leq \mathcal{L}_{20} \leq \cdots \leq \log p(x)$$

- **K=1**: Standard VAE.
- **K $>$ 1**: Tighter approximation of true probability.

IWAE Architecture



Experimental Setup

- **Objective:** Compare VAE ($K = 1$) vs IWAE ($K > 1$).
- **Constraint: Identical Architecture** for fair comparison.

Parameter	**Value**
Dataset	MNIST (Binarized)
Encoder	MLP ($784 \to 400 \to 20$)
Decoder	MLP ($20 \to 400 \to 784$)
Optimizer	Adam ($lr = 1e - 3$)
K (Samples)	**1, 5, 20**

Implementation Detail

Used `torch.logsumexp` to avoid numerical underflow.

Result 1: Log-Likelihood Estimation

- **Metric:** Negative Log-Likelihood (Lower is better) or ELBO (Higher is better).

Figure 1: Estimated Log-Likelihood over Epochs

Observation: $\mathcal{L}_{20} > \mathcal{L}_5 > \mathcal{L}_1$. Increasing K strictly tightens the bound.

Result 2: Latent Utilization

- **Metric:** Active Units (Dimensions where $KL > \epsilon$).

Figure 2: KL Divergence per Latent Dimension

Observation: IWAE ($K = 20$) utilizes more of the latent space, reducing the "Posterior Collapse" often seen in standard VAEs.

Result 3: Sample Quality

VAE (K=1)

IWAE (K=20)

IWAE samples typically show sharper strokes and fewer "averaged" blurry digits.

Trade-offs Summary

Metric	VAE (K=1)	IWAE (K=20)
Bound Tightness	Loose	**Tight**
Latent Usage	Risk of Collapse	**Rich**
Gradient Variance	High	**Low**
Compute Cost	**Low**	High ($\times K$)

Thanks!

Q & A