IWAE vs. VAE
Tighter Bounds, Richer Latents?

Cedric Damais, Yacine Benihaddadene, Amine Mike El Maalouf,
Leon Ayral, Oscar Le Dauphin

GAIDM
EPITA

January 20, 2026

Introduction & Theory
0000

The IWAE Solution
000000

Methodology
00

Results
000000000

Conclusion
000

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

4 Results

5 Conclusion

Introduction & Theory
●○○○

The IWAE Solution
○○○○○○

Methodology
○○

Results
○○○○○○○○○

Conclusion
○○○

1. **Introduction & Theory**

2. The IWAE Solution

3. Methodology

4. Results

5. Conclusion

Introduction & Theory
○●○○

The IWAE Solution
○○○○○○

Methodology
○○

Results
○○○○○○○○○

Conclusion
○○○

The Goal: Latent Variable Inference

The Objective: Given data x, we want to learn the posterior distribution of the latent variables z:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

- $p(x|z)$: Likelihood (Decoder)
- $p(z)$: Prior (e.g., $\mathcal{N}(0, I)$)
- $p(x)$: **Marginal Likelihood (The Evidence)**

The Problem: Intractability

We can't simply use Bayes Formula because p(x) is intractable. To calculate the denominator $p(x)$, we must marginalize out z:

$$p(x) = \int p(x|z)p(z) \, dz$$

Introduction & Theory
○○○●

The IWAE Solution
○○○○○○

Methodology
○○

Results
○○○○○○○○○

Conclusion
○○○

Standard VAE: The Objective

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

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

$$\log p(x) \geq \mathcal{L}_{\mathsf{VAE}} = \mathbb{E}_{z \sim q}\left[\log \frac{p(x, z)}{q(z|x)}\right]$$
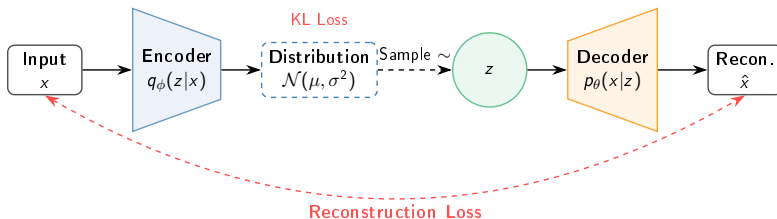
The Problem: The "Gap"

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

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

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

Introduction & Theory
oooo●

The IWAE Solution
oooooo

Methodology
oo

Results
ooooooooo

Conclusion
ooo

Model Architecture

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

4 Results

5 Conclusion

Novelty 1: Strictly Tighter Bounds

Theorem (Burda et al., 2015): The IWAE bound $\mathcal{L}_K$ is monotonically increasing with K.

$$\mathcal{L}_1 \leq \mathcal{L}_2 \leq \cdots \leq \mathcal{L}_K \leq \cdots \leq \log p(x)$$

Why?

Using **Jensen's Inequality** on the concave log function:

$$\mathbb{E}\left[\log\left(\frac{1}{K}\sum w_i\right)\right] \leq \log\left(\mathbb{E}\left[\frac{1}{K}\sum w_i\right]\right) = \log p(x)$$

Implication:
- As $K \to \infty$, the estimator converges to the true marginal likelihood $\log p(x)$.
- Even with a finite K, we are guaranteed a better objective

Novelty 2: Richer Implicit Posteriors

The IWAE Solution: Weighting as Filtering

Normalized Weight: $\quad \tilde{w}_i = \dfrac{w_i}{\sum_{j=1}^{k} w_j} \quad$ where $\quad w_i = \dfrac{p(x, z_i)}{q(z_i|x)}$

$$\underbrace{\tilde{q}(z|x)}_{\substack{\text{Implicit} \\ \text{Multi-modal}}} \approx \sum_{i=1}^{k} \tilde{w}_i \delta(z - z_i)$$

$$\nabla \mathcal{L}_k = \mathbb{E}_{\epsilon} \left[\sum_{i=1}^{k} \tilde{w}_i \nabla_{\theta} \log w(x, z_i, \theta) \right]$$

Why Sampling More Helps? (The Mechanics)

Case $K = 1$ (Risky)

If we draw just **one** bad sample z_1:

$$w_1 \approx 0$$

$$\log(w_1) \to -\infty$$

Gradient Explodes

Case $K > 1$ (Hedging)

If we draw many bad samples, but just **one good one**:

$$\log(0 + 0 + \cdots + w_{good})$$

$$\approx \log(w_{good}) > -\infty$$

Stable Training

The Insight: The sum acts as a safety net. The Encoder is allowed to make mistakes, as long as it gets it right *once*.

Introduction & Theory
○○○○

The IWAE Solution
○○○○○●○

Methodology
○○

Results
○○○○○○○○○

Conclusion
○○○

The Mechanism: Importance Weighting

1. The Definition

- Calculate raw weight:

$$w_i = \frac{p(x, z_i)}{q(z_i|x)}$$

- Normalize:

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^{k} w_j}$$

2. The Intuition

Sample z_i is **Good**
$\rightarrow p(x, z_i)$ is High
$\rightarrow \tilde{w}_i \approx 1$

Sample z_j is **Bad**
$\rightarrow p(x, z_j)$ is Low
$\rightarrow \tilde{w}_j \approx 0$

Key Result: The Filter Effect

The gradient update effectively **ignores** bad samples:

$$k$$

Introduction & Theory
oooo

The IWAE Solution
oooooo●

Methodology
oo

Results
ooooooooo

Conclusion
ooo

IWAE Architecture

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

4 Results

5 Conclusion

Experimental Setup

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

Parameter	Value
Dataset	MNIST (Binarized)
Encoder	MLP ($784 \rightarrow 400 \rightarrow 20$)
Decoder	MLP ($20 \rightarrow 400 \rightarrow 784$)
Optimizer	Adam ($lr = 1e - 3$)
K (Samples)	**1, 5, 20, 30, 50, 100**

Implementation Detail

Used `torch.logsumexp` to avoid numerical underflow.

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

4 Results
   Log-Likelihood
   Latent Utilization
   K-Analysis
   Sample Quality

5 Conclusion

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

4 Results
Log-Likelihood
Latent Utilization
K-Analysis
Sample Quality

5 Conclusion

Result 1: Estimated Log-Likelihood vs Number of Samples (K)

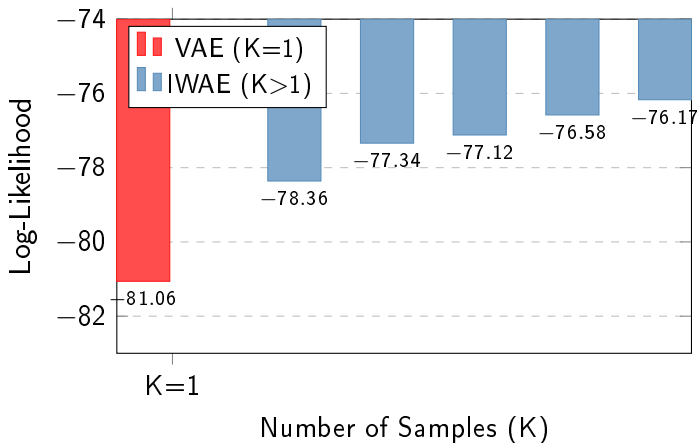- **Metric:** Estimated Log-Likelihood (Higher is better).

Figure 1: Estimated Log-Likelihood on MNIST Test Set (Higher is Better)

Introduction & Theory
oooo

The IWAE Solution
oooooo

Methodology
oo

Results
oooo●ooooo

Conclusion
ooo

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

4 Results
Log-Likelihood
Latent Utilization
K-Analysis
Sample Quality

5 Conclusion

Result 2: Latent Utilization

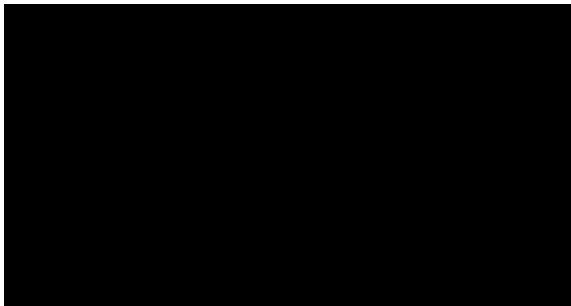- **Metric:** Active Units (Dimensions where $KL > \epsilon$).

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.

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

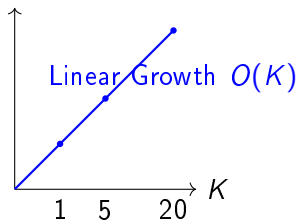**4** Results
   Log-Likelihood
   Latent Utilization
   K-Analysis
   Sample Quality

5 Conclusion

Analysis: Impact of K on Complexity

Training Time vs. K

Time / Memory

Linear Growth $O(K)$

1   5   20 K

Trade-off Analysis:

- **Compute:** Training time and memory usage scale **linearly** with K.

- **Performance:** Returns diminish as K increases (gap between $K = 5$ and $K = 20$ is smaller than $1 \rightarrow 5$).

- **Gradient Variance:** Higher K leads to lower variance gradients (DReG estimator), stabilizing training.

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

4 Results
Log-Likelihood
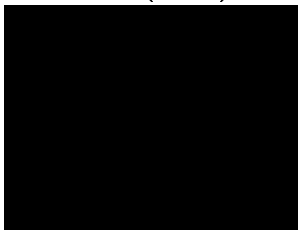Latent Utilization
K-Analysis
Sample Quality

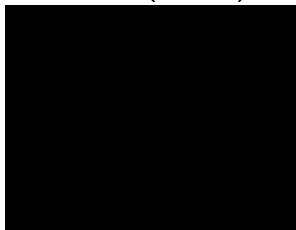**5** Conclusion

Result 3: Sample Quality

VAE (K=1)   IWAE (K=20)

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

1 Introduction & Theory

2 The IWAE Solution

3 Methodology

4 Results

5 Conclusion

Introduction & Theory
0000

The IWAE Solution
000000

Methodology
00

Results
000000000

Conclusion
0●0

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 ($\times K$)

Introduction & Theory
○○○○

The IWAE Solution
○○○○○○

Methodology
○○

Results
○○○○○○○○○

Conclusion
○○●

Thanks!

Q & A