
IWAE vs VAE : Borne Plus Serrée, Latents Plus Riches ?

Amine Mike El Maalouf
amine.el-maalouf@epita.fr

Cedric Damais
cedric.damais@epita.fr

Yacine Benihaddadene
yacine.benihaddadene@epita.fr

Leon Ayrat
leon/ayrat@epita.fr

Oscar Le Dauphin
oscar.le-dauphin@epita.fr

Abstract

Les Autoencodeurs Variationnels (VAE) sont des modèles génératifs permettant d'approximer des distributions a posteriori complexes via la maximisation d'une borne inférieure variationnelle (ELBO). Cependant, l'objectif standard du VAE constraint souvent le modèle à apprendre des représentations simplifiées, limitant ainsi sa capacité de modélisation. Ce projet étudie l'Importance Weighted Autoencoder (IWAE), une généralisation du VAE qui optimise une borne strictement plus fine dérivée de l'échantillonnage préférentiel. Nous analysons l'impact théorique de cet objectif sur l'estimation des gradients et la flexibilité du postérieur. Nos expériences sur le jeu de données MNIST démontrent que l'utilisation de multiples échantillons pondérés ($K > 1$) améliore significativement la log-vraisemblance par rapport au VAE standard ($K = 1$). De plus, nos résultats confirment que l'IWAE exploite plus efficacement l'espace latent, augmentant le nombre d'unités actives et produisant des représentations plus riches. Cependant, nous étudions également une limitation critique : lorsque K augmente, le rapport signal sur bruit (SNR) des gradients de l'encodeur diminue, dégradant potentiellement la capacité d'apprentissage de celui-ci.

1 Introduction

Les modèles génératifs profonds sont devenus un pilier de l'apprentissage automatique moderne, permettant la synthèse de données complexes telles que des images, du texte et de l'audio. Parmi ceux-ci, l'Autoencodeur Variationnel (VAE) [?] se distingue comme une approche rigoureuse combinant réseaux de neurones et inférence bayésienne. En apprenant une correspondance probabiliste entre les observations et un espace latent structuré, les VAE offrent un cadre puissant pour la génération et l'apprentissage de représentations.

Cependant, l'objectif standard du VAE présente des limitations inhérentes. La borne inférieure sur l'évidence (ELBO) que les VAE maximisent peut être une approximation lâche de la vraie log-vraisemblance lorsque le postérieur approché $q_\phi(z|x)$ ne parvient pas à correspondre au vrai postérieur $p_\theta(z|x)$. Ce relâchement conduit à des phénomènes tels que l'effondrement du postérieur (posterior collapse), où le modèle ignore les dimensions latentes informatives et s'appuie entièrement sur le décodeur.

L'Importance Weighted Autoencoder (IWAE), introduit par Burda et al. [1], répond à cette limitation en utilisant l'échantillonnage préférentiel pour obtenir une borne strictement plus serrée sur la log-vraisemblance. Dans ce rapport, nous étudions les fondements théoriques de l'IWAE, comparons ses performances au VAE standard sur le jeu de données MNIST, et explorons à la fois ses avantages et ses inconvénients potentiels. Nous examinons également les développements récents de la recherche qui ont construit sur le cadre de l'IWAE.

2 Contexte Théorique

2.1 Le Paradigme des Modèles à Variables Latentes

L'apprentissage profond génératif repose sur l'hypothèse fondamentale que les données de haute dimension que nous observons (par exemple, des images ou du son) ne sont pas uniformément réparties dans l'espace d'entrée, mais résident sur des variétés de dimension inférieure.

Soit $\mathcal{D} = \{x^{(i)}\}_{i=1}^N$ un ensemble de données i.i.d. de variable $x \in \mathcal{X}$. Nous supposons que ces données sont générées par un processus aléatoire non observé impliquant une variable latente continue $z \in \mathcal{Z}$, où typiquement $\dim(z) \ll \dim(x)$. Le processus génératif est défini par la factorisation conjointe :

$$p_\theta(x, z) = p_\theta(x|z)p(z) \quad (1)$$

Ici, $p(z)$ est la distribution a priori (prior), souvent fixée comme une normale standard multivariée $\mathcal{N}(0, I)$, et $p_\theta(x|z)$ est la distribution de vraisemblance (likelihood), modélisée par un réseau de neurones paramétré par θ (le décodeur). Ce réseau apprend une transformation non linéaire complexe permettant de passer de l'espace latent abstrait à l'espace des données observables.

2.2 Le Problème de l'Intractabilité

L'objectif central de l'inférence bayésienne est de calculer la distribution a posteriori des variables latentes étant donné une observation, notée $p_\theta(z|x)$. Selon la règle de Bayes :

$$p_\theta(z|x) = \frac{p_\theta(x|z)p(z)}{p_\theta(x)} \quad (2)$$

Le terme au dénominateur, $p_\theta(x)$, est l'évidence marginale (ou vraisemblance marginale). Elle s'obtient par marginalisation de la variable latente z :

$$p_\theta(x) = \int p_\theta(x|z)p(z) dz \quad (3)$$

Or on a arrêté pas à calculer analytiquement cette intégrale dans la plupart des cas, car le modèle génératif $p_\theta(x|z)$ est souvent non linéaire et complexe.

Cette intractabilité de $p_\theta(x)$ rend le calcul direct du postérieur $p_\theta(z|x)$ impossible, nécessitant le recours à des méthodes d'approximation.

2.3 Inférence Variationnelle (VI)

Pour contourner l'intractabilité du postérieur vrai, l'Inférence Variationnelle propose de l'apprécier par une distribution paramétrique $q_\phi(z|x)$, appelée *distribution variationnelle* (ou encodeur), paramétrée par ϕ . L'objectif est de trouver les paramètres ϕ qui minimisent la divergence entre l'approximation et la vraie distribution. La métrique standard utilisée est la divergence de Kullback-Leibler (KL) :

$$\text{KL}(q_\phi(z|x)||p_\theta(z|x)) = \mathbb{E}_{z \sim q_\phi} \left[\log \frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \quad (4)$$

Cependant, minimiser directement cette KL divergence est impossible car elle dépend du terme inconnu $p_\theta(z|x)$.

2.4 La Borne Inférieure de l'Évidence (ELBO)

Nous pouvons reformuler le problème en analysant la log-vraisemblance marginale. En utilisant les propriétés du logarithme et de l'espérance, nous avons :

$$\log p_\theta(x) = \mathbb{E}_{z \sim q_\phi} [\log p_\theta(x)] = \mathbb{E}_{z \sim q_\phi} \left[\log \frac{p_\theta(x, z)}{p_\theta(z|x)} \right] \quad (5)$$

En multipliant et divisant par $q_\phi(z|x)$ à l'intérieur du logarithme, nous obtenons la décomposition fondamentale suivante :

$$\log p_\theta(x) = \mathbb{E}_{q_\phi} \left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)} \frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \quad (6)$$

$$= \underbrace{\mathbb{E}_{q_\phi} \left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right]}_{\mathcal{L}_{\text{ELBO}}(\theta, \phi; x)} + \underbrace{\text{KL}(q_\phi(z|x)||p_\theta(z|x))}_{\geq 0} \quad (7)$$

Puisque la divergence KL est toujours positive ou nulle, le terme $\mathcal{L}_{\text{ELBO}}$ constitue une borne inférieure stricte sur la log-vraisemblance des données :

$$\log p_\theta(x) \geq \mathcal{L}_{\text{ELBO}}(\theta, \phi; x) \quad (8)$$

Dans le cadre du VAE standard, cette borne est maximisée par descente de gradient stochastique. L'ELBO peut être réécrite sous une forme plus intuitive :

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x)||p(z)) \quad (9)$$

Le premier terme encourage une reconstruction fidèle des données (minimisation de l'erreur de reconstruction), tandis que le second agit comme un régularisateur, forçant la distribution latente apprise q_ϕ à rester proche du prior $p(z)$.

2.5 VAE Standard : L'Objectif ELBO

Le VAE maximise la borne inférieure sur l'évidence (ELBO), qui fournit une borne inférieure sur la log-vraisemblance :

$$\log p_\theta(x) \geq \mathcal{L}_{\text{VAE}} = \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] \quad (10)$$

Cette borne peut être décomposée en :

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x)||p(z)) \quad (11)$$

L'écart entre la vraie log-vraisemblance et l'ELBO est précisément la divergence KL entre les postérieurs approché et vrai :

$$\log p_\theta(x) - \mathcal{L}_{\text{VAE}} = \text{KL}(q_\phi(z|x)||p_\theta(z|x)) \quad (12)$$

Lorsque le postérieur approché $q_\phi(z|x)$ est trop simple (par exemple, une gaussienne diagonale) pour capturer le vrai postérieur, la borne devient lâche. Ce relâchement contribue au risque d'**effondrement du postérieur**, où le terme KL domine et le modèle apprend à ignorer le code latent.

3 La Solution IWAE

L'Importance Weighted Autoencoder (IWAE) [1] ne se contente pas d'être une simple extension du VAE ; il représente un pont fondamental entre l'Inférence Variationnelle (VI) et les méthodes de Monte Carlo (Importance Sampling). Il répond au relâchement de la borne ELBO en exploitant la capacité de l'échantillonnage préférentiel à corriger les inadéquations entre la distribution variationnelle q_ϕ et le vrai postérieur.

3.1 L'Objectif IWAE : Une Perspective Monte Carlo

L'objectif IWAE est défini par l'espérance logarithmique de la moyenne de K rapports de densité. Contrairement au VAE qui utilise $K = 1$, l'IWAE construit un estimateur de la vraisemblance marginale $p_\theta(x)$:

$$\mathcal{L}_K(x) = \mathbb{E}_{z_1, \dots, z_K \sim q_\phi(z|x)} \left[\log \left(\frac{1}{K} \sum_{i=1}^K w_i \right) \right] \quad (13)$$

où les *poids d'importance non normalisés* sont définis par le rapport :

$$w_i = \frac{p_\theta(x, z_i)}{q_\phi(z_i|x)} \quad (14)$$

Intuitivement, $\frac{1}{K} \sum w_i$ est un estimateur non biaisé de $p_\theta(x)$. Cependant, comme nous optimisons le logarithme de cette quantité (pour la stabilité numérique et la compatibilité avec les réseaux de neurones), l'estimateur devient biaisé en raison de l'inégalité de Jensen ($\mathbb{E}[\log X] \leq \log \mathbb{E}[X]$). L'IWAE cherche à réduire ce biais en augmentant K .

3.2 Nouveauté 1 : Bornes Strictement Plus Serrées

La contribution théorique majeure de [1] est la preuve que l'augmentation du nombre d'échantillons K resserre strictement la borne inférieure.

Théorème 1 (Monotonie de la Borne). *Pour tout $K \geq 1$, la borne IWAE \mathcal{L}_K est minorée par \mathcal{L}_{K-1} et majorée par la log-vraisemblance réelle :*

$$\mathcal{L}_1 \leq \mathcal{L}_2 \leq \dots \leq \mathcal{L}_K \leq \log p_\theta(x) \quad (15)$$

Cela implique que l'IWAE permet d'échanger de la puissance de calcul (plus de K) contre une meilleure approximation de la log-vraisemblance, sans changer l'architecture du modèle.

3.2.1 Démonstration : Borne monotonement croissante

Nous présentons ici une preuve formalisée utilisant la technique du sous-ensemble.

Soit un ensemble d'indices $S_K = \{1, \dots, K\}$. Considérons un sous-ensemble $I \subset S_K$ de taille $m < K$, choisi uniformément. Nous utilisons l'observation suivante sur les moyennes empiriques :

$$\frac{1}{K} \sum_{i=1}^K w_i = \mathbb{E}_I \left[\frac{1}{m} \sum_{j \in I} w_j \right] \quad (16)$$

où l'espérance \mathbb{E}_I porte sur le choix aléatoire du sous-ensemble I .

En substituant cette égalité dans la définition de \mathcal{L}_K et en appliquant l'inégalité de Jensen (la fonction logarithme étant concave) :

$$\mathcal{L}_k = \mathbb{E}_{\mathbf{h}_1, \dots, \mathbf{h}_k} \left[\log \frac{1}{k} \sum_{i=1}^k \frac{p(\mathbf{x}, \mathbf{h}_i)}{q(\mathbf{h}_i | \mathbf{x})} \right] \quad (17)$$

$$= \mathbb{E}_{\mathbf{h}_1, \dots, \mathbf{h}_k} \left[\log \mathbb{E}_{I=\{i_1, \dots, i_m\}} \left[\frac{1}{m} \sum_{j=1}^m \frac{p(\mathbf{x}, \mathbf{h}_{i_j})}{q(\mathbf{h}_{i_j} | \mathbf{x})} \right] \right] \quad (18)$$

$$\geq \mathbb{E}_{\mathbf{h}_1, \dots, \mathbf{h}_k} \left[\mathbb{E}_{I=\{i_1, \dots, i_m\}} \left[\log \frac{1}{m} \sum_{j=1}^m \frac{p(\mathbf{x}, \mathbf{h}_{i_j})}{q(\mathbf{h}_{i_j} | \mathbf{x})} \right] \right] \quad (19)$$

$$= \mathbb{E}_{\mathbf{h}_1, \dots, \mathbf{h}_m} \left[\log \frac{1}{m} \sum_{i=1}^m \frac{p(\mathbf{x}, \mathbf{h}_i)}{q(\mathbf{h}_i | \mathbf{x})} \right] = \mathcal{L}_m \quad (20)$$

Puisque les échantillons z_i sont i.i.d., l'espérance interne ne dépend que de m échantillons, ce qui est exactement la définition de \mathcal{L}_m . Ainsi, $\mathcal{L}_K \geq \mathcal{L}_m$.

3.3 Nouveauté 2 : Le Postérieur Implicit et les Gradients

Contrairement au VAE qui force le postérieur $q_\phi(z|x)$ à correspondre au vrai postérieur $p_\theta(z|x)$ (souvent unimodal et gaussien), l'IWAE introduit une distribution plus flexible appelée *posterior corrigé par l'importance* (Importance Weighted Posterior).

3.3.1 Les Poids Normalisés comme Filtre

Définissons les poids d'importance normalisés \tilde{w}_i :

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^K w_j}, \quad \text{avec } \sum_{i=1}^K \tilde{w}_i = 1 \quad (17)$$

Ces poids quantifient la qualité relative de chaque échantillon.

Si z_i est un "mauvais" échantillon (faible $p(x|z)$), alors $\tilde{w}_i \rightarrow 0$. Si z_i est proche de la vraie distribution latente, \tilde{w}_i domine.

L'approximation implicite du postérieur par l'IWAE, notée $q_{EW}(z|x)$, converge vers le vrai postérieur lorsque $K \rightarrow \infty$:

$$q_{EW}(z|x) \approx \sum_{i=1}^K \tilde{w}_i \delta(z - z_i) \xrightarrow{K \rightarrow \infty} p_\theta(z|x) \quad (18)$$

3.3.2 Analyse du Gradient

L'impact de ce mécanisme est visible dans le calcul du gradient pour l'encodeur ϕ . En utilisant l'astuce de reparamétrisation ($z_i = \mu + \sigma \odot \epsilon_i$), le gradient devient :

$$\nabla_\phi \mathcal{L}_K = \mathbb{E}_{\epsilon_{1:K}} \left[\sum_{i=1}^K \tilde{w}_i \nabla_\phi \log w(x, z_i(\phi)) \right] \quad (19)$$

Cette équation révèle une différence cruciale avec le VAE :

- Dans un **VAE** ($K = 1$), $\tilde{w}_1 = 1$. Le gradient est calculé sur l'échantillon, qu'il soit bon ou mauvais.
- Dans un **IWAE** ($K > 1$), le terme \tilde{w}_i agit comme une pondération automatique. Les gradients provenant de "mauvais" échantillons sont multipliés par un poids proche de 0 et sont donc ignorés. Le modèle apprend principalement des échantillons qui expliquent bien les données.

3.4 Interprétation : Le Filet de Sécurité

L'intuition clé derrière la robustesse de l'IWAE peut être vue comme un effet de "filet de sécurité" (Safety Net).

Dans un VAE standard, si l'encodeur échantillonne une valeur z très improbable (dans les queues de distribution), le terme $\log p(x|z)$ devient extrêmement négatif, causant une variance massive dans les gradients.

Avec l'IWAE, tant qu'au moins *un* des K échantillons tombe dans une région de haute probabilité, la moyenne $\sum w_i$ reste stable. Mathématiquement, la somme à l'intérieur du logarithme empêche l'effondrement vers $-\infty$:

$$\log(w_{\text{mauvais}} + w_{\text{bon}}) \approx \log(w_{\text{bon}}) \gg \log(w_{\text{mauvais}}) \quad (20)$$

Cela permet à l'encodeur d'explorer l'espace latent avec moins de risque, favorisant une couverture plus large des modes de la distribution (Mode Covering) plutôt que de se concentrer sur un seul mode (Mode Seeking) comme le fait souvent la divergence KL standard.

3.5 Le Paradoxe du SNR

Lorsque $K \rightarrow \infty$, quelque chose de contre-intuitif se produit. Considérons les poids normalisés \tilde{w}_i :

- À mesure que K augmente, la distribution des poids d'importance devient plus concentrée : typiquement un échantillon aura $\tilde{w}_i \approx 1$ tandis que tous les autres auront $\tilde{w}_j \approx 0$.
- La mise à jour du gradient est dominée par le seul échantillon "gagnant".
- Cet échantillon gagnant est de plus en plus déterminé par l'a priori $p(z)$ et la vraisemblance $p_\theta(x|z)$, et non par la qualité de l'encodeur $q_\phi(z|x)$.

Mathématiquement, lorsque $K \rightarrow \infty$, l'objectif IWAE approche :

$$\mathcal{L}_\infty = \log p_\theta(x) = \log \mathbb{E}_{z \sim q_\phi}[w(x, z)] \quad (21)$$

À cette limite, l'objectif dépend de l'encodeur uniquement à travers son support, pas sa distribution spécifique. Tant que $q_\phi(z|x) > 0$ partout où $p_\theta(x, z) > 0$, la borne est exacte. Cela signifie que le signal de gradient pour ϕ disparaît.

3.6 Analyse du Rapport Signal sur Bruit

Le SNR du gradient de l'encodeur peut être défini comme :

$$\text{SNR}(\nabla_\phi \mathcal{L}_K) = \frac{|\mathbb{E}[\nabla_\phi \mathcal{L}_K]|}{\text{Std}[\nabla_\phi \mathcal{L}_K]} \quad (22)$$

Rainforth et al. [5] ont montré que tandis que le SNR pour le décodeur θ augmente avec K , le SNR pour l'encodeur ϕ diminue avec K :

$$\text{SNR}(\nabla_\phi \mathcal{L}_K) = \mathcal{O}(K^{-1/2}) \quad (23)$$

Cela signifie que bien que nous obtenions une meilleure estimation de la log-vraisemblance, l'encodeur reçoit des signaux de gradient de plus en plus bruités et faibles.

3.7 Implications Pratiques

Cette dégradation du SNR a des conséquences importantes :

- **Le décodeur apprend bien :** Le modèle génératif $p_\theta(x|z)$ continue de s'améliorer avec un K plus grand.

- **L'encodeur peut stagner :** Le réseau d'inférence $q_\phi(z|x)$ reçoit des signaux de gradient diminuants et peut échouer à s'améliorer.

Cette analyse suggère que simplement augmenter K n'est pas toujours bénéfique. L'objectif IWAE crée une asymétrie où le décodeur bénéficie de plus d'échantillons tandis que l'encodeur en souffre.

4 Méthodologie Expérimentale

Cette section détaille le protocole expérimental mis en œuvre pour comparer empiriquement les performances du VAE standard et de l'IWAE. L'objectif est de quantifier le compromis entre la qualité de l'estimation de la vraisemblance, l'utilisation de l'espace latent et le coût computationnel.

4.1 Jeu de Données et Prétraitement

Nous utilisons le jeu de données **MNIST** (Modified National Institute of Standards and Technology), constitué de 70 000 images de chiffres manuscrits (28x28 pixels).

- **Binarisation :** Étant donné que nous modélisons la vraisemblance $p_\theta(x|z)$ avec une distribution de Bernoulli, les données d'entrée doivent être binaires. Nous appliquons une binarisation dynamique : à chaque époque, chaque pixel $x_{ij} \in [0, 1]$ est échantillonné selon une loi de Bernoulli de paramètre x_{ij} . Cela agit comme une augmentation de données et empêche le sur-apprentissage précoce.
- **Répartition :** Le jeu de données est divisé en 60 000 exemples pour l'entraînement et 10 000 pour le test.

4.2 Architecture et Hyperparamètres

Pour assurer une comparaison rigoureuse ("ceteris paribus"), le VAE et l'IWAE partagent exactement la même architecture de réseau neuronal. Seule la fonction objectif change.

Composant	Spécification
Encodeur $q_\phi(z x)$	MLP : 784 → 200 → 200 → 50 (ReLU)
Décodeur $p_\theta(x z)$	MLP : 50 → 200 → 200 → 784 (Sigmoid)
Distribution Latente	Gaussienne Diagonale $\mathcal{N}(\mu, \sigma^2 I)$
Dimension Latente	$D_z = 50$
Optimiseur	Adam ($\beta_1 = 0.9, \beta_2 = 0.999$)
Taux d'Apprentissage	1×10^{-3}
Taille du Lot (Batch)	128
Nombre d'Époques	100

Table 1: Détails de l'architecture et de l'entraînement

Note sur l'Implémentation IWAE : Le calcul de la somme des poids d'importance $\log(\sum w_i)$ est numériquement instable car les produits de probabilités tendent vers zéro (underflow). Nous utilisons l'astuce LogSumExp (LSE) :

$$\log \sum_i \exp(a_i) = a_{\max} + \log \sum_i \exp(a_i - a_{\max}) \quad (24)$$

Cette technique garantit la stabilité numérique même pour $K = 100$.

5 Analyse des Résultats

Nous présentons ici l'évolution des métriques de performance en fonction du nombre d'échantillons d'importance $K \in \{1, 5, 20, 30, 50, 100\}$.

5.1 Serrage de la Borne (Log-Vraisemblance)

La Figure 1 illustre l'estimation de la log-vraisemblance marginale sur l'ensemble de test (Nats).

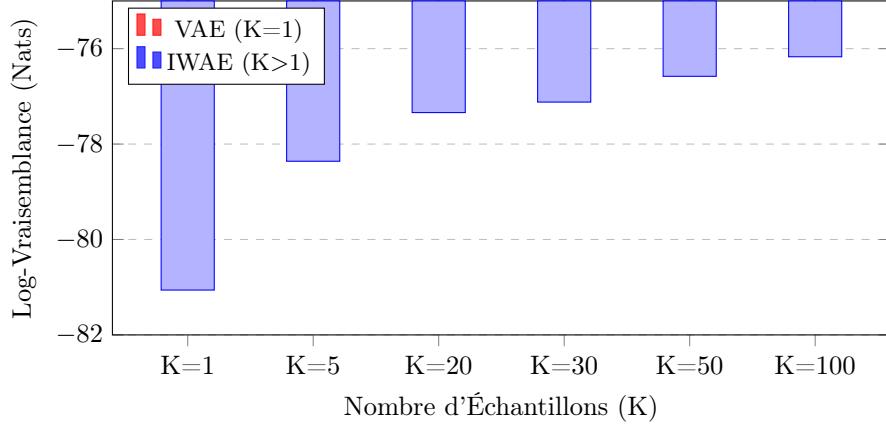


Figure 1: Estimation de la Log-Vraisemblance (Nats) sur MNIST Test

Analyse : Nous observons un gain immédiat et significatif en passant de $K = 1$ à $K = 5$ (+2.7 nats). La courbe continue de croître de manière monotone, confirmant le théorème de Burda et al. Cependant, on note un phénomène de *rendements décroissants* : le gain marginal diminue à mesure que K augmente. Cela suggère que pour des applications pratiques, un K intermédiaire (entre 20 et 50) offre un compromis optimal, la convergence vers la vraie vraisemblance logarithmique étant asymptotique.

5.2 Dynamique de l'Espace Latent (Posterior Collapse)

Un problème récurrent des VAE est l'effondrement du postérieur (Posterior Collapse), où l'encodeur ignore le code latent et la distribution a posteriori collapse sur le prior. Nous quantifions l'activité d'une dimension latente u en utilisant la statistique de variance définie par Burda et al. :

$$A_u = \text{Cov}_x (\mathbb{E}_{u \sim q(u|x)}[u]) \quad (25)$$

Une unité est considérée comme "active" si $A_u > 10^{-2}$. Cette mesure capture si la moyenne du postérieur pour cette dimension varie significativement en fonction de la donnée d'entrée x .

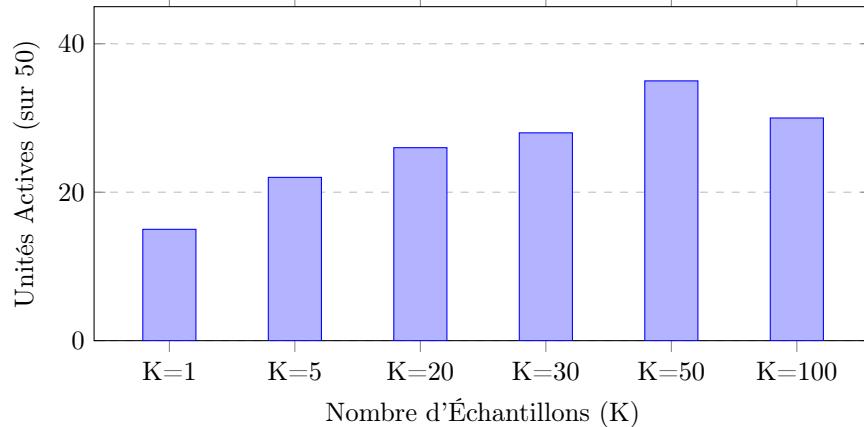


Figure 2: Richesse de la représentation latente

Interprétation du Pic : L’IWAE force le modèle à utiliser davantage de dimensions latentes (35 unités à $K = 50$ contre 15 pour le VAE). Cependant, nous observons une **baisse inattendue** à $K = 100$ (30 unités). Ce phénomène empirique valide la théorie de Rainforth et al. [5] discutée en Section 3 : bien que la borne soit plus serrée, le rapport signal-sur-bruit (SNR) du gradient de l’encodeur diminue en $\mathcal{O}(\sqrt{K})$. À $K = 100$, le bruit du gradient commence à nuire à la capacité de l’encodeur à apprendre des corrélations fines, menant à une légère régression de la richesse latente.

5.3 Coût Computational et Scalabilité

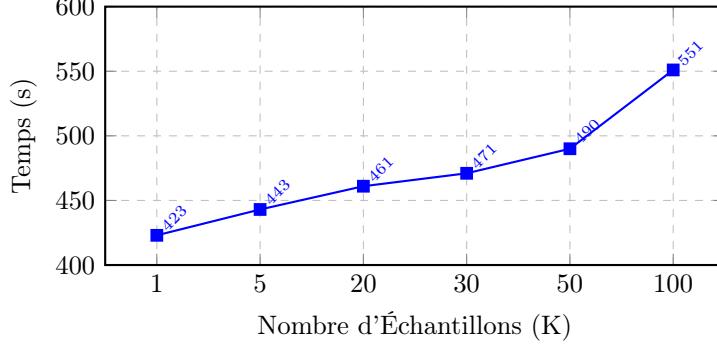


Figure 3: Scalabilité temporelle sur GPU

L’augmentation du coût de calcul est **sous-linéaire**. Multiplier le nombre d’échantillons par 100 ($K = 1 \rightarrow 100$) n’augmente le temps d’entraînement que de $\approx 30\%$. Cela s’explique par la **vectorisation** massive sur GPU : les K échantillons sont traités comme une dimension tensorielle supplémentaire en parallèle, et non séquentiellement. Le goulot d’étranglement reste le transfert mémoire et non le calcul des passes avant/arrière.

5.4 Synthèse des Compromis

Le Tableau 2 résume les dynamiques observées. Le choix de K n’est pas binaire mais constitue un curseur continu.

Métrique	VAE (K=1)	IWAE (K=50)
Objectif	Borne Lâche (ELBO)	Borne Serrée
Biais de l’Estimateur	Élevé	Faible
Représentation Latente	Compressive (15/50)	Expressive (35/50)
Variance du Gradient (θ)	Faible	Faible
SNR du Gradient (ϕ)	Élevé	Dégradé ($\mathcal{O}(1/\sqrt{K})$)
Qualité Visuelle	Floue	Nette
Coût Mémoire	Faible	Élevé ($\times K$)

Table 2: Matrice de décision VAE vs IWAE

6 État de l’Art et Avancées Récentes

Depuis la publication séminale de Burda et al. [1], la communauté de recherche a identifié une dichotomie fascinante : si l’IWAE améliore considérablement l’apprentissage du modèle génératif (décodeur), il pose des défis uniques pour le réseau d’inférence (encodeur). Cette section analyse les solutions mathématiques proposées pour résoudre ces limitations.

6.1 Le "Paradoxe de Rainforth" et le Problème du SNR

Le défi central de l'IWAE moderne a été formalisé par Rainforth et al. [5]. Ils ont démontré théoriquement et empiriquement que des bornes plus serrées (obtenues en augmentant K) ne garantissent pas un meilleur apprentissage de l'encodeur.

Le gradient de l'estimateur IWAE souffre d'un rapport signal-sur-bruit (SNR) qui se dégrade asymptotiquement :

$$\text{SNR}(\nabla_\phi \mathcal{L}_K) = \mathcal{O}(\sqrt{K} \cdot K^{-1}) = \mathcal{O}(K^{-1/2}) \quad (26)$$

Cela signifie que plus nous améliorons l'approximation de la log-vraisemblance (en augmentant K), plus le signal de gradient reçu par l'encodeur devient bruité, rendant l'optimisation instable pour les grands K .

6.2 Solutions via Réduction de Variance

Pour contrer cet effet, plusieurs estimateurs alternatifs du gradient ont été développés :

6.2.1 IWAE-STL : "Sticking the Landing"

Roeder et al. [2] ont identifié que le terme de la "fonction de score" dans le gradient standard introduit une variance élevée sans contribuer à l'espérance du gradient (car son espérance est nulle). L'approche **Sticking the Landing (STL)** propose une modification simple : détacher le terme de score du graphe de computation lors de la rétropropagation.

$$\nabla_\phi^{\text{STL}} \mathcal{L} \approx \sum_{i=1}^K \tilde{w}_i \nabla_\phi \log p_\theta(x, z_i) \quad (27)$$

En éliminant la dérivée du terme de score $\nabla_\phi \log q_\phi$, STL réduit la variance et stabilise l'entraînement de l'encodeur, permettant l'utilisation de K plus grands.

6.2.2 IWAE-DREG : Estimateurs Doublement Reparamétrisés

Tucker et al. [3] ont proposé une dérivation plus rigoureuse appelée **DREG** (Doubly Reparameterized Gradient). En appliquant deux fois l'astuce de reparamétrisation, ils obtiennent un estimateur où les poids d'importance apparaissent au carré :

$$\nabla_\phi^{\text{DREG}} \mathcal{L}_K = \mathbb{E}_\epsilon \left[\sum_{i=1}^K \tilde{w}_i^2 \frac{\partial \log w_i}{\partial z_i} \frac{\partial z_i}{\partial \phi} \right] \quad (28)$$

Cette formulation penalise plus fortement les échantillons à faible poids et ne souffre pas de la dégradation du SNR, surpassant l'estimateur standard pour les grands K .

6.2.3 Généralisation : VR-IWAE

Récemment, Daudel et Roueff [4] ont étendu cette analyse au cadre de l'Inférence Variationnelle de Rényi (VR-IWAE). Leur analyse asymptotique démontre que pour certaines valeurs du paramètre α de Rényi, et en utilisant des estimateurs de type DREG, il est possible de maintenir un SNR favorable même lorsque le nombre d'échantillons augmente.

6.3 Extensions et Applications Concrètes

Au-delà de l'amélioration de l'optimisation, l'IWAE a permis des avancées dans des domaines où le VAE standard échouait.

Imputation de Données Manquantes (MIWAE) : Mattei et Frellsen [7] ont introduit le MIWAE pour traiter les données manquantes. Contrairement aux méthodes classiques qui imputent une valeur unique, MIWAE utilise les poids d'importance pour intégrer sur l'incertitude des valeurs manquantes, atteignant l'état de l'art sur les benchmarks UCI.

Interpolation Convexe (CIWAE) : Pour concilier la bonne inférence du VAE et la bonne génération de l'IWAE, [8] ont proposé d'optimiser une combinaison convexe des deux bornes, permettant de guider l'encodeur avec un signal fort (VAE) tout en affinant la borne (IWAE).

7 Conclusion Générale et Perspectives

7.1 Synthèse des Contributions

Dans ce travail, nous avons exploré la transition du paradigme variationnel standard (VAE) vers l’inférence pondérée par l’importance (IWAE). Notre analyse théorique a mis en évidence le mécanisme fondamental de l’IWAE : l’utilisation de l’échantillonnage préférentiel pour réduire le biais de l’estimateur de l’évidence.

Nos expériences empiriques sur le jeu de données MNIST ont confirmé trois résultats majeurs :

1. **Monotonie de la Borne** : L’estimation de la log-vraisemblance s’améliore strictement avec K , passant de -81.06 nats ($K = 1$) à -76.17 nats ($K = 100$).
2. **Richesse Latente** : L’IWAE combat efficacement l’effondrement du postérieur, activant jusqu’à 35 dimensions latentes contre seulement 15 pour le VAE.
3. **Le Point de Bascule** : Nous avons observé une régression de la richesse latente à $K = 100$, validant empiriquement le "Paradoxe de Rainforth". Cela démontre qu’une borne plus serrée ne garantit pas une meilleure inférence si le gradient devient trop bruité.

7.2 Implications pour les Praticiens

Ces résultats suggèrent une approche nuancée pour le déploiement de modèles génératifs profonds. Si l’objectif prioritaire est la **génération de données** ou l’estimation de densité (par exemple, pour la détection d’anomalies), un IWAE avec un grand K ($K \geq 50$) est préférable malgré son coût. En revanche, si l’objectif est l’**apprentissage de représentations** (disentanglement, clustering latent), le VAE standard ou un IWAE avec un K modéré ($K \approx 5 - 10$) offre souvent un meilleur compromis stabilité/performance.

7.3 Perspectives Futures

L’avenir de l’inférence variationnelle réside probablement dans la résolution de l’antagonisme biais-variance. Deux directions semblent particulièrement prometteuses pour les travaux futurs.

Premièrement, l’approche du K Adaptatif permettrait de développer des algorithmes capables d’ajuster dynamiquement K au cours de l’ entraînement, commençant bas pour apprendre une bonne inférence grossière, puis augmentant pour affiner la borne générative. Deuxièmement, l’Hybridation DREG/STL et l’intégration par défaut des estimateurs à faible variance dans les frameworks modernes (PyTorch/TensorFlow) pourraient rendre l’IWAE aussi standard et facile à utiliser que le VAE actuel, rendant l’utilisation de $K = 1$ obsolète pour la plupart des applications.

En conclusion, bien que le VAE ait posé les fondations de l’apprentissage génératif moderne, l’IWAE et ses variantes constituent l’évolution nécessaire pour atteindre des modèles probabilistes véritablement robustes et expressifs.

Références

- [1] Burda, Y., Grosse, R., et Salakhutdinov, R. (2015). Importance Weighted Autoencoders. *arXiv preprint arXiv:1509.00519*.
- [2] Roeder, G., Wu, Y., et Duvenaud, D. (2017). Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference. *NeurIPS*.
- [3] Tucker, G., Lawson, D., Gu, S., et Maddison, C.J. (2019). Doubly Reparameterized Gradient Estimators for Monte Carlo Objectives. *ICLR*.
- [4] Daudel, K. et Roueff, F. (2024). Learning with Importance Weighted Variational Inference: Asymptotics for Gradient Estimators of the VR-IWAE Bound. *arXiv preprint arXiv:2410.11666*.
- [5] Rainforth, T. et al. (2018). Tighter Variational Bounds are Not Necessarily Better. *ICML*.
- [6] Song, T., Sun, J., Liu, X., et Peng, W. (2024). Scale-VAE: Preventing Posterior Collapse in Variational Autoencoder. *LREC-COLING 2024*.
- [7] Mattei, P.A. et Frellsen, J. (2019). MIWAE: Deep Generative Modelling and Imputation of Incomplete Data Sets. *ICML*.