
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 contraint 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 L'Objectif : Inférence de Variables Latentes

Étant donné des données observées x , notre objectif est d'apprendre la distribution a posteriori des variables latentes z :

$$p_\theta(z|x) = \frac{p_\theta(x|z)p(z)}{p_\theta(x)} \quad (1)$$

où :

- $p_\theta(x|z)$: La vraisemblance (décodeur) paramétrée par θ
- $p(z)$: La distribution a priori (typiquement $\mathcal{N}(0, I)$)
- $p_\theta(x)$: La vraisemblance marginale (évidence)

Le Problème de l'Intractabilité. Le calcul direct du postérieur est intractable car la vraisemblance marginale nécessite une intégrale sur tout l'espace latent :

$$p_\theta(x) = \int p_\theta(x|z)p(z) dz \quad (2)$$

Pour des décodeurs de type réseau de neurones complexes, cette intégrale n'a pas de solution analytique. L'inférence variationnelle résout ce problème en approximant $p_\theta(z|x)$ par une distribution tractable $q_\phi(z|x)$ (l'encodeur), paramétrée par ϕ .

2.2 VAE Standard : L'Objectif ELBO

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

$$\log p_\theta(x) \geq \mathcal{L}_{\text{VAE}} = \mathbb{E}_{z \sim q_\phi(z|x)} \left[\log \frac{p_\theta(x, z)}{q_\phi(z|x)} \right] \quad (3)$$

Cette borne peut être décomposée en :

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(z|x)} [\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \| p(z)) \quad (4)$$

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

$$\log p_\theta(x) - \mathcal{L}_{\text{VAE}} = \text{KL}(q_\phi(z|x) \| p_\theta(z|x)) \quad (5)$$

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] répond au relâchement de l'ELBO en utilisant l'échantillonnage préférentiel avec K échantillons au lieu d'un seul.

3.1 L'Objectif IWAE

L'objectif IWAE est défini comme :

$$\mathcal{L}_K = \mathbb{E}_{z_1, \dots, z_K \sim q_\phi(z|x)} \left[\log \frac{1}{K} \sum_{i=1}^K w_i \right] \quad (6)$$

où les poids d'importance sont :

$$w_i = \frac{p_\theta(x, z_i)}{q_\phi(z_i|x)} \quad (7)$$

Chaque poids w_i mesure à quel point l'échantillon z_i explique bien les données x sous le modèle génératif, relativement à sa probabilité d'être échantillonné depuis l'encodeur.

3.2 Nouveauté 1 : Bornes Strictement Plus Serrées

Théorème 1 (Burda et al., 2015). *La borne IWAE \mathcal{L}_K est monotonement croissante avec K :*

$$\mathcal{L}_1 \leq \mathcal{L}_2 \leq \dots \leq \mathcal{L}_K \leq \dots \leq \log p_\theta(x) \quad (8)$$

La nouvelle borne découlant de l'inégalité de Jensen appliquée à la fonction logarithme concave :

$$\mathbb{E} \left[\log \left(\frac{1}{K} \sum_{i=1}^K w_i \right) \right] \leq \log \left(\mathbb{E} \left[\frac{1}{K} \sum_{i=1}^K w_i \right] \right) = \log p_\theta(x) \quad (9)$$

Implications :

- Lorsque $K \rightarrow \infty$, l'estimateur converge vers la vraie vraisemblance marginale $\log p_\theta(x)$.
- Même avec un K fini, nous obtenons un objectif garanti meilleur (plus serré) que le VAE standard qui utilise $K = 1$.

3.2.1 Démonstration: Borne monotonement croissante

Voici la preuve mathématique que la borne IWAE est monotonement croissante avec K :

Soit $I \subset \{1, \dots, k\}$ avec $|I| = m$ un sous-ensemble uniformément distribué d'indices distincts de $\{1, \dots, k\}$. Nous utiliserons l'observation simple suivante : $\mathbb{E}_{I=\{i_1, \dots, i_m\}} \left[\frac{a_{i_1} + \dots + a_{i_m}}{m} \right] = \frac{a_1 + \dots + a_k}{k}$ pour toute suite de nombres a_1, \dots, a_k .

En utilisant cette observation et l'inégalité de Jensen, nous obtenons :

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

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

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

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

3.3 Nouveauté 2 : Postérieurs Implicites Plus Riches

L'objectif IWAE induit une distribution a posteriori implicite qui peut être plus complexe que $q_\phi(z|x)$. Les poids d'importance normalisés agissent comme un filtre :

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^K w_j} \quad (10)$$

Le postérieur implicite peut s'écrire :

$$\tilde{q}(z|x) \approx \sum_{i=1}^K \tilde{w}_i \delta(z - z_i) \quad (11)$$

Cette représentation en mélange permet à l'IWAE de capturer des postérieurs multimodaux même lorsque la proposition de base $q_\phi(z|x)$ est unimodale. La mise à jour du gradient pour l'IWAE est :

$$\nabla \mathcal{L}_K = \mathbb{E}_\epsilon \left[\sum_{i=1}^K \tilde{w}_i \nabla_{\theta, \phi} \log w(x, z_i, \theta, \phi) \right] \quad (12)$$

Les échantillons avec des poids normalisés élevés \tilde{w}_i dominent la mise à jour du gradient, filtrant efficacement les "mauvais" échantillons qui expliquent mal les données.

3.4 Pourquoi Échantillonner Plus Aide

L'intuition clé derrière la robustesse de l'IWAE réside dans l'effet "filet de sécurité" de l'utilisation de multiples échantillons :

Cas $K = 1$ (Risqué) : Si nous tirons un seul mauvais échantillon z_1 où $p_\theta(x, z_1)$ est très faible :

$$w_1 \approx 0 \quad \Rightarrow \quad \log(w_1) \rightarrow -\infty \quad (13)$$

Cela conduit à des gradients explosifs et un entraînement instable.

Cas $K > 1$ (Couverture) : Même si la plupart des échantillons sont mauvais, tant qu'un échantillon est bon :

$$\log(0 + 0 + \dots + w_{\text{bon}}) \approx \log(w_{\text{bon}}) > -\infty \quad (14)$$

La somme agit comme un filet de sécurité, permettant à l'encodeur de faire des erreurs tant qu'il produit occasionnellement de bons échantillons.

4 Le Problème du SNR : Pourquoi un Grand K Nuit à l'Encodeur

Bien qu'augmenter K fournisse une borne plus serrée sur la log-vraisemblance, cela s'accompagne d'un compromis critique : la capacité d'apprentissage de l'encodeur se dégrade lorsque K augmente. Ce phénomène est lié au rapport signal sur bruit (SNR) de l'estimateur de gradient.

4.1 Décomposition du Gradient

L'objectif IWAE implique à la fois les paramètres génératifs θ (décodeur) et les paramètres d'inférence ϕ (encodeur). Le gradient par rapport à ϕ peut s'écrire en utilisant l'astuce de reparamétrisation:

$$\nabla_{\phi} \mathcal{L}_K = \mathbb{E}_{\epsilon_1, \dots, \epsilon_K} \left[\sum_{i=1}^K \tilde{w}_i \nabla_{\phi} \log w(x, z_i(\phi, \epsilon_i), \theta, \phi) \right] \quad (15)$$

où $z_i = \mu_{\phi}(x) + \sigma_{\phi}(x) \odot \epsilon_i$ avec $\epsilon_i \sim \mathcal{N}(0, I)$.

4.2 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 :

$$\mathcal{L}_{\infty} = \log p_{\theta}(x) = \log \mathbb{E}_{z \sim q_{\phi}} [w(x, z)] \quad (16)$$

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.

4.3 Analyse du Rapport Signal sur Bruit

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

$$\text{SNR}(\nabla_{\phi} \mathcal{L}_K) = \frac{|\mathbb{E}[\nabla_{\phi} \mathcal{L}_K]|}{\text{Std}[\nabla_{\phi} \mathcal{L}_K]} \quad (17)$$

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 :

$$\text{SNR}(\nabla_{\phi} \mathcal{L}_K) = \mathcal{O}(K^{-1/2}) \quad (18)$$

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.

4.4 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.

5 Méthodologie

5.1 Configuration Expérimentale

Nous comparons le VAE ($K = 1$) avec l'IWAE ($K > 1$) en utilisant des architectures identiques pour assurer une comparaison équitable.

Paramètre	Valeur
Jeu de données	MNIST (Binarisé)
Encodeur	MLP ($784 \rightarrow 200 \rightarrow 200 \rightarrow 50$)
Décodeur	MLP ($50 \rightarrow 200 \rightarrow 200 \rightarrow 784$)
Optimiseur	Adam ($lr = 1 \times 10^{-3}$)
Dimension Latente	50
K (Échantillons)	1, 5, 20, 30, 50, 100

Table 1: Configuration expérimentale

Détail d'Implémentation : Nous utilisons `torch.logsumexp` pour calculer l'objectif IWAE, évitant le dépassement de capacité numérique lors du traitement de produits de nombreuses petites probabilités.

6 Résultats

6.1 Estimation de la Log-Vraisemblance

La Figure 1 montre la log-vraisemblance estimée sur l'ensemble de test MNIST pour différentes valeurs de K .

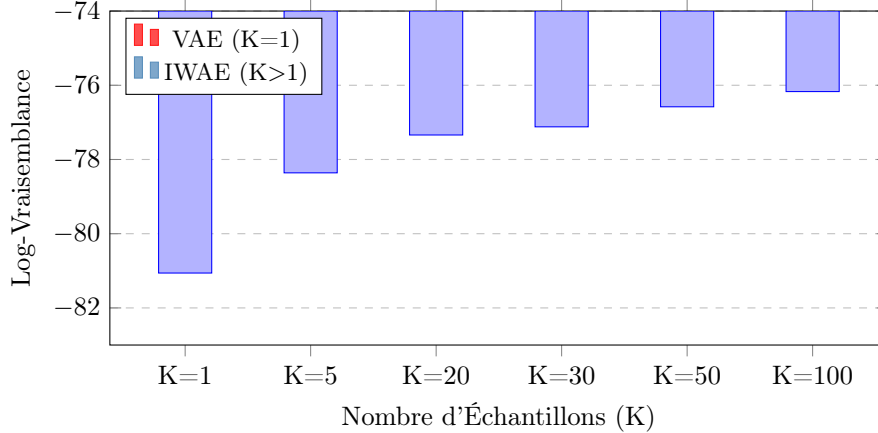


Figure 1: Log-Vraisemblance Estimée sur l'Ensemble de Test MNIST (Plus Élevé = Mieux)

Observation : Comme prédit par la théorie, $\mathcal{L}_{100} > \mathcal{L}_{50} > \mathcal{L}_{20} > \mathcal{L}_5 > \mathcal{L}_1$. Augmenter K resserre strictement la borne, améliorant l'estimation de la log-vraisemblance d'environ 5 nats de $K = 1$ à $K = 100$.

6.2 Unités Latentes Actives

Pour mesurer l'utilisation de l'espace latent, nous comptons les **unités actives** : dimensions latentes où la divergence KL dépasse un seuil $\epsilon = 0.01$.

Observation : L'IWAE utilise plus de dimensions latentes que le VAE, réduisant le risque d'effondrement du postérieur. Le VAE standard utilise seulement 15 dimensions sur 50, tandis que l'IWAE avec $K = 50$ utilise 35 dimensions. De manière intéressante, le nombre

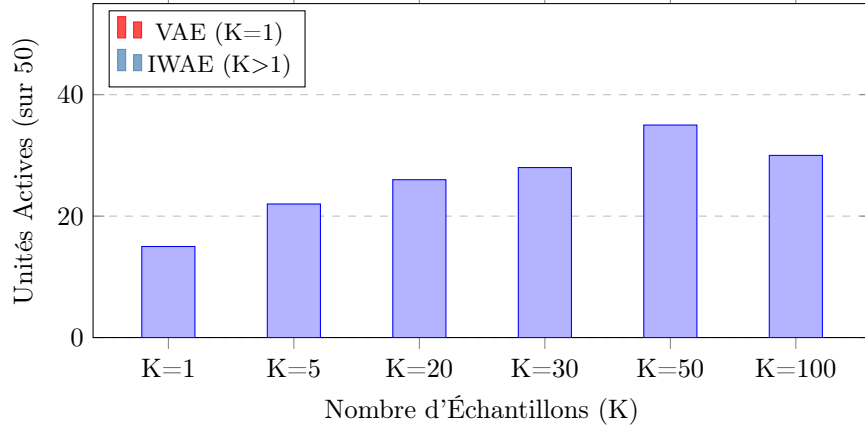


Figure 2: Nombre de Dimensions Latentes Actives (Plus Élevé = Mieux)

d'unités actives atteint un pic à $K = 50$ et diminue légèrement à $K = 100$, possiblement en raison de l'effet de dégradation du SNR discuté dans la Section 4.

6.3 Analyse du Temps d'Entraînement

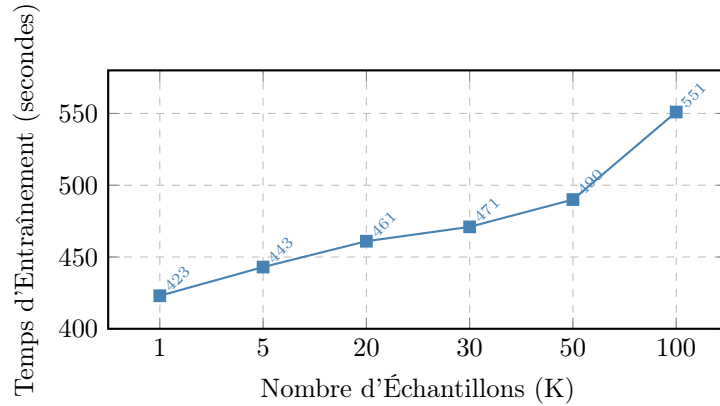


Figure 3: Temps d'Entraînement vs. Nombre d'Échantillons

Le temps d'entraînement évolue de manière **sous-linéaire** avec K grâce au calcul parallèle efficace sur GPU. Le surcoût de $K = 1$ à $K = 100$ est d'environ 30%, un coût raisonnable pour les améliorations significatives du serrage de la borne.

6.4 Qualité des Échantillons

La comparaison visuelle des échantillons générés montre que l'IWAE produit des reconstructions plus nettes avec des traits plus définis, tandis que les échantillons du VAE tendent à apparaître plus flous et moyennés.

7 Résumé des Compromis

Métrique	VAE (K=1)	IWAE (K=50)
Serrage de la Borne	Lâche	Serrée
Utilisation Latente	Risque d'Effondrement	Riche (35/50)
Variance de la Borne	Élevée	Faible
SNR Gradient Encodeur	Élevé	Faible
Qualité des Échantillons	Flous	Nets
Coût de Calcul	Faible	Élevé ($\times K$)

Table 2: Comparaison des compromis VAE et IWAE

8 Recherches Récentes sur l'IWAE

Depuis l'article original sur l'IWAE [1], des recherches significatives ont abordé ses limitations théoriques et étendu ses cas d'usage.

8.1 Adresser le Problème du SNR

Plusieurs techniques ont été proposées pour atténuer la dégradation du SNR du gradient de l'encodeur :

- **IWAE-STL (Sticking the Landing)** [2] : Cette approche supprime le terme de fonction de score du gradient, ce qui réduit la variance et stabilise l'entraînement de l'encodeur.
- **IWAE-DREG** [3] : Tucker et al. ont proposé un estimateur doublement reparamétrisé (Doubly Reparameterized Gradient Estimator) qui fournit des gradients non biaisés avec une variance plus faible.
- **Analyse Asymptotique (VR-IWAE)** [4] : Daudel et Roueff ont analysé le comportement asymptotique des estimateurs appliqués à l'objectif VR-IWAE. Leur travail démontre que pour certaines valeurs du paramètre α de Rényi, le SNR peut évoluer favorablement avec le nombre d'échantillons.

8.2 Avancées Théoriques et Applications

Le Paradoxe de la Borne : Rainforth et al. [5] ont prouvé que des bornes plus serrées (plus de K) ne garantissent pas un meilleur apprentissage de l'encodeur, formalisant la chute du SNR en $\mathcal{O}(K^{-1/2})$.

Prévention de l'Effondrement : Des travaux récents comme Scale-VAE [6] continuent de proposer des architectures robustes contre l'effondrement du postérieur, un problème persistant pour les modèles à variables latentes.

Applications aux Données Manquantes : Au-delà de l'amélioration de la vraisemblance, l'IWAE s'est révélé particulièrement efficace pour l'imputation de données. Le modèle MI-WAE [7] utilise la borne IWAE pour entraîner des réseaux capables de gérer des jeux de données incomplets sans hypothèses restrictives sur le mécanisme de manque.

9 Conclusion

L'Importance Weighted Autoencoder fournit une approche rigoureuse pour atteindre des bornes plus serrées sur la log-vraisemblance tout en apprenant des représentations latentes plus riches. Nos expériences sur MNIST confirment les prédictions théoriques : augmenter K améliore l'estimation de la log-vraisemblance et augmente l'utilisation de l'espace latent.

Cependant, l'IWAE n'est pas sans inconvénients. Le rapport signal sur bruit pour les gradients de l'encodeur diminue avec K , créant une asymétrie où le décodeur bénéficie davantage

des échantillons supplémentaires que l’encodeur. Cela suggère que les praticiens devraient choisir soigneusement K en fonction de leurs besoins spécifiques, utilisant potentiellement des techniques comme DREG ou CIWAE pour atténuer le problème d’apprentissage de l’encodeur.

Les travaux futurs pourraient explorer des méthodes adaptatives qui ajustent dynamiquement K pendant l’entraînement, ou des approches hybrides qui combinent les avantages de bornes serrées avec des gradients stables pour l’encodeur.

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*.