Wasserstein GAN are minimax optimal distribution estimators

Arthur Stéphanovitch [1] & Eddie Aamari [2] & Clément Levrard [3]

[1] *Université Paris Cité, Sorbonne Université, CNRS*
Laboratoire de Probabilités, Statistique et Modélisation
Paris, France
[2] *Ecole Normale Supérieure, Université PSL, CNRS*
Département de Mathématiques et Applications
F-75005 Paris, France
[3] *Université de Rennes, CNRS*
Institut de recherche mathématique de Rennes
F-35000 Rennes, France

Résumé. Nous étudions les taux de convergence non asymptotiques de l'estimateur Wasserstein Generative Adversarial Networks (WGAN). Précisément, on construit des classes de réseaux neuronaux représentant les générateurs et les discriminateurs qui donnent un GAN atteignant le taux minimax optimal pour l'estimation d'une certaine mesure de probabilité μ avec un support dans $\mathbb{R}^p$. La probabilité μ est considérée comme la poussée en avant de la mesure de Lebesgue sur le tore d-dimensionnel $\mathbb{T}^d$ par une application $g^\star : \mathbb{T}^d \to \mathbb{R}^p$ de régularité $\beta + 1$. En mesurant l'erreur avec l'IPM γ-Hölder , nous obtenons, à des facteurs logarithmiques près, le taux minimax optimal $O(n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-\frac{1}{2}})$, où n est la taille de l'échantillon, β détermine la régularité de la mesure cible μ, γ est la régularité de l'IPM ($\gamma = 1$ dans le cas de Wasserstein), et $d \leq p$ est la dimension intrinsèque de μ. Dans le processus, nous montrons une inégalité d'interpolation entre les IPM Hölder. Ce résultat de la théorie des espaces de fonctions généralise les inégalités d'interpolation classiques au cas où les mesures impliquées ont des densités sur des variétés différentes.

Mots-clés. Taux minimax, modèle génératif, estimation de mesure, sous-variété, inégalité d'interpolation

Abstract. We provide non asymptotic rates of convergence of the Wasserstein Generative Adversarial networks (WGAN) estimator. We build neural networks classes representing the generators and discriminators which yield a GAN that achieves the minimax optimal rate for estimating a certain probability measure μ with support in $\mathbb{R}^p$. The probability μ is considered to be the push forward of the Lebesgue measure on the d-dimensional torus $\mathbb{T}^d$ by a map $g^\star : \mathbb{T}^d \to \mathbb{R}^p$ of smoothness $\beta + 1$. Measuring the error with the γ-Hölder Integral Probability Metric (IPM), we obtain up to logarithmic factors, the minimax optimal rate $O(n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-\frac{1}{2}})$ where n is the sample size, β determines the smoothness of the target measure μ, γ is the smoothness of the IPM ($\gamma = 1$ is the Wasserstein case) and $d \leq p$ is the intrinsic dimension of μ. In the process, we derive a sharp interpolation inequality between Hölder IPMs. This novel result of theory of functions spaces generalizes classical interpolation inequalities to the case where the measures involved have densities on different manifolds.

Keywords. Minimax rate, generative model, distribution estimation, manifold, interpolation inequality

1 Résumé long

Soient $X_1, ..., X_n$ des points aléatoires i.i.d. tirés d'une mesure de probabilité μ avec un support dans $\mathbb{R}^p$. L'inférence de μ est un problème fondamental en statistique et en apprentissage automatique, pour lequel de nombreuses méthodes ont été développées (Tsybakov, 2004). Ces dernières années ont vu l'avènement de méthodologies génératives basées sur les réseaux antagonistes génératifs (GAN) (Goodfellow et al., 2014), avec des réalisations exceptionnelles dans les domaines de l'image (Karras et al., 2021), de la vidéo (Vondrick et al., 2016), et de la génération de texte (Yu et al., 2017). Dans cet article, nous nous concentrons sur l'approche GAN de Wasserstein (WGAN) de (Arjovsky et al., 2017), qui utilise la distance de Wasserstein 1 comme alternative à la divergence de Jensen-Shannon mise en œuvre dans le GAN traditionnel. Au fil des ans, les WGAN et leurs dérivés ont gagné en popularité dans la communauté de l'apprentissage automatique. Ils sont aujourd'hui considérés comme l'une des techniques génératives les plus réussies, obtenant des résultats de pointe dans des problèmes difficiles (Karras et al., 2021), tout en améliorant la stabilité et en éliminant des problèmes désagréables tels que le collapsus de mode (Gulrajani et al., 2017). Bien que les WGAN aient montré d'excellentes propriétés dans de nombreuses études empiriques rapportées dans la littérature sur l'apprentissage automatique (Liu et al. (2019), Luo and Lu (2018), Stanczuk et al. (2021)), bon nombre de leurs propriétés théoriques restent à étudier.

Le présent travail établit l'optimalité minimax de l'estimateur WGAN. Pour mettre en place la notation, le problème génératif consiste à utiliser les données $X_1, ..., X_n$ pour apprendre μ et, simultanément, être capable d'échantillonner à partir d'une distribution proche de celle-ci. Afin de résoudre ce problème, un réseau antagoniste génératif se compose d'une classe de fonctions génératrices $\mathcal{G}$ et d'une classe de discriminateurs $\mathcal{D}$. Étant donné une distribution facile à échantillonner ν sur un espace latent $\mathcal{Z}$, le générateur $\mathcal{G} \ni g : \mathcal{Z} \to \mathbb{R}^p$ approxime μ en essayant de minimiser sur $\mathcal{G}$ une certaine métrique de probabilité intégrale (IPM) (Müller, 1997) :

$$d_{\mathcal{D}}(\mu, g_{\#\nu}) := \sup_{D \in \mathcal{D}} \mathbb{E}_\mu[D(X)] - \mathbb{E}_\nu[D(g(Z))], \tag{1}$$

où $g_{\#\nu}$ représente la mesure de transfert de ν par g. L'objectif du discriminateur $\mathcal{D} \ni D : \mathbb{R}^p \to \mathbb{R}$ est de distinguer entre la vraie distribution et la fausse $g_{\#\nu}$, en maximisant sur $\mathcal{D}$ la quantité

$$L(g, D) := \mathbb{E}_\mu[D(X)] - \mathbb{E}_\nu[D(g(Z))].$$

Le problème min-max des réseaux antagonistes génératifs peut alors être écrit comme

$$\inf_{g \in \mathcal{G}} \sup_{D \in \mathcal{D}} \mathbb{E}\mu[D(X)] - \mathbb{E}\nu[D(g(Z))]. \tag{2}$$

On peut voir la classe $\mathcal{D}$ comme un sous-ensemble d'une classe plus large $\mathcal{F}$, avec $d_{\mathcal{F}}$ comme une métrique sur les distributions. Divers types de classes $\mathcal{F}$ ont été utilisés dans la littérature sur les GAN. Cela inclut les fonctions continues de Lipschitz (WGAN, (Arjovsky et al., 2017)), les fonctions de Sobolev (Sobolev GAN, (Mroueh et al., 2017)) et l'espace de Hilbert à noyau reproducteur (MMD GAN, (Li et al., 2017)). Ces différences sont à contraster avec celles plus historiquement utilisées dans l'estimation de densité non paramétrique classique, telles que la distance L^p, la distance de Hellinger et la divergence de Kullback-Leibler (Tsybakov, 2004), qui ne sont applicables que sous un modèle de domination. Dans le présent article, nous travaillons dans un cadre général où la mesure cible μ peut avoir une structure de basse dimension rendant ces mesures de divergence habituelles non pertinence. Nous choisissons la classe discriminative $\mathcal{F}$ comme étant la classe Hölder $\mathcal{H}_1^\gamma(\mathbb{R}^p, \mathbb{R})$ correspondant à la boule unité de fonctions de régularité $\gamma \geq 1$. On remarque que le cas $\gamma = 1$ est équivalent au cas de Wasserstein lorsque μ a un support compact.

On sait que les taux optimaux d'estimation d'une mesure μ avec un support dans $\mathbb{R}^p$ décroissent de manière exponentielle à mesure que la dimension ambiante p augmente. Pour surmonter cette "curse of dimensionality", certaines hypothèses structurelles de basse dimension sur μ doivent être imposées. Dans ce travail, nous supposons qu'il existe une application $g^\star \in \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^p)$ avec $\mathbb{T}^d$ le tore d-dimensionnel, telle que $\mu = g^\star_{\#U}$ avec $U \sim \mathcal{U}([0,1]^d)$ une variable aléatoire uniforme sur le cube. En particulier, il existe $Y_1, ..., Y_n$ i.i.d. tels que $Y_i \sim \mathcal{U}([0,1]^d)$ et $g^\star(Y_i) = X_i$. Notons que les Y_i sont inconnus, nous n'avons accès qu'aux X_i. Dans ce contexte, le problème d'inférence consiste à essayer de trouver un estimateur $\hat{\mu}$ de $\mu = g^\star_{\#U}$ basé sur l'échantillon $(X_i)i = 1, ..., n$, de telle sorte que l'erreur attendue

$$\mathbb{E}_{X_i \sim g^\star_{\#U}}[d_{\mathcal{H}_1^\gamma}(g^\star_{\#U}, \hat{\mu}(X_1, ..., X_n))],$$

soit aussi petite que possible. L'estimateur $\hat{\mu}$ est dit minimax optimal s'il n'existe aucun estimateur qui atteint un meilleur taux de convergence uniforme sur le modèle. Formellement, cela signifie qu'il existe une constante $C > 0$ indépendante de n telle que

$$\begin{aligned} &\sup_{g^\star \in \mathcal{H}_K^{\beta+1}} \mathbb{E}_{X_i \sim g^\star_{\#U}}[d_{\mathcal{H}_1^\gamma}(g^\star_{\#U}, \hat{\mu}(X_1, ..., X_n))] \\ &\leq C \inf_{\hat{T}} \sup_{g^\star \in \mathcal{H}_K^{\beta+1}} \mathbb{E}_{X_i \sim g^\star_{\#U}}[d_{\mathcal{H}_1^\gamma}(g^\star_{\#U}, \hat{T}(X_1, ..., X_n))]. \end{aligned}$$

Dans Divol (2021), l'auteur fournit un estimateur minimax (non génératif) des densités dans le cadre de la variété pour la distance de Wasserstein. Il utilise un estimateur local polynomial de la variété provenant de (Aamari and Levrard, 2017) couplé avec un estimateur de densité à noyau. Cependant, l'estimateur polynomial local est très coûteux en termes de calcul et ne peut donc pas être utilisé en haute dimension. Dans Tang and Yang (2022), l'auteur fournit également un estimateur minimax des densités dans le cadre de la variété mais pour la distance $d_{\mathcal{H}_1^\gamma}$. L'estimateur utilise une régularisation de la mesure empirique $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ avec un estimateur de variété couplé à une régularisation des ondelettes tronquée. Cet estimateur est purement théorique et serait extrêmement coûteux à mettre en œuvre en pratique. Par conséquent, l'existence d'un estimateur minimax calculable est toujours une question ouverte et cruciale à résoudre pour fournir des outils efficaces pour des applications concrètes.

Comme l'approche Wasserstein GAN (Arjovsky et al., 2017) s'est avérée facilement implémentable et a fourni des résultats de pointe dans divers domaines, nous nous concentrons dans cet article sur l'estimateur GAN

$$\hat{g} \in \arg\min_{g\in\mathcal{G}} \sup_{D\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n} D(X_i) - D(g(U_i)) \tag{3}$$

pour $U_i \sim \mathcal{U}([0,1]^d)$ i.i.d., $\mathcal{G} \subset \mathcal{H}_K^{\beta+1}(\mathbb{T}^d, \mathbb{R}^p)$ et $\mathcal{D} \subset \mathcal{H}_1^{\gamma}(\mathbb{R}^p, \mathbb{R})$ avec $\gamma \in [1, \beta+1]$. L'estimateur $\hat{g}$ doit être compris comme une approximation empirique de la solution de (2) basée sur les données $X_1, ..., X_n$. Les mesures de probabilité $\hat{g}_{\#U}$ sont alors naturellement nos estimateurs de la cible $g^\star_{\#U}$. Rappelons qu'étant donné que $\mathcal{D} \subset \mathcal{H}^\gamma$, le cas $\gamma = 1$ correspond à l'estimateur classique WGAN. L'une des forces de l'estimateur GAN (3) est qu'il effectue à la fois des estimations de support et de densité en même temps, ce qui permet en particulier d'éviter l'utilisation de tout estimateur de variété comme dans Divol (2021) et Tang and Yang (2022).

Contributions principales

Nous construisons des classes de réseaux neuronaux calculable $\mathcal{G}$ et $\mathcal{D}$, telles que

$$\sup_{g^\star\in\mathcal{H}_K^{\beta+1}} \mathbb{E}_{X_i\sim g^\star_{\#U}}[d_{\mathcal{H}_1^\gamma}(g^\star_{\#U}, \hat{g}_{\#U})] \leq C_1(\log n)^{C_2}\left(n^{-\frac{\beta+\gamma}{2\beta+d}} \vee n^{-\frac{1}{2}}\right), \tag{4}$$

avec $C_1, C_2 > 0$ des constantes indépendantes de n. Ce taux a été prouvé comme étant minimax optimal dans Tang and Yang (2022) jusqu'aux facteurs logarithmiques. À notre connaissance, il s'agit de la première étude montrant que l'estimateur GAN atteint les taux minimax pour les distances Wasserstein/Hölder. Ce résultat améliore, en particulier, les taux obtenus dans Chen et al. (2020), Schreuder et al. (2021) et Chae (2022). Les taux de convergence minimax de la version classique du GAN (Goodfellow et al., 2014) ont récemment été obtenus par Belomestny et al. (2023) pour la divergence Jensen–Shannon. Leur résultat traite uniquement du cadre de dimension pleine $d = p$, car la divergence Jensen–Shannon n'est non triviale que lorsque les mesures à comparer ne sont pas singulières l'une par rapport à l'autre.

Nous obtenons le taux de convergence minimax (4) sur les mesures $\mu = g^\star_{\#U}$ qui ont une densité par rapport à la mesure du volume d'une sous-variété inconnue. Dans le processus, des taux minimax sont également obtenus pour deux modèles statistiques intermédiaires :

- Nous traitons d'abord un cadre d basse dimension où la mesure cible peut avoir des atomes et son support n'est pas nécessairement une variété. Dans ce cas, la classe des discriminateurs $\mathcal{D}$ utilisée est théorique, ce qui signifie qu'elle n'est pas calculable en pratique. Ce modèle est étudié pour discuter des limites des hypothèses du cas général.
- Nous prouvons des taux minimax dans le cadre de la dimension pleine $p = d$, où la mesure cible a une densité par rapport à la mesure de Lebesgue p-dimensionnelle. Ce cas est traité pour comprendre dans le cadre plus simple de la dimension pleine comment des hypothèses supplémentaires peuvent nous aider à obtenir un estimateur calculabe.

- En adaptant la méthode développée dans le cas de dimension pleine au cas de la sous-variété, nous proposons un estimateur GAN calculable atteignant des taux minimax pour tous les $\gamma \in [1, \beta + 1]$ simultanément.

Le résultat principal est démontré à l'aide d'une nouvelle inégalité d'interpolation qui borne la distance $d_{\mathcal{H}_1^\gamma}(g_{\#U}, g^\star_{\#U})$ par la quantité $d_{\mathcal{H}_1^{\beta+1}}(g_{\#U}, g^\star_{\#U})^{\frac{\beta+\gamma}{2\beta+1}}$ et un facteur logarithmique.

References

Eddie Aamari and Clément Levrard. Non-asymptotic rates for manifold, tangent space, and curvature estimation, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y.W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. PMLR, 2017.

Denis Belomestny, Eric Moulines, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Rates of convergence for density estimation with generative adversarial networks, 2023.

Minwoo Chae. Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks. *arXiv preprint arXiv:2202.02890*, 2022.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*, 2020.

Vincent Divol. Measure estimation on manifolds: an optimal transport approach, 2021.

I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A.C. Courville. Improved training of Wasserstein GANs. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777. Curran Associates, Inc., 2017.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *CoRR*, abs/2106.12423, 2021.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network, 2017.

Yufei Liu, Yuan Zhou, Xin Liu, Fang Dong, Chang Wang, and Zihong Wang. Wasserstein gan-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering*, 5(1):156–163, 2019.

Yun Luo and Bao-Liang Lu. Eeg data augmentation for emotion recognition using a conditional wasserstein gan. In *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 2535–2538. IEEE, 2018.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan, 2017.

A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pages 1051–1071. PMLR, 2021.

Jan Stanczuk, Christian Etmann, Lisa Maria Kreusser, and Carola-Bibiane Schönlieb. Wasserstein gans work because they fail (to approximate the wasserstein distance). *arXiv preprint arXiv:2103.01678*, 2021.

Rong Tang and Yun Yang. Minimax rate of distribution estimation on unknown submanifold under adversarial losses, 2022.

Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 613–621. Curran Associates, Inc., 2016.

L. Yu, W. Zhang, J. Wang, and Y. Yu. SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2852–2858. AAAI Press, 2017.