# Estimation d'une distance de Wasserstein par transport optimal entropique

Jérémie Bigot [1], Paul Freulon [2], Boris P.Hejblum [3] & Arthur Leclaire [4]

[1] *Université de Bordeaux, Bordeaux, 33000, France. Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), 33400 Talence, France.*
*jeremie.bigot@math.u-bordeaux.fr*

[2] *EPFL, Institut de Mathématiques, CH-1015 Lausanne, Suisse. Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), 33400 Talence, France.*
*paul.freulon@epfl.ch*

[3] *Université de Bordeaux, Bordeaux, 33000, France. Bordeaux Population Health Research Center Inserm U1219, Inria SISTM, 33000 Bordeaux, France. Vaccine Research Institute (VRI), 94010 Créteil, France.*
*boris.hejblum@u-bordeaux.fr*

[4] *LTCI, Télécom Paris, IP Paris, 91120 Palaiseau, France. Institut de Mathématiques de Bordeaux et CNRS (UMR 5251), 33400 Talence, France.*
*arthur.leclaire@telecom-paris.fr*

**Résumé.** Le transport optimal entropique a été introduit en 2013 par M.Cuturi pour accélérer le calcul des distances de transport optimal. Dans cette présentation, on cherche à estimer une distance de Wasserstein en utilisant ce transport optimal régularisé. L'estimateur étudié est une distance de transport entropique où les deux mesures ont été remplacées par leurs versions empiriques. On présente des résultats de convergence de cet estimateur vers la distance Wasserstein évaluée entre les mesures sous-jacentes aux observations. Dans un premier temps, on essaye de motiver ce problème par une question classique de statistique : comment ajuster un modèle statistique à des observations ? Dans une deuxième partie, on fait l'état de l'art de résultats permettant de contrôler l'erreur de l'estimateur considéré. Enfin, on présente un nouvelle borne d'erreur pour l'estimateur étudié. De cette borne, on déduit un choix du paramètre de régularisation.

**Mots-clés.** Transport optimal, régularisation entropique, vitesse de convergence.

**Abstract.** For a faster computation of the optimal transport problem, M.Cuturi introduced in 2013 the entropic optimal transport. In this communication, we estimate a Wasserstein distance thanks to this regularized optimal transport. More precisely, we substitute the compared measures by their empirical counterparts in the regularized Wasserstein distance. We present rates of convergence of this estimator towards the Wasserstein distance between the underlying measures. Firstly, we try to motivate this estimation problem with a standard statistical question: how to fit a statistical model to some observations? Secondly, we review some known results that enable us to control the estimation error. Finally, we present a new bound on the estimation error. From this new bound, we deduce a choice for the regularization parameter.

**Keywords.** Optimal transport, entropic regularization, rate of convergence.

# 1 Notations, motivations, et problème étudié

Dans ce texte, on travaille dans l'espace euclidien à $d$ dimensions $\mathbb{R}^d$. Soit $Y_1, \dots, Y_n \in \mathbb{R}^d$ une série de $n$ observations supposées indépendantes et identiquement distribuées. Un problème classique est d'approcher la loi de $Y_1, \dots, Y_n$ par une collection de mesures $\mathcal{P} := \{\mu_\theta \ : \ \theta \in \Theta\}$. En termes statistiques, on parle d'un modèle $\mathcal{P}$ paramétré par l'ensemble $\Theta$. Une approche basée sur la fonction de vraisemblance permet d'aborder ce problème. Sous certaines hypothèses, la fonction de vraisemblance s'interprète comme une version empirique de la divergence de Kullback-Leibler [13, Thm. 9.13].

**Definition 1.1.** *Soit $\mu$ et $\nu$ deux mesures de probabilité sur $\mathbb{R}^d$. La divergence de Kullback-Leibler entre $\mu$ et $\nu$ est définie par*

$$\mathrm{KL}(\mu|\nu) = \int_{\mathbb{R}^d} \log\Big(\frac{d\mu}{d\nu}\Big) d\mu,$$

*si $\mu$ est absolument continue par rapport à $\nu$. Sinon, $\mathrm{KL}(\mu|\nu) = +\infty$.*

Les approches reposant sur la fonction de vraisemblance, ou la divergence de Kullback-Leibler, nécessitent des hypothèses d'absolue continuité. Retirer ces hypothèses motive l'utilisation des distances de transport en statistique [1, 7].

**Definition 1.2.** *Soit $\mu$ et $\nu$ deux mesures de probabilité sur $\mathbb{R}^d$ admettant des moments d'ordre deux. En notant $\Pi(\mu, \nu)$ l'ensemble des mesures sur $\mathbb{R}^d \times \mathbb{R}^d$ ayant pour marginales $\mu$ et $\nu$, la 2-distance de Wasserstein $W_0(\mu, \nu)$ est définie par*

$$W_0(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x - y\|^2 d\pi(x, y), \tag{1.1}$$

*où $\|x - y\|$ est la distance euclidienne entre $x$ et $y$.*

La quantité $W_0$ introduite en équation (1.1) permet de définir une distance entre mesures de probabilité [12]. Le calcul numérique de la distance $W_0$ est relativement lent. C'est pour en accélérer le calcul que M.Cuturi propose de régulariser le problème d'optimisation (1.1). À notre connaissance, c'est dans l'article [4] qu'est introduit pour la première fois la distance de Wasserstein régularisée.

**Definition 1.3.** *Soit $\mu$ et $\nu$ deux mesures de probabilité sur $\mathbb{R}^d$; et $\lambda \geq 0$ un paramètre de régularisation. La distance de Wasserstein régularisée $W_\lambda$ est définie par le problème*

$$W_\lambda(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x - y\|^2 d\pi(x, y) + \lambda \,\mathrm{KL}(\pi|\mu \otimes \nu). \tag{1.2}$$

Remarquons que lorsque $\lambda = 0$, on retrouve la distance de Wasserstein classique (1.1). D'autres régularisations que la divergence de Kullback-Leibler ont été proposées [8]. Pour le distinguer des autres régularisations, le problème (1.2) est parfois désigné par transport optimal entropique. Enfin, même si on parle de distance lorsque $\lambda > 0$, le problème de

optimal régularisé ne définit plus une distance. En effet, sur la droite réel $\mathbb{R}$, en posant $\mu = (\delta_0 + \delta_1)/2$, on a $W_\lambda(\mu, \mu) > 0$ lorsque $\lambda > 0$.

Les critères d'écart entre lois de probabilité étant introduits, revenons à des considérations statistiques. On va se concentrer sur l'utilisation des distances de transport (1.1) et (1.2). Supposons nos observations $Y_1, \dots, Y_n$ identiquement distribuées selon une mesure inconnue $\nu$. En utilisant une distance de transport comme critère d'écart, approcher la mesure $\nu$ par le modèle $\{\mu_\theta \mid \theta \in \Theta\}$ nécessite de résoudre le problème $\min_{\theta\in\Theta} W_0(\mu_\theta, \nu)$.

Dans ce texte; plutôt que d'étudier le problème $\min_{\theta\in\Theta} W_0(\mu_\theta, \nu)$, on va aborder une question préliminaire: estimer $W_0(\mu, \nu)$. Comme nous sommes dans un problème statistique, la mesure $\nu$ est inconnue. Dans certains cas [6], la mesure $\mu$ n'est pas connue non plus. C'est ce qu'on va supposer dans la suite de ce texte. Soit $X_1, \dots, X_n \sim \mu$; et soit $Y_1, \dots, Y_n \sim \nu$. On va substituer $\mu$ et $\nu$ par leurs mesures empiriques respectivement définies par

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i} \quad \text{et} \quad \hat{\nu}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_i}.$$

De plus, afin d'accélérer le calcul, on va remplacer la distance $W_0$ par sa version régularisée $W_\lambda$. Ainsi, alors que l'on souhaite étudier la quantité $W_0(\mu, \nu)$, on va calculer $W_\lambda(\hat{\mu}_n, \hat{\nu}_n)$. Autrement dit, $W_\lambda(\hat{\mu}_n, \hat{\nu}_n)$ est un estimateur de la quantité d'intérêt $W_0(\mu, \nu)$. C'est pourquoi nous cherchons à contrôler l'écart

$$\mathbb{E}[|W_0(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|]. \tag{1.3}$$

Dans la suite de ce texte, nous allons discuter de cette question. En section 2 on examine certains résultats déjà établis permettant de contrôler (1.3). Puis, en section 3, on propose de nouveaux résultats permettant de majorer l'erreur (1.3). Finalement, on déduit de la majoration obtenue un choix de paramètre de régularisation. Tous nos résultats nécessitent de supposer que les mesures $\mu$ et $\nu$ ont des supports bornés. Plus précisément, on va supposer que les supports de $\mu$ et $\nu$ sont inclus dans la boule centrée de rayon $R$. On note cette boule $B(0, R) := \{x \in \mathbb{R}^d \ : \ \|x\| \leq R\}$. De plus, toutes les observations considérées sont supposées indépendantes.

# 2 État de l'art

Une première étape pour contrôler l'erreur (1.3) est de la décomposer entre un terme d'approximation et un terme d'estimation. On commence donc par l'inégalité

$$|W_0(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)| \leq \underbrace{|W_0(\mu, \nu) - W_\lambda(\mu, \nu)|}_{\text{Approximation}} + \underbrace{|W_\lambda(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|}_{\text{Estimation}}. \tag{2.1}$$

Pour l'erreur d'approximation, on exploite le résultat suivant.

**Theorem 2.1.** *[5, Thm. 1] Soit $\lambda > 0$. Si $\mu$ et $\nu$ ont leur supports inclus dans la boule centrée de rayon $R$, alors*

$$|W_0(\mu,\nu) - W_\lambda(\mu,\nu)| \leq 2d\lambda \log\left(\frac{8e^2R^2}{\sqrt{d}\lambda}\right). \tag{2.2}$$

Dans le même article est établi l'erreur d'estimation (désignée par "sample complexity" dans [5]) que l'on rappelle ci-dessous.

**Theorem 2.2.** *[5, Thm. 3] Soit $\lambda > 0$. Si $X_1, \dots, X_n$ et $Y_1, \dots, Y_n$ sont respectivement distribuées selon $\mu$ et $\nu$, alors*

$$\mathbb{E}[|W_\lambda(\mu,\nu) - W_\lambda(\hat{\mu}_n,\hat{\nu}_n)|] \lesssim \left(1 + \frac{1}{\lambda^{\lfloor d/2 \rfloor}}\right)\frac{1}{\sqrt{n}}, \tag{2.3}$$

*le symbole $\lesssim$ cachant une constante multiplicative dépendant de $R$ et $d$.*

Remarquons que dans [5], un facteur $e^{R^2/\lambda}$ était présent dans la majoration (2.3). Ce facteur $e^{R^2/\lambda}$ a été enlevé dans les travaux ultérieurs [10, 3]. En utilisant la décomposition (2.1), ainsi que les théorèmes 2.1 et 2.2, si la dimension $d$ est paire, on obtient le contrôle

$$\mathbb{E}\left[|W_0(\mu,\nu) - W_\lambda(\hat{\mu}_n,\hat{\nu}_n)|\right] \lesssim \lambda \log(\lambda^{-1}) + \lambda^{-d/2} n^{-1/2}.$$

Si l'on essaye d'optimiser en $\lambda$ le membre de droite de cette dernière inégalité, on obtient $\mathbb{E}\left[|W_0(\mu,\nu) - W_\lambda(\hat{\mu}_n,\hat{\nu}_n)|\right] \lesssim n^{-1/(d+2)} \log(n)$. Cette vitesse de convergence est largement plus lente que celle obtenue avec l'estimateur *non* régularisé $W_0(\hat{\mu}_n,\hat{\nu}_n)$. La vitesse de convergence de cet estimateur a été établi par Chizat et al. en 2020. Dans un soucis de concision, on rappelle ce résultat uniquement dans le cas où la dimension $d$ est strictement supérieure à quatre.

**Theorem 2.3.** *[3, Thm. 2] Soit $X_1, \dots, X_n$ et $Y_1, \dots, Y_n$ deux séries de $n$ observations respectivement distribuées selon $\mu$ et $\nu$. Si la dimension $d$ des observations est strictement supérieure à quatre, on a alors*

$$\mathbb{E}[|W_0(\mu,\nu) - W_0(\hat{\mu}_n,\hat{\nu}_n)|] \lesssim n^{-2/d}.$$

Cette vitesse de convergence décroît lorsque la dimension augmente. On peut donc s'interroger sur l'existence d'un estimateur convergeant plus rapidement. Sans hypothèses supplémentaires sur les mesures considérées, la réponse à cette question est négative.

**Theorem 2.4.** *[9, Thm. 22] Sous l'hypothèse que $n$ observations indépendantes sont disponibles pour chaque mesure $\mu$ et $\nu$, on a*

$$(n\log(n))^{-2/d} \lesssim \inf_{\widehat{W}_n} \sup_{\substack{\mu\in\mathcal{P}(\mathbb{R}^d),\\ \nu\in\mathcal{P}(\mathbb{R}^d)}} \mathbb{E}\left[|W_0(\mu,\nu) - \widehat{W}_n|\right], \tag{2.4}$$

*où la borne inférieure est calculée sur l'ensemble des estimateurs de $W_0(\mu,\nu)$.*

Le théorème 2.4 est un résultat de type "minimax". On en déduit que la vitesse de convergence en $n^{-2/d}$ de l'estimateur non régularisé $W_0(\hat{\mu}_n,\hat{\nu}_n)$ est, en considérant le pire scénario, optimal.

# 3 Nouveaux résultats

Cette section présente nos principaux résultats concernant la convergence de $W_{\lambda_n}(\hat{\mu}_n, \hat{\nu}_n)$ vers $W_0(\mu, \nu)$. On va montrer qu'un paramètre de régularisation adapté permet d'atteindre (à un facteur logarithmique près) la vitesse minimax $n^{-2/d}$. Pour un texte plus détaillé, la lecture de la prépublication [2] est proposée.

L'amélioration de la vitesse de convergence que l'on obtient repose sur un nouveau contrôle de l'erreur d'estimation $|W_\lambda(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|$. Ce contrôle ne dépend pas du paramètre de régularisation.

**Proposition 3.1.** *[2, Prop. 3.1] Soit $\lambda \geq 0$. Si $n$ observations indépendantes sont disponibles pour les mesures $\mu$ et $\nu$, dans le cas où la dimension $d$ des observations est strictement supérieure à quatre, on a*

$$\mathbb{E}[|W_\lambda(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|] \lesssim n^{-2/d}, \tag{3.1}$$

*où $\lesssim$ cache une constante multiplicative dépendant de $R$ et de la dimension $d$.*

En combinant ce nouveau contrôle de l'erreur d'estimation avec l'erreur d'approximation donnée par le théorème 2.1, on obtient l'inégalité

$$\mathbb{E}[|W_0(\mu, \nu) - W_\lambda(\hat{\mu}_n, \hat{\nu}_n)|] \lesssim n^{-2/d} + \lambda \log(\lambda^{-1}). \tag{3.2}$$

Dans cette dernière inégalité (3.2), nous n'avons pas de contrôle sur l'erreur d'estimation $n^{-2/d}$. En revanche, l'erreur d'approximation de l'ordre $\lambda \log(\lambda^{-1})$ est contrôlée par le paramètre de régularisation. On va donc choisir ce paramètre $\lambda$ de façon à obtenir une erreur d'approximation de même ordre de grandeur que l'erreur d'estimation.

**Theorem 3.1.** *[2, Prop. 4.2] Si $n$ observations indépendantes sont disponibles pour chacune des deux mesures $\mu$ et $\nu$, alors*

$$\mathbb{E}[|W_0(\mu, \nu) - W_{\lambda_n}(\hat{\mu}_n, \hat{\nu}_n)|] \lesssim n^{-2/d} \log(n) \quad \textit{avec} \quad \lambda_n = n^{-2/d},$$

*où $\lesssim$ cache une constante multiplicative dépendant uniquement de $R$ et $d$.*

Ce dernier résultat montre qu'un choix de paramètre $\lambda_n$ dépendant du nombre d'observations disponibles permet atteindre, à un facteur logarithmique près, la vitesse minimax avec l'estimateur $W_{\lambda_n}(\hat{\mu}_n, \hat{\nu}_n)$. On pourrait choisir un paramètre de régularisation encore plus petit que $n^{-2/d}$; mais un tel choix ralentirait le calcul numérique de $W_\lambda(\hat{\mu}_n, \hat{\nu}_n)$. L'algorithme de Sinkhorn [11] permet le calcul de coût de transport régularisé $W_\lambda$ dans le cas où les mesures comparées sont discrètes. La vitesse de convergence de cet algorithme ralentit lorsque le paramètre de régularisation décroit.

# References


[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[2] J. Bigot, P. Freulon, B. P. Hejblum, and A. Leclaire. On the potential benefits of entropic regularization for smoothing wasserstein estimators. *arXiv preprint arXiv:2210.06934*, 2022.

[3] L. Chizat, P. Roussillon, F. Léger, F. Vialard, and G. Peyré. Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. In *Proc. NeurIPS'20*, 2020.

[4] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[5] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

[6] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[7] M. Hallin, G. Mordant, and J. Segers. Multivariate goodness-of-fit tests based on Wasserstein distance. *Electronic Journal of Statistics*, 15(1):1328 – 1371, 2021.

[8] D. A. Lorenz, P. Manns, and C. Meyer. Quadratically regularized optimal transport. *Applied Mathematics & Optimization*, 83(3):1919–1949, 2021.

[9] T. Manole and J. Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *The Annals of Applied Probability*, 34(1B):1108–1135, 2024.

[10] G. Mena and J. Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[12] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[13] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2004.