Central Limit Theorem For Empirical Transportation Cost In General Dimensions

Navid Ardeshir

Introduction: A fundamental question in the theory of optimal transport is to characterize a limiting distribution for transportation cost of an empirical distribution to some arbitrary target measure. More precisely, let us define $X_i \overset{\text{i.i.d}}{\sim} \mu \in \mathcal{P}(\mathbb{R}^d)$ for $1 \leq i \leq n$ (denote μ_n as the empirical measure) and a target measure $\nu \in \mathcal{P}(\mathbb{R}^d)$. We are interested in the rate of convergence of $\mathcal{W}_2(\mu_n, \nu)$ to $\mathcal{W}_2(\mu, \nu)$. Indeed much work has been devoted to provide results for the case when the target measure is the true underlying measure of samples (i.e. $\nu = \mu$). However, in this article we will paraphrase the work by [1] with a more statistical emphasis which yields meaningful results when $\nu \neq \mu$. Moreover, they demonstrate that the rate at which one obtains a non degenrate distribution is $\sqrt{n}$ (dimension-free) under mild assumptions and the limiting distribution is Gaussian. This is essentially different from the results obtained in [2] for 1-dimensional case for which they prove a non-gaussian limit.

Problem Formulation: We start by asserting the duality for 2-Wasserstein distance which holds true for any probability measures with finite second moments:

$$\mathcal{W}_2^2(\mu, \nu) \coloneqq \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \pi(dx, dy) = \max_{\varrho = \varrho^{**}} \int_{\mathbb{R}^d} (\|x\|^2 - 2\varrho(x))\mu(dx) + \int_{\mathbb{R}^d} (\|y\|^2 - 2\varrho^*(y))\nu(dy)$$

Where $\varrho^*(y) = \sup_{x \in \mathbb{R}^d} \langle y, x \rangle - \varrho(x)$ is the Fenchel dual of the potential function ϱ. Note that the constraint $\varrho = \varrho^{**}$ reads into ϱ being a convex and lower semi continuous function.
Due to Law of Large Numbers one can argue that the empirical measure converges weakly to the true underlying distribution as the sample size grows, i.e. $\mu_n \rightharpoonup \mu$. Therefore, by triangle inequality in wasserstein space we conclude $|\mathcal{W}_2(\mu_n, \nu) - \mathcal{W}_2(\mu, \nu)| \leq \mathcal{W}_2(\mu_n, \mu) \to 0$ since wasserstein distance metrizes weak convergence. In order to characterize the rate of convergence one needs to find sequences of real numbers $(a_n)_{n \geq 1}, (b_n)_{n \geq 1}$ such that the following quantity has a non-degenrate asymptotic distribution:

$$a_n(\mathcal{W}_2^2(\mu_n, \nu) - b_n) \rightharpoonup Z$$

A natural candidate to consider would be $b_n = \mathbb{E}[\mathcal{W}_2^2(\mu_n, \nu)]$ and $a_n = (\text{Var}(\mathcal{W}_2^2(\mu_n, \nu)))^{-1/2}$ to prove such results.

Theorem 1 (CLT)**.** *Suppose the following assumptions hold:*

1. *ν is absolutely continuous with respect to Lebesgue measure.*
2. *μ, ν have finite $4 + \delta$ moments for some $\delta > 0$.*
3. *ν has a positive density inside the convex hull of its support.*

Then, $n\,Var(\mathcal{W}_2^2(\mu_n, \nu)) \to \sigma^2(\mu, \nu) \coloneqq Var(\|X_1\|_2^2 - 2\varrho(X_1))$ where ϱ is the dual maximizer for transporting μ to ν. Moreover, the asymptotic distribution is a Gaussian:

$$\sqrt{n}(\mathcal{W}_2^2(\mu_n, \nu) - \mathbb{E}[\mathcal{W}_2^2(\mu_n, \nu)]) \rightharpoonup \mathcal{N}(0, \sigma^2(\mu, \nu))$$

Remark 1. *the asymptotic distribution might differ under different regularity assumptions on μ and ν. For instance, if one assumes μ, ν has a countable support then the final distribution will not be gaussian; Though, the rate of convergence is the same.*

Remark 2. *The first assumption is essential for existence of a monge map from the target measure to the empirical which allows a rich structure on the wasserstein distance. The second condition is conjectured to be not sharp and finite fourth moment should be sufficient. The third assumption is rather technical in order to ensure uniqueness (up to constants) of convex potentials.*

Proof Strategy: The main ideas of the proof can be divided into three main parts:

(a) **Variance Bounds:** Efron-Stein's inequality is a powerful tool in measure concentration which states variance is a quantity with tensorization property along the dimension. Using this inequality we find the rate at which $\text{Var}(\mathcal{W}_2^2(\mu_n, \nu))$ converges to zero.

(b) **Linearization:** Functional $\mu \mapsto \mathcal{W}_2^2(\mu, \nu)$ is a non-linear map and can be linearized (in the sense of Fréchet derivative) around a fix measure. However, in light of next step, one can ensure the order of approximation mathces with the intended asymptotic rate computed in the previous step, hence we can substitute wasserstein distance with it's first order approximation and obtain the same asymptotic behaviour.

(c) **Stability:** Optimal Transport maps (if exists) are in general not stable. However, due to the rich structure of $\mathcal{W}_2$ under mild assumptions one can exploit ideas from convex analysis and geometry to prove pointwise convergence for optimal potentials. For the sake of brevity, we are going to omit proofs regarding stability and mainly state pertitent results.

Theorem 2 (Efron-Stein's Variance Inequality)**.** *Let $(X_i)_{i\leq n} \in \mathbb{R}^d$ be i.i.d random variables with distribution μ and $(X_i^{'})_{i\leq n}$ be an independent copy. Furthermore, $f : \mathbb{R}^d \to \mathbb{R} \in \mathbf{L}_2(\mu)$ then:*

$$Var(f(X_1, ..., X_n)) \leq \sum_{i=1}^{n} \mathbb{E}[(f(X_1, ..., X_i^{'}, ..., X_n) - f(X_1, ..., X_n))_+^2]$$

where $(x)_+$ is zero whenever x is non-positive and equal to x otherwise.

The proof of this theorem is omitted and can be found in [4] but the gist of it is to generate a path from $(X_1, ..., X_n)$ to $(X_1^{'}, ..., X_n^{'})$ by swapping coordinates sequentially.

We may now start proving $\text{Var}(\mathcal{W}_2^2(\mu_n, \nu))$ is $O(\frac{1}{n})$ by considering $f(X_1, ..., X_n) = \mathcal{W}_2^2(\mu_n, \nu)$. Since $\nu \ll \mathcal{L}^d$ then due to Brenier's polar factorization theorem the optimal coupling between μ_n, ν is unique and supported on a cyclically monotone graph $\Gamma_n = \{(T_n(y), y) : y \in \mathbb{R}^d\}$ where $T_n = \nabla \varrho_n^*$ ν-a.e. for the optimal potential ϱ_n (differentiability is ensured ν-a.s.). Now, by considering a suboptimal transport where $T_n^{-1}(X_j)$ is transported to X_j for $j \neq i$ and $T_n^{-1}(X_i)$ transported to $X_i^{'}$ we may upper bound the cost of transportation from ν to the purturbed empirical measure:

$$\begin{aligned}
f(X_1, ..., X_i^{'}, ..., X_n) - f(X_1, ..., X_n) &\leq \int_{T_n^{-1}(X_i)} (\|y - X_i^{'}\|_2^2 - \|y - X_i\|_2^2)\nu(dy) \\
&\leq \|X_i^{'} - X_i\| \int_{T_n^{-1}(X_i)} (\|X_i\| + \|X_i^{'}\| + 2\|y\|)\nu(dy) \\
\text{(Using } (a+b)^2 \leq 2(a^2+b^2)) &\leq (\|X_i^{'} - X_i\|^2 \frac{4(\|X_i\|^2 + \|X_i^{'}\|^2)}{n^2} + 8\|X_i^{'} - X_i\|^2 (\int_{T_n^{-1}(X_i)} \|y\|\nu(dy))^2)^{\frac{1}{2}}
\end{aligned}$$

Note that we have used the fact that $\int_{T_n^{-1}(X_i)} 1\nu(dy) = \frac{1}{n}$ in the last inequality. Furthermore, using Holder's inequality combined with finite fourth moment asuumption:

$$\begin{aligned}
&\mathbb{E}[(f(X_1, ..., X_i^{'}, ..., X_n) - f(X_1, ..., X_n))_+^2] - \mathbb{E}[\|X_i^{'} - X_i\|^2 \frac{4\|X_i\|^2}{n^2}] \\
\text{(Using Cauchy Schwartz)} &\leq 8(\mathbb{E}[\|X_i^{'} - X_i\|^4]\mathbb{E}[(\int_{T_n^{-1}(X_i)} \|y\|\nu(dy))^4])^{\frac{1}{2}} \\
\text{(Using Holder Inequality)} &\leq 8(\mathbb{E}[\|X_i^{'} - X_i\|^4]\mathbb{E}[(\int_{T_n^{-1}(X_i)} 1\nu(dy))^3 \int_{T_n^{-1}(X_i)} \|y\|^4\nu(dy)])^{\frac{1}{2}} \\
&= (\frac{64}{n^3}\mathbb{E}[\|X_i^{'} - X_i\|^4]\mathbb{E}[\int_{T_n^{-1}(X_i)} \|y\|^4\nu(dy)])^{\frac{1}{2}} \\
\text{(By Symmetry)} &= (\frac{64}{n^4}\mathbb{E}[\|X_i^{'} - X_i\|^4] \int_{\mathbb{R}^d} \|y\|^4\nu(dy))^{\frac{1}{2}} = O(\frac{1}{n^2})
\end{aligned}$$

Combining these inequalities for $1 \leq i \leq n$ one can conclude the transportation cost has a variance of order $\frac{1}{n}$ which suggest (contingent on upper bound being sharp) the rate of convergence should be captured by $a_n = \sqrt{n}$. Note that derivations above can be repeated to prove $n(f(X_1, ..., X_i^{'}, ..., X_n) - f(X_1, ..., X_n))_+$ has uniformly finite $2 + \frac{\delta}{2}$ moment which implies the sequence is uniformly integrable by de la Vallée Poussin's theorem.

Let us now shift our focus to linearization and present the intuition. For simplicity if we restrict our attention into a compact space $\Omega \subset \mathbb{R}^d$ and let $\varphi^c(y) = \sup_{x \in \mathbb{R}^d} \|x - y\|_2^2 - \varphi(x)$. We then have $\mu \mapsto \mathcal{W}_2^2(\mu, \nu) = \sup\{\langle \varphi, \mu \rangle + \int \varphi^c d\nu : \varphi \in C^0(\Omega)\}$ is a lower semi-continuous convex functional over the domain $\mathcal{M}(\mathbb{R}^d)$ ($+\infty$-valued when $\mu \notin \mathcal{P}(\Omega)$) which can be viewed as the dual functional of $H(\varphi) = -\int \varphi^c d\nu$ over the Banach space $C^0(\Omega)$ (see section 7.2.1 in [3]). Therefore, it's subdifferential can be characterized by the set of functions φ that achieves the maximum. In other words, optimal potentials also play the role of first variations and if they are unique (up to additive constants) one expects a smooth behaviour locally (under the topolgy of weak convergence). Moreover, since we are on a compact set we may use Arzela-Ascoli's theorem to prove uniform convergence for optimal potentials (by selecting a potential that is zero at some specific point). Roughly speaking, if we let ϱ to be the unique optimal potential for transporting μ to ν then for $\hat{\mu} \in \mathcal{P}(\Omega)$ close enough to μ:

$$R_n = \mathcal{W}_2^2(\hat{\mu}, \nu) - \int (\|x\|^2 - 2\varrho(x))\hat{\mu}(dx) - \int (\|y\|^2 - 2\varrho^*)\nu(dy) \approx 0$$

Therefore, so long as the approximation error $R_n - \mathbb{E}[R_n]$ is $o_p(\frac{1}{\sqrt{n}})$ we may work with the linearization term instead. Using the standard CLT we have:

$$\sqrt{n} \int (\|x\|^2 - 2\varrho(x))(\mu_n(dx) - \mu(dx)) \rightharpoonup \mathcal{N}(0, \sigma^2(\mu, \nu))$$

Hence by Slutsky's theorem we obtain the desired central limit theorem. In order to show the approximation error is $o_p(\frac{1}{n})$ it is enough to show that $n\text{Var}(R_n - R_n^{'})_+ \to 0$ where $R_n^{'}$ is approximation error obtained with samples $(X_1^{'}, X_2, ..., X_n)$ by Efron-Stein's inequality. To that end, we use transport maps to embed the random variables in a more convenient space to prove $n^2(R_n - R_n^{'})_+^2 \overset{\text{a.s.}}{\to} 0$ and uniform integrability which implies $\mathbf{L}_1$ convergence. Note that, uniform integrability is implied since:

$$n(R_n - R_n^{'}) = \underbrace{n(\mathcal{W}_2^2(\mu_n, \nu) - \mathcal{W}_2^2(\mu_n^{'}, \nu))}_{\text{is U.I.}} - \underbrace{(\|X_1\|^2 - \varrho(X_1)) - (\|X_1^{'}\|^2 - \varrho(X_1^{'}))}_{\text{have finite second moment}}$$

Now, using Brenier's theorem we may assume $X_i = \nabla\varrho^*(Y_i)$ and $X_1^{'} = \nabla\varrho^*(Y_1^{'})$ where $(Y_i)_{i \leq n}, Y_1^{'} \overset{\text{i.i.d}}{\sim} \nu$. We then have:

$$\begin{aligned}
R_n - R_n^{'} &\leq R_n - \int (\|x\|^2 - 2\varrho_n(x))\mu_n^{'}(dx) - \int (\|y\|^2 - 2\varrho_n^*(y))\nu(dy) + \int (\|x\|^2 - 2\varrho)\mu_n^{'}(dx) \\
&= \int (\|x\|^2 - 2\varrho_n(x))(\mu_n(dx) - \mu_n^{'}(dx)) - \int (\|x\|^2 - 2\varrho)(\mu_n(dx) - \mu_n^{'}(dx)) \\
&= \frac{2}{n}(\int (\varrho - \varrho_n)d\mu_n - \int (\varrho - \varrho_n)d\mu_n^{'}) \\
&= \frac{2}{n}(\varrho(\nabla\varrho^*(Y_1)) - \varrho_n(\nabla\varrho^*(Y_1)) - \varrho(\nabla\varrho^*(Y_1^{'})) + \varrho_n(\nabla\varrho^*(Y_1^{'}))) \overset{?}{=} o(\frac{1}{n})
\end{aligned}$$

The proof would be essentially done if we knew pointwise convergence of optimal potentials, i.e. $\varrho_n \overset{\nu-\text{a.s.}}{\to} \varrho$ which was true if we were working on a compact space. However, in order to lift this assumption a different approach is required which we briefely mention and only state the main result from [1].

Theorem 3. *Suppose assumptions (1) and (3) in Theorem 1. holds. Let ϱ_n (ϱ resp.) be the optimal potential for transporting μ_n (μ resp.) to ν and we have $\mathcal{W}_2^2(\mu_n, \mu) \to 0$. Then, one can find suitable constants a_n such that if $\tilde{\varrho}(x) = \varrho(x) - a_n$ then:*

$$\lim_{n \to \infty} (\tilde{\varrho}_n^*(x), \nabla\tilde{\varrho}_n^*(x)) = (\varrho^*(x), \nabla\varrho^*(x)) \quad \textit{and} \quad \lim_{n \to \infty} \tilde{\varrho}_n(\nabla\varrho^*(x)) = \tilde{\varrho}(\nabla\varrho^*(x))$$

for a.e. x in the interior of the support of ν.

The proof idea is to exploit convexity of optimal potentials and work with a different notion of convergence over sets called graphical convergence (see [1] for definition) which preserves cyclically-monotonicity property (see Theorem 2.6 in [1]). It is well known that optimality of a transport plan can be characterized by cyclically monotonicity of it's support, hence this notion of convergence is natural to consider. More precisely, they show that the support of transport plans $\Gamma_n \subset \partial\varrho_n$ graphically converges to $\Gamma \subset \partial\varrho$ and uses properties of this convergence to show pointwise convergence.

References

[1] Del Barrio, E. and Loubes, J.M., 2019. Central limit theorems for empirical transportation cost in general dimension. The Annals of Probability, 47(2), pp.926-951.

[2] Del Barrio, E., Giné, E. and Matrán, C., 1999. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. Annals of Probability, pp.1009-1071.

[3] Santambrogio, F., 2015. Optimal transport for applied mathematicians. Birkäuser, NY, 55(58-63), p.94.

[4] van Handel, R., 2014. Probability in high dimension. PRINCETON UNIV NJ.