Influence de l'Echantillonage sur la Distance de Wasserstein Empirique

June 7, 2023

Semaine 5 du 7 Juin au 14 Juin

Soit $X_1, ..., X_n$ réalisation i.i.d d'une probabilité $\mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$. On note $\widehat{\mu}_1^n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ la réalisation empirique de μ_1. Cette semaine, on s'intéresse à la distance $W_2^2(\mu_1, \widehat{\mu}_1^n)$. Notamment, on sait que:

$$W_2^2(\mu_1, \widehat{\mu}_1^n) \underset{n\to\infty}{\to} 0 \tag{1}$$

Cette propriéte vient du faite que $\widehat{\mu}_1^n$ converge faiblement vers μ_1 et donc la distance de Wasserstein converge vers 0.

Soit $\epsilon > 0$, on va chercher à savoir le nombre n (c.a.d. la quantité de réalisation i.i.d) nécésaire à ce que:

$$W_2^2(\mu_1, \widehat{\mu}_1^n) < \epsilon \tag{2}$$

Cette quantité n dépent de plein de paramètres telque la dimension, l'existence de moment pour la distributions,...

De manière général, on cherche à écrire:

$$W_2^2(\widehat{\mu}_1^n, \mu_1) \leq \phi(n, d) \tag{3}$$

avec ϕ une fonction décroissante en n et qui tend vers 0 avec n.

On écrit $W_2^2(\widehat{\mu}_1^n, \mu_1) \lesssim \phi(n, d)$ quand il y a une constante $C > 0$ telque $W_2^2(\widehat{\mu}_1^n, \mu_1) \leq C\phi(n, d)$.

Finalement, on appel vitesse de convergence la quantité:

$$W_2^2(\widehat{\mu}_1^n, \mu_1) \tag{4}$$

et sample complexity la quantité:

$$|W_2^2(\widehat{\mu}_1^n, \widehat{\mu}_2^n) - W_2^2(\mu_1, \mu_2)| \tag{5}$$

où $\widehat{\mu}_2^n$ est une réalisation empirique de μ_2 avec n atomes.

Lecture.

- Partie 8 Introduction et partie 8.4.1 (Rates for OT) [PC+19], pour une vue d'ensemble du problème
- Vidéo sur la sample complexity `https://www.youtube.com/watch?v=glHy9z19mI0`
- On the rate of convergence in Wasserstein distance of the empirical measure [FG15]. Lire la partie 1.1, 1.2, 1.4 pour des résultats portant sur les moments des distributions
- Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance [WB19]. Lire la partie 1,2, 4.1, 4.3 et 6. Pour des résultats de convergence avec la dimension intrinsèque et sur la concentration des estimateurs.

Petit résumé de 1 page.
Les points clés sont:

- Estimateur plug-in
- Vitesse de convergence et sample complexity
- Concentration

Faire quelques expérimentations avec les Notebook

- Vérifier numériquement la vitesse de convergence sur le sampling de Gaussienne et de distributions uniforme
- Vérifier le cas particulier où les distributions sont Gaussiennes et la distance de Wasserstein est calculée avec la distance de Bures
- Vérifier numériquement la concentration

Points à réfléchir.

- Montrer que l'on peut avoir une borne sup pour la sample complexity si l'on a une borne sup pour la vitesse de convergence
- Faire un tableau à double entré qui calcule le n nécéssaire à ce que $|W_2^2(\widehat{\mu}_1^n, \widehat{\mu}_2^n) - W_2^2(\mu_1, \mu_2)| < \epsilon$, pour $\epsilon \in \{1, 1^{-1}, 1e^{-10}\}$ et $d \in \{2, 10, 100, 1000, 65536\}$
- De quoi dépendent les constantes C dans les équations de sample complexity?
- Faire un liste des hypothèses posées pour démontrer les résultats de sample complexity (existence de densité, distributions bornés, existence de moment)
- Les mesures Gaussiennes font elles parties des probabilités qui respectent les hypothèses? Et les mesures uniforme sur un ensemble compact?
- Quels autres paramètres que l'existence de moment et la dimension pourraient influencer la vitesse de convergence?

References

[FG15] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[WB19] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. 2019.