Influence de l'Echantillonage sur la Distance de Wasserstein Empirique

June 14, 2023

Semaine 6 du 14 Juin au 21 Juin

On a vu que la sample complexity de l'estimateur plug-in de la distance de Wasserstein dépendait fortement de la dimension [CRL+20]:

$$|\mathbb{E}W_2^2(\mu_n, \nu_n) - W_2^2(\mu, \nu)| \asymp n^{-2/d} \qquad d > 4 \tag{1}$$

avec $\mu_n = \frac{1}{n}\sum \delta_{x_i}$ (resp. $\nu_n = \frac{1}{n}\sum \delta_{y_i}$), $x_i \sim \mu$ i.i.d. (resp. $y_i \sim \nu$ i.i.d.).

Ce résultat est négatif car il indique que la convergence de $\mathbb{E}W_2^2(\mu_n, \nu_n)$ vers $W_2^2(\mu, \nu)$ dépent de façon exponentielle en la dimension. De plus cela ne peut pas être améliorer pour l'estimateur plug-in.

Un autre résultat important est que peu importe l'estimateur choisit pour estimer la distance, la sample complexity ne sera jamais mieux que:

$$(n \log n)^{-1/d} \tag{2}$$

Cette quantité s'appel une borne minimax.

Il reste intéressant de chercher le meilleur estimateur. On peut se tourner vers d'autres façon de sampler les distributions μ et ν.

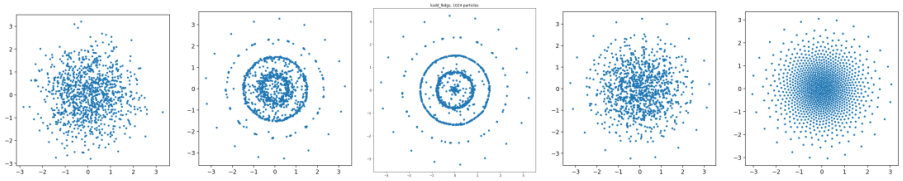

Figure 1: Différents Echantillonnage d'une Gaussienne [XKS22]

Dans cette figure on peut voir qu'il existe différentes mesures empiriques pour la distribution Gaussienne, notamment quand les atomes de la distrubtions empirique ne sont plus determinées de façon i.i.d.

Cette semaine nous allons voir le resultat de la borne minimax. Nous allons ensuite voir si la façon de sampler a un impact sur la sample complexity. On peut essayer de determiner de façon empirique une formule pour $|W_2^2(\tilde{\mu}_n, \tilde{\nu}_n)_2 - W_2^2(\mu, \nu)|$, avec $\tilde{\mu}_n$ et $\tilde{\nu}_n$ des distrubtions empiriques issues du sampling.
Lecture.

- Partie 7 jusqu'au theorem 12 (inclut) de [NWR22]

Faire quelques expérimentations avec les Notebook

- Quelle est la sample complexity observé empiriquement
- Quelle est la concentration observé empiriquement

Petit résumé de 1 page. Les points clés sont:

- Estimateur minimax
- Observation de l'effet du sampling sur l'estimation de $W_2^2(\mu, \nu)$

References

[CRL+20] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33:2257–2269, 2020.

[NWR22] Jonathan Niles-Weed and Philippe Rigollet. Estimation of wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688, 2022.

[XKS22] Lantian Xu, Anna Korba, and Dejan Slepcev. Accurate quantization of measures via interacting particle-based optimization. In *International Conference on Machine Learning*, pages 24576–24595. PMLR, 2022. https://github.com/xulant/accurate-quantization-and-nsvgd.