Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance

Kimia Nadjahi[1] Alain Durmus[2] Umut Şimşekli[1,3] Roland Badeau[1]

[1] Télécom Paris [2] ENS Paris-Saclay [3] University of Oxford

Minimum Distance Estimation

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{D}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}, \boldsymbol{\mu}_{\boldsymbol{\theta}})$$

$\mathbf{D}$: distance between distributions

$\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}$: empirical distribution of **data points** $Y_1, \ldots, Y_n$ i.i.d from $\mu_\star$

$\boldsymbol{\mu}_{\boldsymbol{\theta}}$: distribution parametrized by $\theta \in \Theta$

Minimum Distance Estimation

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{D}(\hat{\mu}_n, \mu_\theta)$$

$\mathbf{D}$: distance between distributions

$\hat{\mu}_n$: empirical distribution of **data points** $Y_1, \ldots, Y_n$ i.i.d from $\mu_\star$

μ_θ: distribution parametrized by $\theta \in \Theta$

Example: **Generative Modeling**

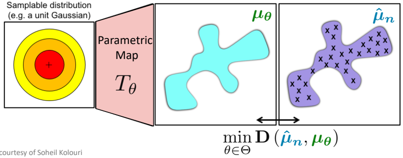


Figure courtesy of Soheil Kolouri

Minimum Expected Distance Estimation

Directly optimizing $\boldsymbol{\mu_\theta}$ is often **not possible** (e.g. GANs)

$$\hat{\theta}_{n,m} = \mathrm{argmin}_{\theta \in \Theta}\, \mathbb{E}\left[\mathbf{D}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta},\boldsymbol{m}}) \mid Y_{1:n}\right]$$

$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta},\boldsymbol{m}}$: empirical distribution of a sample $Z_1, \ldots, Z_m$ i.i.d. from $\boldsymbol{\mu_\theta}$

Optimal Transport

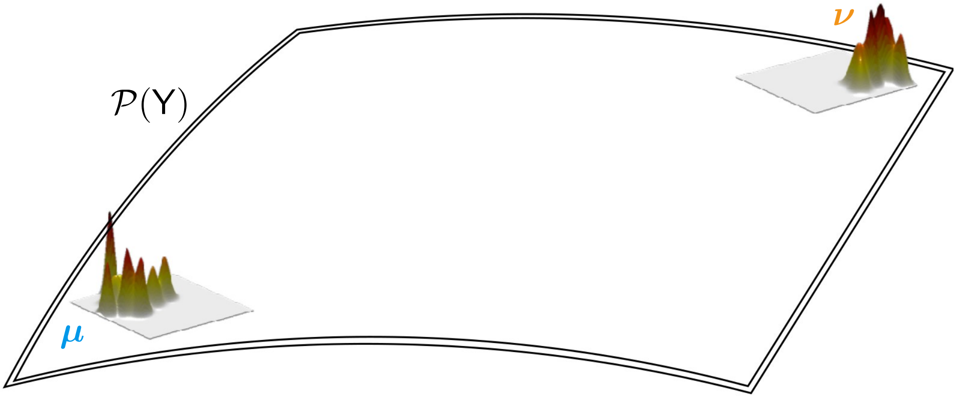


Figure courtesy of Marco Cuturi

Optimal Transport

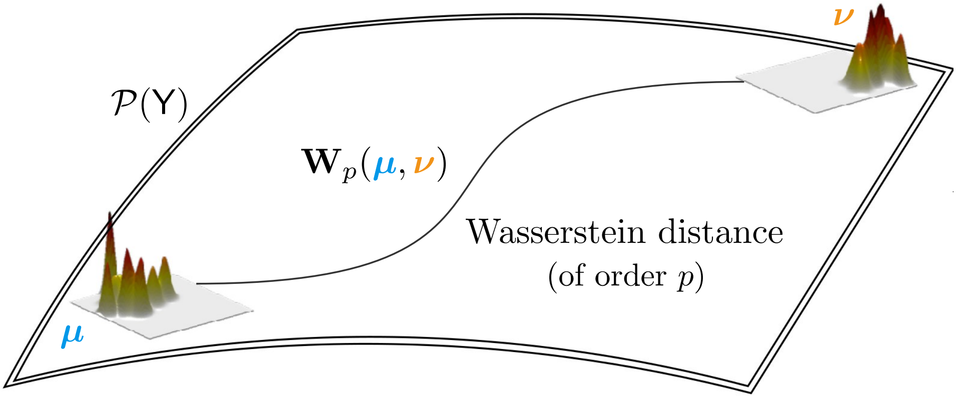


Figure courtesy of Marco Cuturi

Minimum Wasserstein Estimation

Choose $\mathbf{D} = \mathbf{W}_p$ (Wasserstein distance of order $p \geq 1$)

- ✓ Robust and increasingly popular estimators: Wasserstein GAN [1], Wasserstein auto-encoders [2]
- ✓ Asymptotic guarantees [3]

[1] Arjovsky et al., 2017 [2] Tolstikhin et al., 2018 [3] Bernton et al., 2019

Minimum Wasserstein Estimation

Choose $\mathbf{D} = \mathbf{W}_p$ (Wasserstein distance of order $p \geq 1$)

- ✓ Robust and increasingly popular estimators: Wasserstein GAN [1], Wasserstein auto-encoders [2]
- ✓ Asymptotic guarantees [3]

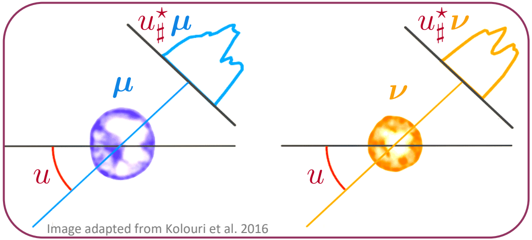
[1] Arjovsky et al., 2017 [2] Tolstikhin et al., 2018 [3] Bernton et al., 2019

- × $\mathbf{W}_p$: expensive + curse of dimensionality
- × Central limit theorem in [3] valid in 1D

Sliced-Wasserstein distance

In 1D, $\mathbf{W}_p$ has an analytical form $\Rightarrow$ Motivates a practical alternative:

$$\mathbf{SW}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} \mathbf{W}_p^p(u_\sharp^\star \mu, u_\sharp^\star \nu) \mathrm{d}\boldsymbol{\sigma}(u)$$



Minimum Sliced-Wasserstein Estimation

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$$
$$\hat{\theta}_{n,m} = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}\left[\mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \mid Y_{1:n}\right]$$

Successful in generative modeling applications:

SW GANs [1, 2], SW Autoencoders [2, 3], SW flows [4]

[1] Deshpande et al., CVPR 2018

[2] Wu et al., CVPR 2019

[3] Kolouri et al., ICLR 2019

[4] Liutkus et al., ICML 2019

Existence and consistency

There exists E with $\mathbb{P}(\mathsf{E}) = 1$ such that, for all $\omega \in \mathsf{E}$:

Existence

There exists $n(\omega)$ such that, for all $n \geq n(\omega)$, the set $\mathrm{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ is non-empty.

Consistency

$$\lim_{n \to +\infty} \inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta \in \Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta),$$

$$\limsup_{n \to +\infty} \operatorname*{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \subset \operatorname*{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta)$$

Central limit theorem

Suppose $\mu_\star = \mu_{\theta_\star}$

Under reasonably mild assumptions on the CDFs of the projected $\mu_{\theta_\star}$, $\hat{\mu}_n$,

$$\sqrt{\boldsymbol{n}} \inf_{\theta \in \Theta} \mathbf{SW}_1(\hat{\mu}_n, \mu_\theta) \xrightarrow{w} \inf_{\theta \in \Theta} \boldsymbol{H}(\theta) \;,$$
$$\sqrt{\boldsymbol{n}}(\hat{\theta}_n - \theta_\star) \xrightarrow{w} \operatorname*{argmin}_{\theta \in \Theta} \boldsymbol{H}(\theta) \qquad \text{as } n \to +\infty,$$

where $\boldsymbol{H}$ is a *random* map that corresponds to the limit (as $n \to +\infty$) of an approximation of $\mathbf{SW}_1(\hat{\mu}_n, \mu_\theta)$ near $\theta_\star$.

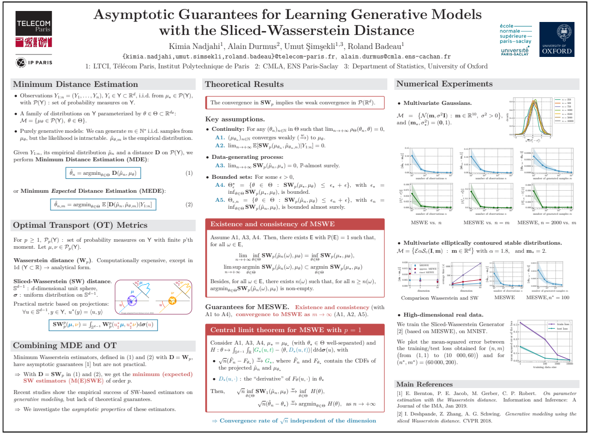
Summary

– Theoretical study of minimum Sliced-Wasserstein estimators:

- Convergence in $\mathbf{SW}_p \Rightarrow$ weak convergence of probability measures
- Existence and consistency of $\hat{\theta}_n$, $\hat{\theta}_{n,m}$
- Central limit theorem for $\hat{\theta}_n$: $\sqrt{n}$ convergence rate for any dimension

– Empirical confirmation on synthetical and real data

Thank you!



Also at **NeurIPS** on **Thu Dec 12th** (Spotlight presentation + Poster)