



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

Kimia Nadjahi¹ Alain Durmus² Umut Şimşekli^{1,3} Roland Badeau¹

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

Minimum Distance Estimation

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

\mathbf{D} : distance between distributions

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

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

Minimum Distance Estimation

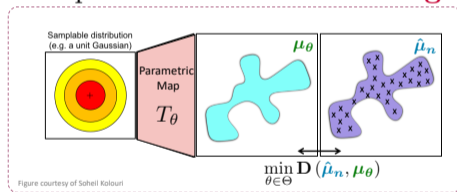
$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{D}(\hat{\mu}_n, \mu_\theta)$$

D: distance between distributions

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

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

Example: **Generative Modeling**



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

K. Nadjahi, A. Durmus, U. Şimşekli, R. Badeau

Minimum **Expected** Distance Estimation

Directly optimizing μ_θ is often **not possible** (e.g. GANs)

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

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

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]

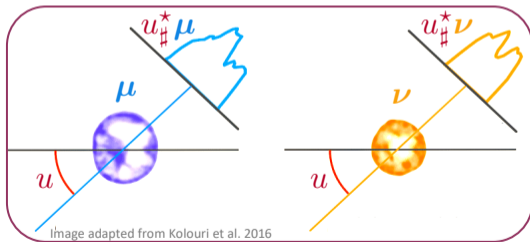
[1] Arjovsky et al., 2017 [2] Tolstikhin et al., 2018 [3] Bernton et al., 2019

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

Sliced-Wasserstein distance

In 1D, W_p has an analytical form \Rightarrow Motivates a practical alternative:

$$\text{SW}_p^p(\mu, \nu) = \int_{\mathbb{S}^{d-1}} W_p^p(u_{\#}^* \mu, u_{\#}^* \nu) d\sigma(u)$$



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

K. Nadjahi, A. Durmus, U. Şimşekli, R. Badeau

Minimum Sliced-Wasserstein Estimation

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

Successful in generative modeling applications (e.g., SW-GAN, Deshpande et al., 2018)

Minimum Sliced-Wasserstein Estimation

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

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

Successful in generative modeling applications (e.g., SW-GAN, Deshpande et al., 2018)

Our contributions:

- 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

Thank you!

Our Poster: East Exhibition Hall B + C #226

TELECOM
PARIS
UNIVERSITY OF OXFORD

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

Kimia Nadjahi¹, Alain Durmus², Umut Şimşekli^{1,3}, Roland Badeau¹
(kimia.nadjahi, umut.simsekli, roland.badeau@telecom-paris.fr, alain.durus@mla.ens-cachan.fr)
1: LTCI, Télécom Paris, Institut Polytechnique de Paris 2: CMLA, ENS Paris-Saclay 3: Department of Statistics, University of Oxford

Minimum Distance Estimation

- Observations $Y_{1:n} = (Y_1, \dots, Y_n)$, $Y_i \in \mathcal{Y} \subset \mathbb{R}^d$, i.i.d. from $\mu_\theta \in \mathcal{P}(\mathcal{Y})$, with $\mathcal{P}(\mathcal{Y})$: set of probability measures on \mathcal{Y} .
- A family of distributions on \mathcal{Y} parametrized by $\theta \in \Theta \subset \mathbb{R}^p$: $\mathcal{M} = \{\mu_\theta \in \mathcal{P}(\mathcal{Y}), \theta \in \Theta\}$.
- Purely generative model: We can generate $n \in \mathbb{N}$ i.i.d. samples from μ_θ , but the likelihood is intractable. $\hat{\mu}_{n,\theta}$ is the empirical distribution.

Given $Y_{1:n}$, its empirical distribution $\hat{\mu}_n$ and a distance \mathbf{D} on $\mathcal{P}(\mathcal{Y})$, we perform **Minimum Distance Estimation (MDE)**:

$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{D}(\hat{\mu}_n, \mu_\theta) \quad (1)$$

or **Minimum Expected Distance Estimation (MEDE)**:

$$\hat{\theta}_{n,\infty} = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}[\mathbf{D}(\hat{\mu}_n, \mu_\theta) | Y_{1:n}] \quad (2)$$

Theoretical Results

The convergence to \mathbf{SW}_p implies the weak convergence in $\mathcal{P}(\mathbb{R}^d)$.

Key assumptions.

- Continuity:** For any $(\theta_n)_n \in \Theta$ such that $\lim_{n \rightarrow \infty} \mu_\theta(\theta_n, \theta) = 0$, $\mathbf{A1}$. $\{\mu_\theta\}_{\theta \in \Theta}$ converge weakly $\left(\frac{1}{n}\right)$ to μ_θ .
- $\mathbf{A2}$. $\lim_{n \rightarrow \infty} \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)] = \mathbf{SW}_p(\mu_\theta, \mu_\theta) = 0$.
- Data-generating process:** $\mathbf{A3}$. $\lim_{n \rightarrow \infty} \mathbb{P}(\mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) = 0, \mathcal{P}\text{-almost surely})$.
- Bounded sets:** For some $\epsilon > 0$, $\mathbf{A4}$. $\Theta_\epsilon = \{\theta \in \Theta : \mathbf{SW}_p(\mu_\theta, \mu_\theta) \leq \epsilon + \epsilon\}$, with $\epsilon_n = \inf_{\theta \in \Theta} \mathbf{SW}_p(\mu_\theta, \mu_\theta)$ is bounded.
- $\mathbf{A5}$. $\Theta_{\epsilon_n} = \{\theta \in \Theta : \mathbf{SW}_p(\mu_\theta, \mu_\theta) \leq \epsilon_n + \epsilon\}$, with $\epsilon_n = \inf_{\theta \in \Theta} \mathbf{SW}_p(\mu_\theta, \mu_\theta)$ is bounded almost surely.

Numerical Experiments

- Multivariate Gaussians.** $\mathcal{M} = \{N(\mu, \Sigma) : \mu \in \mathbb{R}^2, \Sigma \succ 0\}$, and $(\mu_n, \Sigma_n) = (0, 1)$.

Comparison Wasserstein and SW MSWE MSWE, $\alpha = 100$

Optimal Transport (OT) Metrics

For $p \geq 1$, $\mathcal{P}_p(\mathcal{Y})$: set of probability measures on \mathcal{Y} with finite p th moment. Let $\mu, \nu \in \mathcal{P}_p(\mathcal{Y})$.

Wasserstein distance (W_p). Computationally expensive, except in \mathbb{R}^d ($\mathcal{Y} \subset \mathbb{R}$) - analytical form.

Sliced-Wasserstein (SW) distances. \mathbb{R}^{2d-1} : d -dimensional unit sphere, σ : uniform distribution on \mathbb{S}^{2d-1} .

Practical metric based on projection: $\forall \sigma \in \mathbb{S}^{2d-1}, \mu \in \mathcal{P}_p(\mathbb{R}^d), \nu \in \mathcal{P}_p(\mathbb{R}^d)$

$$\mathbf{SW}_p^2(\mu, \nu) = \int_{\mathbb{S}^{2d-1}} W_p^2(\mu_\sigma, \nu_\sigma) d\sigma$$

Existence and consistency of MSWE

Assume $\mathbf{A1}, \mathbf{A3}, \mathbf{A4}$. Then, there exists \mathbb{E} with $\mathbb{P}(\mathbb{E}) = 1$ such that, for all $\omega \in \mathbb{E}$,

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

But $\operatorname{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) \subset \operatorname{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\mu_\omega, \mu_\theta)$

Holds, for all $\omega \in \mathbb{E}$, there exists $\theta(\omega)$ such that, for all $n \geq n(\omega)$, $\operatorname{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ is non-empty.

Combining MDE and OT

Minimum Wasserstein estimators, defined in (1) and (2) with $\mathbf{D} = W_p$, have asymptotic guarantees [5] but are not practical.

\Rightarrow With $\mathbf{D} = \mathbf{SW}_p$ in (1) and (2), we get the **minimum (expected) SW estimators (MDE/SWE)** of order p .

Recent studies show the empirical success of SW-based estimators on generative modeling, but lack of theoretical guarantees.

\Rightarrow We investigate the asymptotic properties of these estimators.

Central limit theorem for MSWE with $p = 1$

Consider $\mathbf{A1}, \mathbf{A3}, \mathbf{A4}, \mathbf{A5}$ (with $\theta_n \in \Theta$ self-separated) and $H: \theta \mapsto \int_{\mathbb{S}^{2d-1}} \int_{\mathbb{R}^d} H(\omega, \tau) |D_\tau \mathbf{D}_p(\omega, \tau)| d\sigma d\tau$, with

- $\sqrt{n}(\hat{\mu}_n - \mu_\theta) \xrightarrow{D} G$, where G_τ and P_τ contain the CDFs of the projected $\hat{\mu}_n$ and μ_θ .
- $D_\tau(\mu_\theta, \cdot)$ the 'derivative' of $\mathbf{D}_p(\omega, \tau)$ in θ .

Then, $\sqrt{n} \inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) \xrightarrow{D} \inf_{\theta \in \Theta} H(\theta)$.

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{D} \operatorname{argmin}_{\theta \in \Theta} H(\theta), \text{ as } n \rightarrow \infty$$

\Rightarrow Convergence rate of \sqrt{n} independent of the dimension

Main References

- [1] E. Bertrand, P. E. Jacob, M. Gerber, C. P. Robert. On parameter estimation with the Wasserstein distance. Information and Inference: A Journal of the IMA, Jan 2019.
- [2] I. Drouot, Z. Zhang, A. G. Schott. Generative modeling using the sliced Wasserstein distance. CVPR 2018.