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

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

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

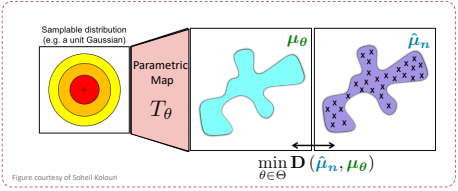
$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{D}(\hat{\boldsymbol{\mu}}_{\boldsymbol{n}}, \boldsymbol{\mu}_{\boldsymbol{\theta}})$$

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 $\theta \in \Theta$

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

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

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

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]

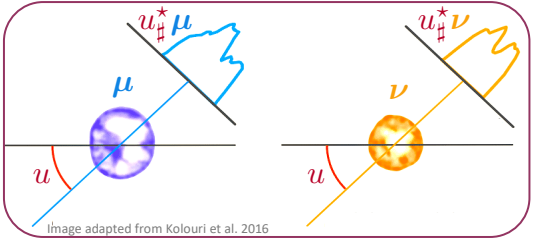
[1] Arjovsky et al., 2017 [2] Tolstikhin et al., 2018 [3] Bernton et al., 2019

- × $\mathbf{W}_p$: 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:

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



Asymptotic Guarantees for Learning Generative Models with the Sliced-Wasserstein Distance.
K. Nadjahi, A. Durmus, U. Şimşekli, R. Badeau

Minimum Sliced-Wasserstein Estimation

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

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

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

Minimum Sliced-Wasserstein Estimation

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

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

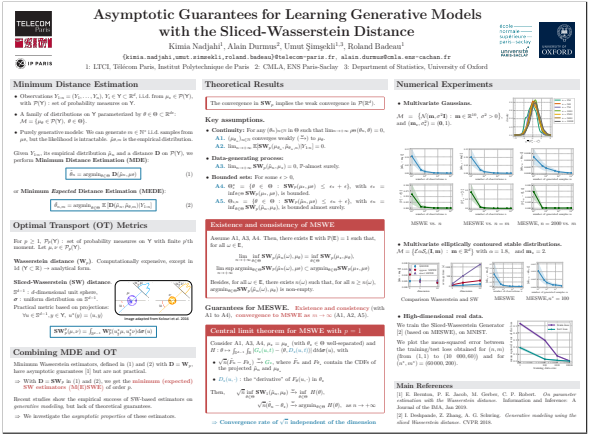
Successful in generative modeling applications (e.g., SW-GAN, Deshpande et al., 2018)

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



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

Kimia Nadjahi[1], Alain Durmus[2], Umut Şimşekli[1,3], Roland Badeau[1]
{kimia.nadjahi,umut.simsekli,roland.badeau}@telecom-paris.fr, alain.durmus@cmla.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 $Y_{1:n} = (Y_1, \dots, Y_n), Y_i \in Y \subset \mathbb{R}^d$, i.i.d. from $\mu_\star \in \mathcal{P}(Y)$, with $\mathcal{P}(Y)$: set of probability measures on Y.
- A family of distributions on Y parameterized by $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$: $M = \{\mu_\theta \in \mathcal{P}(Y), \theta \in \Theta\}$.
- Purely generative model: We can generate $m \in \mathbb{N}^*$ i.i.d. samples from μ_θ, but the likelihood is intractable. $\hat{\mu}_{\theta,m}$ is the empirical distribution.

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

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

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

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

Optimal Transport (OT) Metrics

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

Wasserstein distance ($\mathbf{W}_p$) : Computationally expensive, except in 1d ($Y \subset \mathbb{R}$) $\rightarrow$ analytical form.

Sliced-Wasserstein (SW) distance

$\mathbb{S}^{d-1}$: d-dimensional unit sphere,
σ : uniform distribution on $\mathbb{S}^{d-1}$.
Practical metric based on projections:
$\forall u \in \mathbb{S}^{d-1}, g \in Y,\ u^*(g) = \langle u, g \rangle$

Image adapted from Kolouri et al. 2016

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

Combining MDE and OT

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

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

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

$\Rightarrow$ We investigate the *asymptotic properties* of these estimators.

Theoretical Results

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

Key assumptions.

- **Continuity:** For any $(\theta_n)_{n\in\mathbb{N}}$ in Θ such that $\lim_{n\to+\infty} \rho_\Theta(\theta_n, \theta) = 0$,
  A1. $(\mu_{\theta_n})_{n\in\mathbb{N}}$ converges weakly ($\xrightarrow{w}$) to μ_θ.
  A2. $\lim_{n\to+\infty} \mathbb{E}[\mathbf{SW}_p(\hat{\mu}_{\theta_n,m}, \hat{\mu}_{\theta,m})^p|_{Y_{1:n}}] = 0$.
- **Data-generating process:**
  A3. $\lim_{n\to+\infty} \mathbf{SW}_p(\hat{\mu}_n, \mu_\star) = 0$, $\mathbb{P}$-almost surely.
- **Bounded sets:** For some $\epsilon > 0$,
  A4. $\Theta_\epsilon^\star = \{\theta \in \Theta : \mathbf{SW}_p(\mu_\theta, \mu_\star) \leq \epsilon_0 + \epsilon\}$, with $\epsilon_0 = \inf_{\theta\in\Theta} \mathbf{SW}_p(\mu_\theta, \mu_\star)$, is bounded.
  A5. $\Theta_{\epsilon,m} = \{\theta \in \Theta : \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) \leq \epsilon_0 + \epsilon\}$, with $\epsilon_0 = \inf_{\theta\in\Theta} \mathbf{SW}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$, is bounded almost surely.

Existence and consistency of MSWE

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

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

Besides, for all $\omega \in \mathbb{E}$, there exists $n(\omega)$ 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.

Guarantees for MESWE. Existence and consistency (with A1 to A4), convergence to MSWE as $m \to \infty$ (A1, A2, A5).

Central limit theorem for MSWE with $p = 1$

Consider A1, A3, A4, $\mu_\star = \mu_{\theta_\star}$ (with $\theta_\star \in \Theta$ well-separated) and $H : \theta \mapsto \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} |U_\theta(z,u) - (\theta, P_{\beta_\star}(z))| \mathrm{d}z \mathrm{d}\sigma(u)$, with

- $\sqrt{m}(\hat{F}_\theta - F_{\beta_\star}) \xrightarrow{w} G_\theta$, where $\hat{F}_\theta$ and P_θ contain the CDFs of the projected $\hat{\mu}_\theta$ and μ_θ.
- $U_\theta|_{\theta:\ 1}$: the "derivative" of $F_\theta(u,\cdot)$ in $\theta_\star$

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

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

Numerical Experiments

- **Multivariate Gaussians.**

$M = \{\mathcal{N}(\mathbf{m}, \sigma^2\mathbf{I}) : \mathbf{m} \in \mathbb{R}^{50},\ \sigma^2 > 0\}$, and $(\mathbf{m}_\star, \sigma_\star^2) = (\mathbf{0}, 1)$.

MSWE vs. n MSWE vs. $n = m$ MSWE, $n = 2000$ vs. m

- **Multivariate elliptically contoured stable distributions.**

$M = \{\mathrm{ECS}_\alpha(\mathbf{I}, \mathbf{m}) : \mathbf{m} \in \mathbb{R}^d\}$ with $\alpha = 1.8$, and $\mathbf{m}_\star = 2$.

Comparison Wasserstein and SW MSWE MSWE vs. $n = 100$

- **High-dimensional real data.**

We train the Sliced-Wasserstein Generator [2] (based on MESWE), on MNIST.

We plot the mean-squared error between the training/test loss obtained for (n, m) (from (1,1) to (10 000,600)) and for $(n^\star, m^\star) = (60000, 200)$.

Main References

[1] E. Bernton, 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. Deshpande, Z. Zhang, A. G. Schwing. *Generative modeling using the sliced Wasserstein distance*. CVPR 2018.