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 $Y_{1:n} = (Y_1, \dots, Y_n)$, $Y_i \in \mathsf{Y} \subset \mathbb{R}^d$, i.i.d. from $\mu_\star \in \mathcal{P}(\mathsf{Y})$, with $\mathcal{P}(\mathsf{Y})$: set of probability measures on Y.
- A family of distributions on Y parameterized by $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$: $\mathcal{M} = \{\mu_\theta \in \mathcal{P}(\mathsf{Y}),\ \theta \in \Theta\}$.
- Purely generative models: 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 $\mathcal{P}(\mathsf{Y})$, we perform **Minimum Distance Estimation (MDE)**:

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

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

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

Optimal Transport (OT) Metrics

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

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

Sliced-Wasserstein (SW) distance.

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

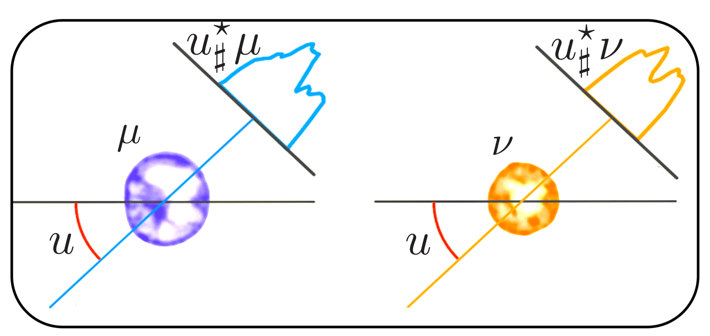


Image adapted from Kolouri et al. 2016

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

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 modeling*, 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. $\lim_{n \to +\infty} \mathbb{E}[\mathbf{SW}_p(\mu_{\theta_n}, \hat{\mu}_{\theta_n, n}) | Y_{1:n}] = 0$.
- **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. $\Theta^\star_\epsilon = \{\theta \in \Theta : \mathbf{SW}_p(\mu_\star, \mu_\theta) \leq \epsilon_\star + \epsilon\}$, with $\epsilon_\star = \inf_{\theta \in \Theta} \mathbf{SW}_p(\mu_\star, \mu_\theta)$, is bounded.
  A5. $\Theta_{\epsilon,n} = \{\theta \in \Theta : \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta) \leq \epsilon_n + \epsilon\}$, with $\epsilon_n = \inf_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n, \mu_\theta)$, is bounded almost surely.

Existence and consistency of MSWE

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

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

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

Besides, for all $\omega \in \mathsf{E}$, there exists $n(\omega)$ such that, for all $n \geq n(\omega)$, $\mathrm{argmin}_{\theta \in \Theta} \mathbf{SW}_p(\hat{\mu}_n(\omega), \mu_\theta)$ 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 $H : \theta \mapsto \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} |G_\star(u,t) - \langle \theta, D_\star(u,t) \rangle| \,\mathrm{d}t \mathrm{d}\boldsymbol{\sigma}(u)$, with

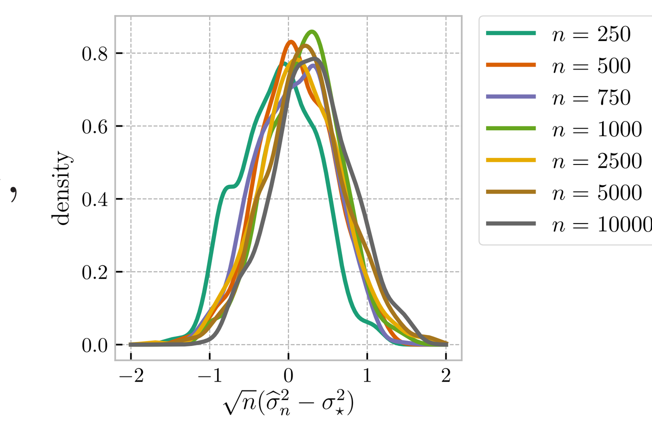
- $\sqrt{n}(\hat{F}_n - F_{\theta_\star}) \xrightarrow{w} G_\star$, where $\hat{F}_n$ and $F_{\theta_\star}$ contain the CDFs of the projected $\hat{\mu}_n$ and $\mu_{\theta_\star}$
- $D_\star(u, \cdot)$: the "derivative" of $F_\theta(u, \cdot)$ in $\theta_\star$

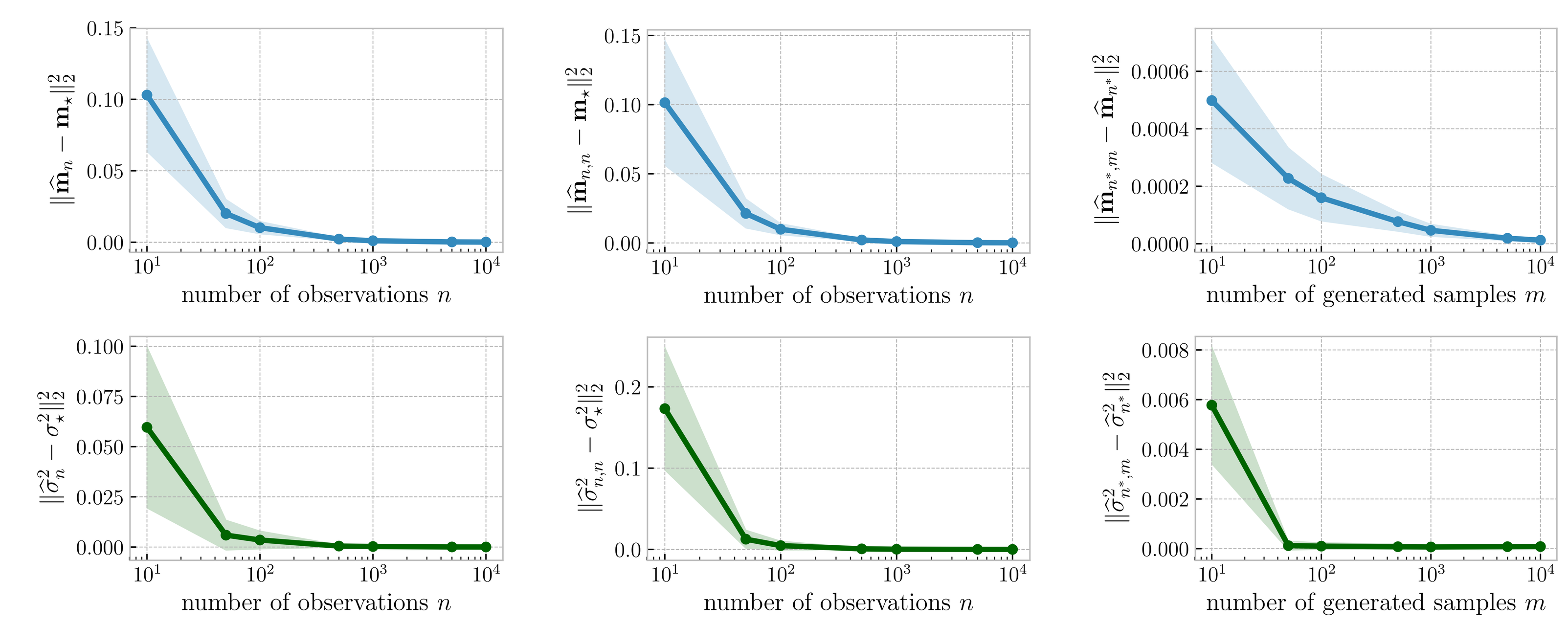
Then,
$$\sqrt{n} \inf_{\theta \in \Theta} \mathbf{SW}_1(\hat{\mu}_n, \mu_\theta) \xrightarrow{w} \inf_{\theta \in \Theta} H(\theta),$$
$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{w} \mathrm{argmin}_{\theta \in \Theta} H(\theta), \quad \text{as } n \to +\infty$$

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

Numerical Experiments

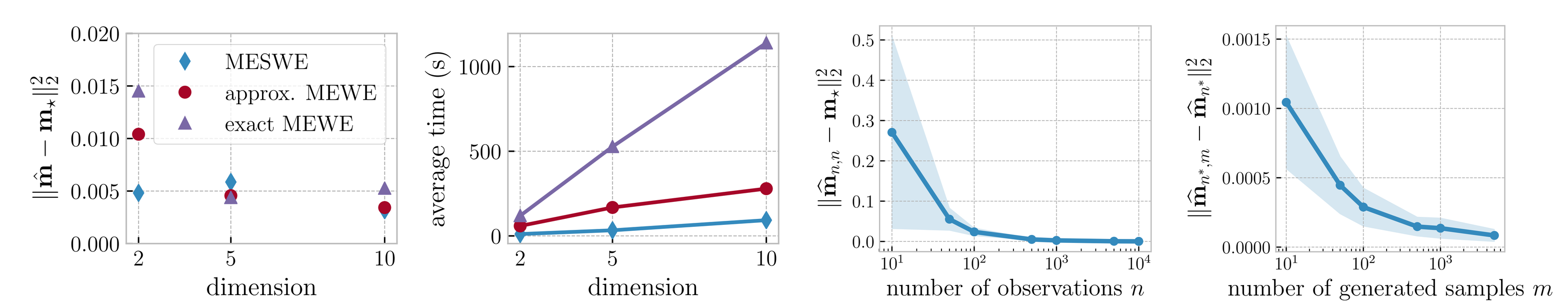
- **Multivariate Gaussians.**
$\mathcal{M} = \{\mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{I}) : \mathbf{m} \in \mathbb{R}^{10},\ \sigma^2 > 0\}$, and $(\mathbf{m}_\star, \sigma^2_\star) = (\mathbf{0}, 1)$.





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

- **Multivariate elliptically contoured stable distributions.**
$\mathcal{M} = \{\mathcal{E}\alpha\mathcal{S}_c(\mathbf{I}, \mathbf{m}) : \mathbf{m} \in \mathbb{R}^d\}$ with $\alpha = 1.8$, and $\mathbf{m}_\star = 2$.

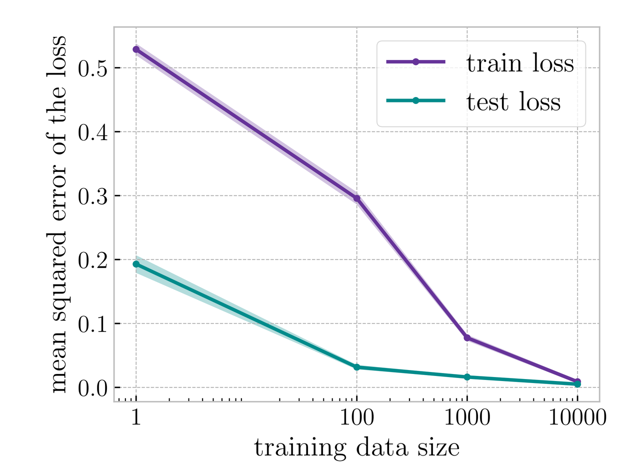


Comparison Wasserstein and SW MESWE MESWE, $n^* = 100$

- **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 $(10\ 000, 60)$) and for $(n^*, m^*) = (60\,000, 200)$.



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.