



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

Kimia Nadjahi<sup>1</sup>   Alain Durmus<sup>2</sup>   Umut Şimşekli<sup>1,3</sup>   Roland Badeau<sup>1</sup>

<sup>1</sup> Télécom Paris   <sup>2</sup> ENS Paris-Saclay   <sup>3</sup> 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  $\mu_\star$

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

# Minimum Distance Estimation

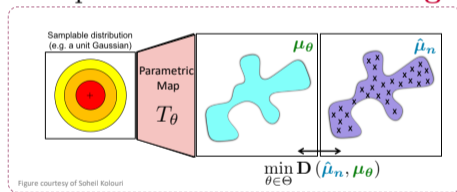
$$\hat{\theta}_n = \operatorname{argmin}_{\theta \in \Theta} \mathbf{D}(\hat{\mu}_n, \mu_\theta)$$

**D**: distance between distributions

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

$\mu_\theta$ : 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  $\mu_\theta$  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  $\mu_\theta$

# Optimal Transport

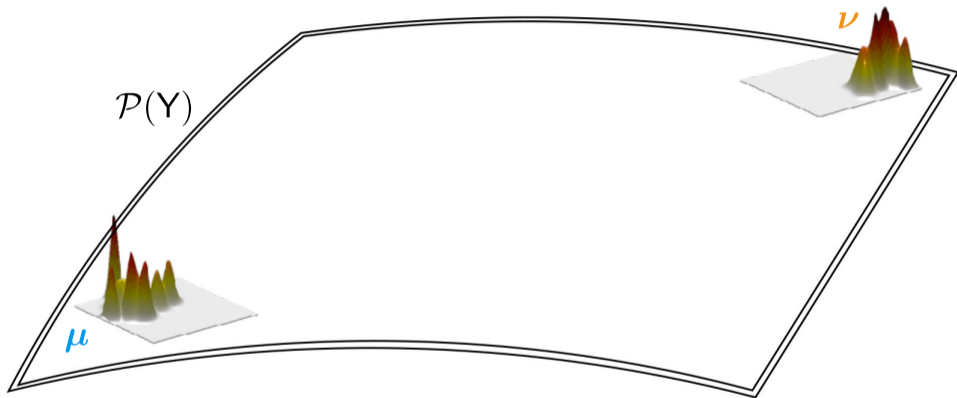


Figure courtesy of Marco Cuturi

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

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

# Optimal Transport

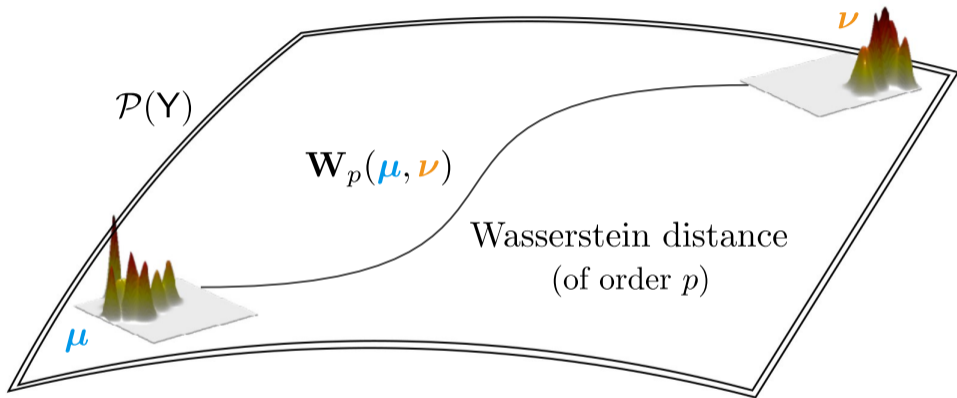


Figure courtesy of Marco Cuturi

# 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

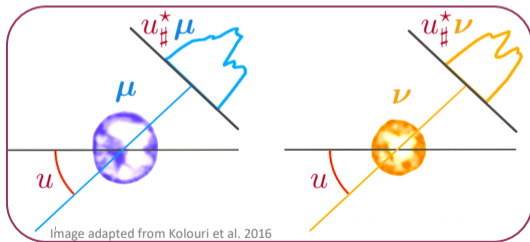
- ✗  $\mathbf{W}_p$ : expensive + curse of dimensionality
- ✗ Central limit theorem in [3] valid in 1D



# Sliced-Wasserstein distance

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

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



# 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:

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

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

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

# Existence and consistency

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

## Existence

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

## Consistency

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

# Central limit theorem

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

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


$$\begin{aligned}\sqrt{n} \inf_{\theta \in \Theta} \mathbf{SW}_1(\hat{\mu}_n, \mu_\theta) &\xrightarrow{w} \inf_{\theta \in \Theta} \mathbf{H}(\theta), \\ \sqrt{n}(\hat{\theta}_n - \theta_\star) &\xrightarrow{w} \operatorname{argmin}_{\theta \in \Theta} \mathbf{H}(\theta) \quad \text{as } n \rightarrow +\infty,\end{aligned}$$

where  $\mathbf{H}$  is a *random* map that corresponds to the limit (as  $n \rightarrow +\infty$ ) 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!


  
IP PARIS

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

Kimia Nadjahi<sup>1</sup>, Alain Durmus<sup>2</sup>, Umut Şimşekli<sup>1,3</sup>, Roland Badeau<sup>1</sup>

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

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

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

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

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

### Theoretical Results

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

#### Key assumptions.

- Continuity:** For any  $(\theta_n)_{n \in \mathbb{N}} \in \Theta$  such that  $\lim_{n \rightarrow \infty} \mu_n(\theta_n, \theta) = 0$ .
  - A1.  $(\mu_n)_{n \in \mathbb{N}}$  converges weakly to  $\mu_\theta$ .
  - A2.  $\lim_{n \rightarrow \infty} \mathbb{E} \operatorname{SW}_p(\hat{\mu}_n, \mu_{\theta_n} | Y_{1:n}) = 0$ .
- Data-generating processes:**
  - A3.  $\lim_{n \rightarrow \infty} \mathbb{P}(\hat{\mu}_n \in \Theta) = 0$ ,  $\mathbb{P}$ -almost surely.
- Bounded sets:** For some  $\epsilon > 0$ .
  - A4.  $\Theta'_\epsilon = \{\theta \in \Theta : \operatorname{SW}_p(\mu_n, \mu_\theta) \leq \epsilon_n + \epsilon\}$ , with  $\epsilon_n = \inf_{\theta \in \Theta} \operatorname{SW}_p(\mu_n, \mu_\theta)$ , is bounded.
  - A5.  $\Theta_{\epsilon, \infty} = \{\theta \in \Theta : \operatorname{SW}_p(\mu_n, \mu_\theta) \leq \epsilon_n + \epsilon\}$ , with  $\epsilon_n = \inf_{\theta \in \Theta} \operatorname{SW}_p(\mu_n, \mu_\theta)$ , 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  $\omega \in \mathbb{E}$ ,

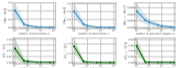

$$\lim_{n \rightarrow \infty} \inf_{\theta \in \Theta} \operatorname{SW}_p(\hat{\mu}_n(\omega), \mu_\theta) = \inf_{\theta \in \Theta} \operatorname{SW}_p(\mu_n, \mu_\theta)$$

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

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

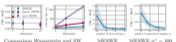
### Numerical Experiments

- Multivariate Gaussians.**  
 $M = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}^d, \sigma^2 > 0\}$ , and  $(\mu, \sigma^2) \in (0, 1)$ .




MSWE vs.  $n$       MSWE vs.  $n, \sigma^2$       MSWE,  $n = 2000$  vs.  $\sigma^2$

- Multivariate elliptically contoured stable distributions.**  
 $M = \{L(\alpha, \beta, \mu) : \alpha \in \mathbb{R}^d, \beta \in \mathbb{R}^d, \mu \in \mathbb{R}^d, \alpha \neq 0, \beta \neq 0\}$ .



Comparison Wasserstein and SW      MSWE      MSWE,  $\sigma^2 = 100$

- 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  $(n, n)$  from (1.1) to  $(10^4, 10^4)$  and for  $(n^*, n^*) = (50000, 200)$ .


---

### 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 ( $W_p$ ).** Computationally expensive, except in  $d (Y \subset \mathbb{R}) \rightarrow$  analytical form.

**Sliced-Wasserstein (SW) distance.**  $\mathbb{R}^{2d-1}$ :  $d$ -dimensional unit sphere,  $\sigma^2$ : uniform distribution on  $\mathbb{S}^{2d-1}$ .  
Practical metric based on projection:  
 $Y \in \mathbb{R}^{2d-1}, y \in Y, u^*(y) = (y, y)$


$$\operatorname{SW}_p(\mu, \nu) = \int_{\mathbb{S}^{2d-1}} W_p(\mu|_{u^*(\cdot)}, \nu|_{u^*(\cdot)}) d\sigma(u)$$

### Guarantees for MESWE: Existence and consistency (with A1 to A5), convergence to MESWE as $n \rightarrow \infty$ (A1, A2, A5).

#### Central limit theorem for MSWE with $p = 1$

Consider A1, A3, A4,  $\mu_n = \mu_\theta$  (with  $\theta_n \in \Theta$  well-separated) and  $H: \theta \mapsto \int_{\mathbb{S}^{2d-1}} \int_Y F_2(y, \tau) (D_\theta D_{\theta_n}(y, \tau)) d\sigma(u) d\mu(y)$ , with

- $\sqrt{n}(F_n - F_\theta) \xrightarrow{d} G_\theta$ , where  $F_n$  and  $F_\theta$  contain the CDFs of the projected  $\hat{\mu}_n$  and  $\mu_\theta$ .
- $D_\theta(y, \cdot)$ : the "derivative" of  $F_2(y, \cdot)$  in  $\theta$ .

Then,  $\sqrt{n} \int_{\mathbb{S}^{2d-1}} \int_Y \operatorname{SW}_p(\hat{\mu}_n, \mu_\theta) d\sigma(u) d\mu(y) \xrightarrow{d} \int_{\mathbb{S}^{2d-1}} \int_Y \operatorname{SW}_p(\theta_n - \theta_n) d\sigma(u) d\mu(y)$ , as  $n \rightarrow +\infty$ .

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

### Combining MDE and OT

Minimum Wasserstein estimators, defined in (1) and (2) with  $D = W_p$ , have asymptotic guarantee [2] but are not practical.

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

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

- We investigate the asymptotic properties of these estimators.

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