Properties of Discrete Sliced Wasserstein Losses

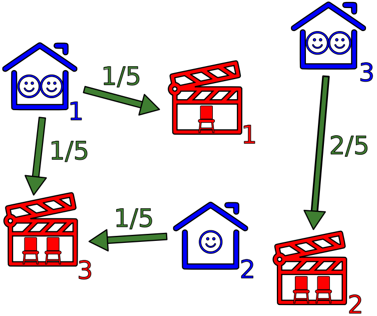
Eloi Tanguy, Rémi Flamary, Julie Delon

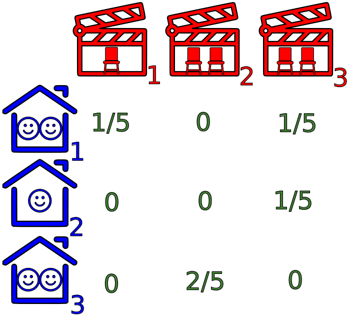
MAP5, Université Paris-Cité

February 2024

1. The Discrete Sliced Wasserstein Distance
2. Optimisation Properties
3. SGD Convergence
4. SGD for Training SW Neural Networks

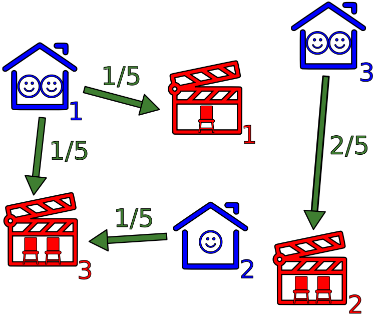
Discrete Optimal Transport

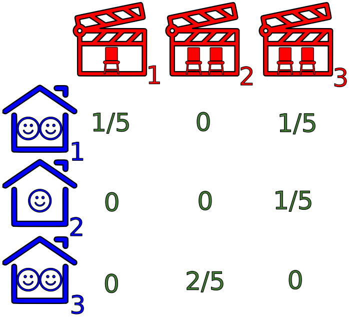




	1	2	3
1	1/5	0	1/5
2	0	0	1/5
3	0	2/5	0

Discrete Optimal Transport





	1	2	3
1	1/5	0	1/5
2	0	0	1/5
3	0	2/5	0

Assignment Cost:

$$\frac{1}{5} \times c(x_1, y_1) + \frac{1}{5} \times c(x_1, y_3) + \frac{1}{5} \times c(x_2, y_3) + \frac{2}{5} \times c(x_3, y_2).$$

Constraints on $\pi \in \mathbb{R}_+^{3\times 3}$: $\pi \mathbf{1} = (2/5, 1/5, 2/5)$, $\pi^\top \mathbf{1} = (1/5, 2/5, 2/5)$.

$$\text{Optimal Transport Cost}: \min_{\pi} \sum_{i,j} c(x_i, y_j)\pi_{i,j}.$$

2-Wasserstein Distance: $c(x, y) = \|x - y\|_2^2$

Measures $\mu = \frac{1}{n}\sum_{i=1}^{n} a_i \delta_{x_i}, \ \nu = \frac{1}{m}\sum_{j=1}^{m} b_j \delta_{y_j}.$

$$\mathrm{W}_2^2(\mu, \nu) = \min_{\substack{\pi \in \mathbb{R}_+^{n \times m} \\ \pi \mathbf{1} = a,\ \pi^\top \mathbf{1} = b}} \sum_{i=1}^{n}\sum_{j=1}^{m} \|x_i - y_j\|_2^2 \pi_{i,j}.$$

2-Wasserstein Distance: $c(x, y) = \|x - y\|_2^2$

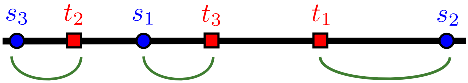
Measures $\mu = \frac{1}{n}\sum_{i=1}^{n} a_i \delta_{x_i}, \ \nu = \frac{1}{m}\sum_{j=1}^{m} b_j \delta_{y_j}$.

$$\mathrm{W}_2^2(\mu, \nu) = \min_{\substack{\pi \in \mathbb{R}_+^{n\times m} \\ \pi \mathbf{1} = a, \ \pi^\top \mathbf{1} = b}} \sum_{i=1}^{n}\sum_{j=1}^{m} \|x_i - y_j\|_2^2 \pi_{i,j}.$$

Continuous case: $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

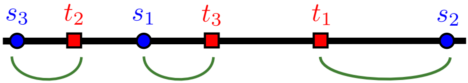
$$\mathrm{W}_2^2(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^{2d}} \|x - y\|_2^2 \mathrm{d}\pi(x, y) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\pi}\left[\|X - Y\|_2^2\right].$$

1D Wasserstein and Sliced Wasserstein

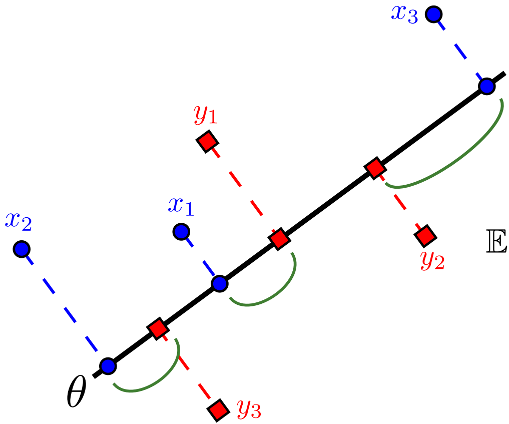


$$\mathrm{W}_2^2(\gamma_S, \gamma_T) = \frac{1}{n}\sum_{i=1}^{n} \left|s_{\sigma(i)} - t_{\tau(i)}\right|^2$$

1D Wasserstein and Sliced Wasserstein



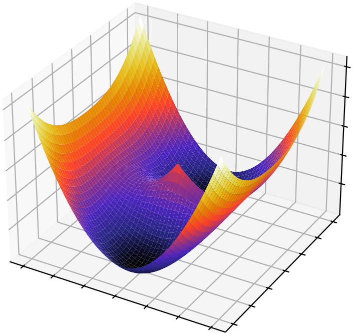
$$\mathrm{W}_2^2(\gamma_S, \gamma_T) = \frac{1}{n}\sum_{i=1}^{n} \left|s_{\sigma(i)} - t_{\tau(i)}\right|^2$$



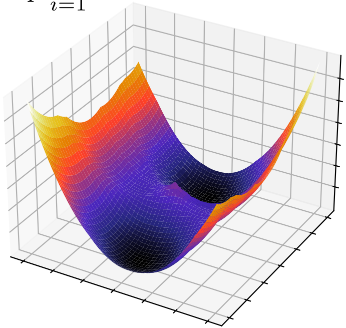
$$\mathrm{SW}_2^2(\gamma_X, \gamma_Y) = \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^d)}\left[\mathrm{W}_2^2(\theta\#\gamma_X, \theta\#\gamma_Y)\right]$$

Monte-Carlo Approximation

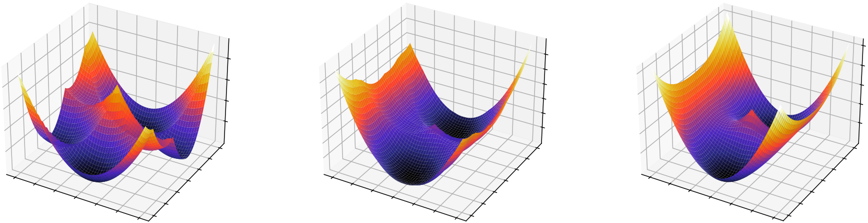
$$\mathcal{E}(X) = \mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^d)}\left[\mathrm{W}_2^2(\theta\#\gamma_X, \theta\#\gamma_Y)\right]$$



$$\mathcal{E}_p(X) := \frac{1}{p}\sum_{i=1}^{p}\mathrm{W}_2^2(\theta_i\#\gamma_X, \theta_i\#\gamma_Y)$$



Statistical Properties

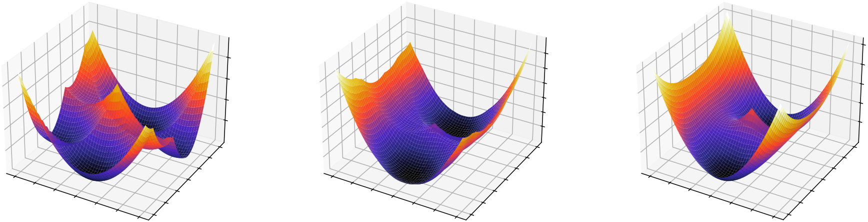


(a) $p = 3$

(b) $p = 10$

(c) $\mathcal{E}$

Statistical Properties

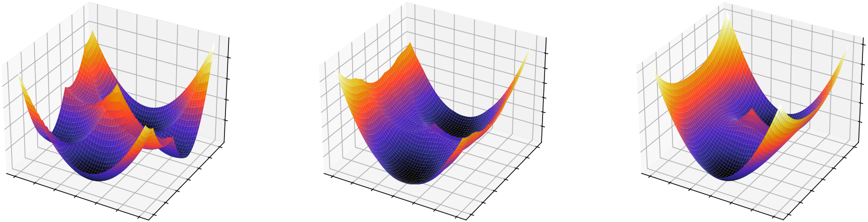


(a) $p = 3$ (b) $p = 10$ (c) $\mathcal{E}$

Uniform Convergence [5]

For $\mathcal{K} \subset \mathbb{R}^{n\times d}$ compact, $\mathbb{P}\left(\|\mathcal{E}_p - \mathcal{E}\|_{\infty,\mathcal{K}} \xrightarrow[p\to+\infty]{} 0\right) = 1.$

Statistical Properties



(a) $p = 3$ (b) $p = 10$ (c) $\mathcal{E}$

Uniform Convergence [5]

For $\mathcal{K} \subset \mathbb{R}^{n \times d}$ compact, $\mathbb{P}\left(\|\mathcal{E}_p - \mathcal{E}\|_{\infty, \mathcal{K}} \xrightarrow[p \longrightarrow +\infty]{} 0 \right) = 1.$

Uniform Central Limit Theorem [5]

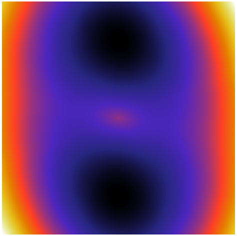
For $\mathcal{K} \subset \mathbb{R}^{n \times d}$ compact, $\sqrt{p}(\mathcal{E}_p - \mathcal{E}) \xrightarrow[p \longrightarrow +\infty]{\mathcal{L},\ \ell^\infty(\mathcal{K})} G.$

1. The Discrete Sliced Wasserstein Distance
2. Optimisation Properties
3. SGD Convergence
4. SGD for Training SW Neural Networks

Global Optima

- SW_2 is a distance:

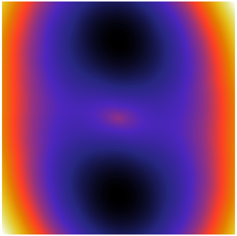
$$\underset{X \in \mathbb{R}^{n \times d}}{\operatorname{argmin}}\ \mathcal{E}(X) = \underset{X \in \mathbb{R}^{n \times d}}{\operatorname{argmin}}\ \mathrm{SW}_2^2(\gamma_X, \gamma_Y) = \{Y \text{ up to a permutation}\}$$



Global Optima

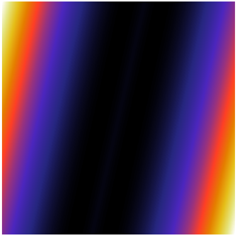
- SW_2 is a distance:

$$\begin{aligned}\underset{X\in\mathbb{R}^{n\times d}}{\mathrm{argmin}}\ \mathcal{E}(X) &= \underset{X\in\mathbb{R}^{n\times d}}{\mathrm{argmin}}\ \mathrm{SW}_2^2(\gamma_X,\gamma_Y) \\ &= \{Y \text{ up to a permutation}\}\end{aligned}$$



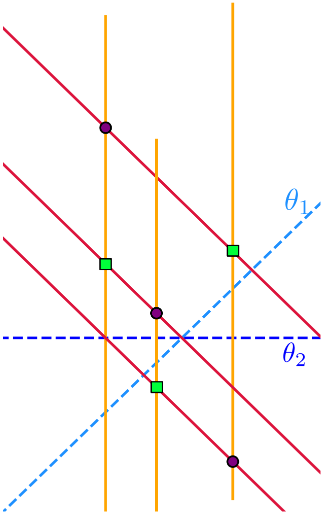
- $\widehat{\mathrm{SW}}_{2,p}$ is **not** a distance:

$$\widehat{\mathrm{SW}}_{2,p}(\gamma,\gamma_Y) = 0 \iff \forall i \in [\![1,p]\!],\ \theta_i\#\gamma = \theta_i\#\gamma_Y\,.$$

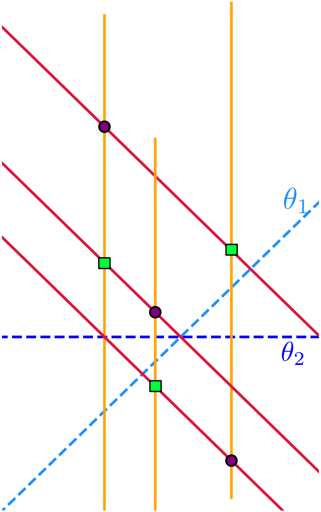


$\mathcal{E}_p$ with $p = 1$.

Reconstruction Problem



Reconstruction Problem



For $P_i : \mathbb{R}^d \longrightarrow \mathbb{R}^{d_i}$,

$$(\mathrm{RP}) : \ \forall i \in [\![1, p]\!],\ P_i\#\gamma = P_i\#\gamma_Y .$$

a.s. Reconstruction [4]

If $\sum_i d_i > d$, for $Y \in \mathbb{R}^{n\times d}$ fixed, $\mathcal{S}_{\mathrm{RP}} = \{\gamma_Y\}$, almost-surely, for random (P_i).

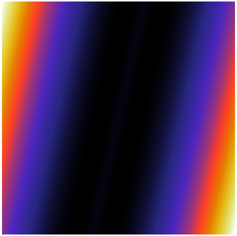
Consequences of the Reconstruction Problem on $\mathcal{E}_p$

If $p \leq d$,

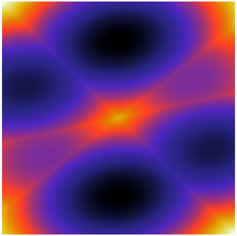
$$\mathcal{E}_p(X) = 0 \;\not\Longrightarrow\; X \in \{Y \text{ up to a permutation}\}.$$

If $p > d$, almost-surely,

$$\mathcal{E}_p(X) = 0 \Longrightarrow X \in \{Y \text{ up to a permutation}\}.$$



$\mathcal{E}_p$ with $p = 1$.



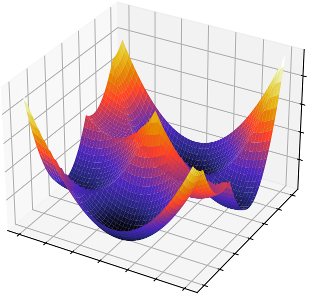
$\mathcal{E}_p$ with $p = 3$.

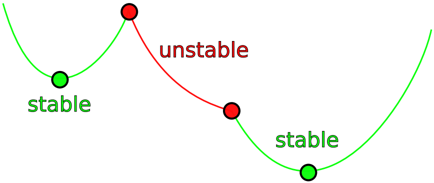
$\mathcal{E}_p$ Cell Decomposition

$$\mathcal{E}_p(X) = \frac{1}{p}\sum_{i=1}^{p} \mathrm{W}_2^2(\theta_i\#\gamma_X, \theta_i\#\gamma_Y) = \min_{(\sigma_1,\cdots,\sigma_p)\in\mathfrak{S}_n^p} \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}(\theta_i^T(x_k - y_{\sigma_i(k)}))^2.$$

$\mathcal{E}_p$ Cell Decomposition

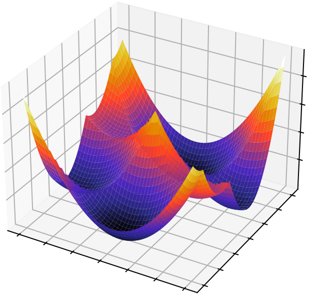
$$\mathcal{E}_p(X) = \frac{1}{p}\sum_{i=1}^{p} \mathrm{W}_2^2(\theta_i\#\gamma_X, \theta_i\#\gamma_Y) = \min_{(\sigma_1,\cdots,\sigma_p)\in\mathfrak{S}_n^p} \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}(\theta_i^T(x_k - y_{\sigma_i(k)}))^2.$$

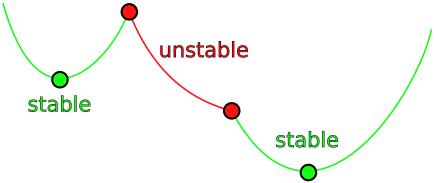




$\mathcal{E}_p$ Cell Decomposition

$$\mathcal{E}_p(X) = \frac{1}{p}\sum_{i=1}^{p} \mathrm{W}_2^2(\theta_i\#\gamma_X, \theta_i\#\gamma_Y) = \min_{(\sigma_1,\cdots,\sigma_p)\in\mathfrak{S}_n^p} \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}(\theta_i^T(x_k - y_{\sigma_i(k)}))^2.$$



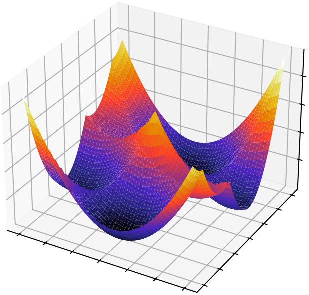


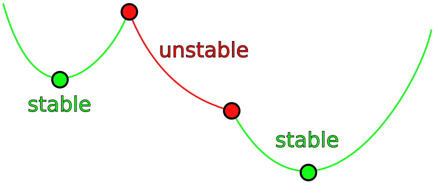
Cell Optima [5]

$\nabla\mathcal{E}_p(X) = 0 \iff X$ is min of a stable cell $\iff X$ is a local min.

$\mathcal{E}_p$ Cell Decomposition

$$\mathcal{E}_p(X) = \frac{1}{p}\sum_{i=1}^{p} \mathrm{W}_2^2(\theta_i\#\gamma_X, \theta_i\#\gamma_Y) = \min_{(\sigma_1,\cdots,\sigma_p)\in\mathfrak{S}_n^p} \frac{1}{np}\sum_{i=1}^{p}\sum_{k=1}^{n}(\theta_i^T(x_k - y_{\sigma_i(k)}))^2.$$



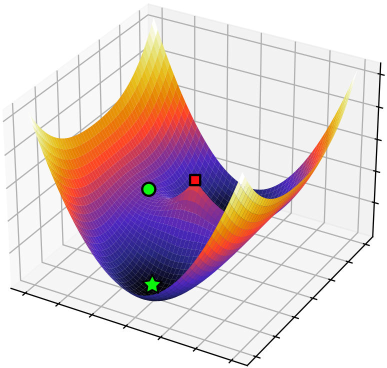


Cell Optima [5]

$$\nabla\mathcal{E}_p(X) = 0 \iff X \text{ is min of a stable cell} \iff X \text{ is a local min.}$$

As $p \longrightarrow +\infty$, $\mathcal{E}_p \approx \mathcal{E}$, more local optima but better optimisation.

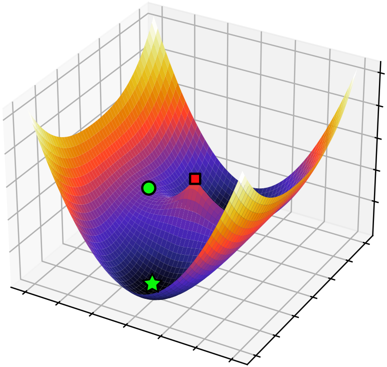
$\mathcal{E}$ Differentiable Critical Points



Critical Points of $\mathcal{E}$ [5]

$$\forall X \in \mathcal{D}_{\mathcal{E}},$$
$$\nabla \mathcal{E}(X) = 0 \iff F(X) = X$$

$\mathcal{E}$ Differentiable Critical Points



Critical Points of $\mathcal{E}$ [5]

$$\forall X \in \mathcal{D}_{\mathcal{E}},$$
$$\nabla \mathcal{E}(X) = 0 \iff F(X) = X$$

Critical Point Approximation [5]

For X_p critical points of $\mathcal{E}_p$, $\quad X_p - F(X_p) \xrightarrow[p \longrightarrow +\infty]{\mathbb{P}} 0.$

1. The Discrete Sliced Wasserstein Distance
2. Optimisation Properties
3. SGD Convergence
4. SGD for Training SW Neural Networks

Preliminary: Stability of the Kantorovich Problem 1/2

Let $\alpha, \beta \in \Sigma_n$, $C \in \mathbb{R}_+^{n\times n}$ and $\Pi(\alpha,\beta) = \{\pi \in \mathbb{R}_+^{n\times n},\ \pi\mathbb{1} = \alpha,\ \pi^T\mathbb{1} = \beta\}$.

$$\mathrm{W}(\alpha,\beta;C) := \inf_{\pi\in\Pi(\alpha,\beta)} \pi \cdot C$$

Preliminary: Stability of the Kantorovich Problem 1/2

Let $\alpha, \beta \in \Sigma_n$, $C \in \mathbb{R}_+^{n\times n}$ and $\Pi(\alpha, \beta) = \{\pi \in \mathbb{R}_+^{n\times n},\ \pi\mathbb{1} = \alpha,\ \pi^T\mathbb{1} = \beta\}$.

$$\mathrm{W}(\alpha, \beta; C) := \inf_{\pi\in\Pi(\alpha,\beta)} \pi \cdot C$$

Stability of the Kantorovich LP [5]

$$\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C})\right| \leq \|C-\overline{C}\|_\infty + \|C\|_\infty(\|\alpha-\overline{\alpha}\|_1 + \|\beta-\overline{\beta}\|_1).$$

Preliminary: Stability of the Kantorovich Problem 1/2

Let $\alpha, \beta \in \Sigma_n$, $C \in \mathbb{R}_+^{n\times n}$ and $\Pi(\alpha, \beta) = \{\pi \in \mathbb{R}_+^{n\times n},\ \pi\mathbb{1} = \alpha,\ \pi^T\mathbb{1} = \beta\}$.

$$\mathrm{W}(\alpha, \beta; C) := \inf_{\pi\in\Pi(\alpha,\beta)} \pi \cdot C$$

Stability of the Kantorovich LP [5]

$$\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}; \overline{C})\right| \leq \|C-\overline{C}\|_\infty + \|C\|_\infty(\|\alpha-\overline{\alpha}\|_1 + \|\beta-\overline{\beta}\|_1).$$

Proof. 1)

$$\begin{aligned}\mathrm{W}(\alpha, \beta, C) - \mathrm{W}(\alpha, \beta, \overline{C}) &= \inf_{\pi\in\Pi(\alpha,\beta)} \pi \cdot C - \inf_{\overline{\pi}\in\Pi(\alpha,\beta)} \overline{\pi} \cdot \overline{C} \\ &\leq \overline{\pi}^* \cdot C - \overline{\pi}^* \cdot \overline{C} \\ &= \sum_{i,j} \overline{\pi}^*_{i,j}(C_{i,j} - \overline{C}_{i,j}) \\ &\leq \|C - \overline{C}\|_\infty \sum_{i,j} \overline{\pi}^*_{i,j} = \|C - \overline{C}\|_\infty.\end{aligned}$$

Preliminary: Stability of the Kantorovich Problem 1/2

Proof. 2)

- Dual expression
  $$\mathrm{W}(\alpha, \beta, C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}, C) = \sup_{f \oplus g \leq C} f^T \alpha + g^T \beta - \sup_{\overline{f} \oplus \overline{g} \leq C} \overline{f}^T \overline{\alpha} + \overline{g}^T \overline{\beta}$$
- Complementary slackness: $\pi^*_{i,j} \neq 0 \Longrightarrow f^*_i + g^*_i = C_{i,j}$
- Bound dual $\|f^*\|_\infty \leq \|C\|_\infty, \ \|g^*\|_\infty \leq \|C\|_\infty$.

Preliminary: Stability of the Kantorovich Problem 1/2

Proof. 2)

- Dual expression
  $$\mathrm{W}(\alpha, \beta, C) - \mathrm{W}(\overline{\alpha}, \overline{\beta}, C) = \sup_{f \oplus g \leq C} f^T \alpha + g^T \beta - \sup_{\overline{f} \oplus \overline{g} \leq C} \overline{f}^T \overline{\alpha} + \overline{g}^T \overline{\beta}$$
- Complementary slackness: $\pi^*_{i,j} \neq 0 \implies f^*_i + g^*_i = C_{i,j}$
- Bound dual $\|f^*\|_\infty \leq \|C\|_\infty$, $\|g^*\|_\infty \leq \|C\|_\infty$.

$$\left|\mathrm{W}(\alpha, \beta; C) - \mathrm{W}(\alpha, \beta; \overline{C})\right| \leq \|C - \overline{C}\|_\infty.$$

Consequence with $C_{k,l} := \|x_k - y_l\|_2^2$ and $X, X' \in \mathcal{K}$:

$$|\mathrm{W}_2^2(\gamma_X, \gamma_Y) - \mathrm{W}_2^2(\gamma_{X'}, \gamma_Y)| \leq c_{\mathcal{K},Y} \max_k \|x_k - x'_k\|_2.$$

Convergence of Interpolated Trajectories

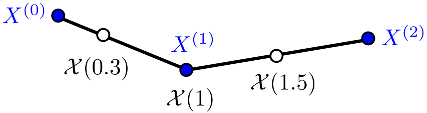
SGD on $\mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^d)}\left[\underbrace{\mathrm{W}_2^2(\theta\#\gamma_X,\theta\#\gamma_Y)}_{w_\theta(X)}\right]$:

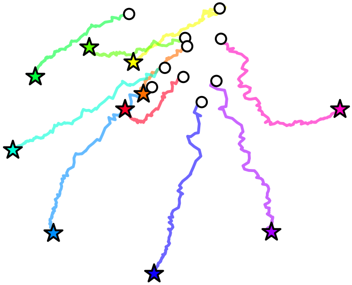
$$X^{(k+1)} = X^{(k)} - \alpha\nabla w_{\theta^{(k+1)}}(X^{(k)})$$

Convergence of Interpolated Trajectories

SGD on $\mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^d)}\left[\underbrace{\mathrm{W}_2^2(\theta\#\gamma_X, \theta\#\gamma_Y)}_{w_\theta(X)}\right]$:

$$X^{(k+1)} = X^{(k)} - \alpha\nabla w_{\theta^{(k+1)}}(X^{(k)})$$

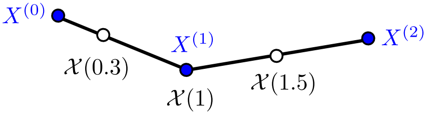


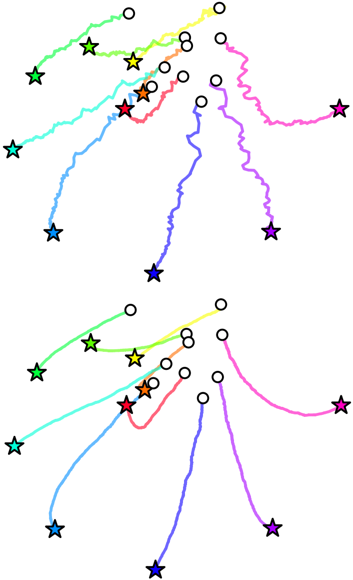


Convergence of Interpolated Trajectories

SGD on $\mathbb{E}_{\theta\sim\mathcal{U}(\mathbb{S}^d)}\left[\underbrace{\mathrm{W}_2^2(\theta\#\gamma_X,\theta\#\gamma_Y)}_{w_\theta(X)}\right]$:

$$X^{(k+1)} = X^{(k)} - \alpha\nabla w_{\theta^{(k+1)}}(X^{(k)})$$





Interpolations Converge [5]

$$d(\mathcal{X}_\alpha,\mathcal{S}) \xrightarrow[\alpha\to 0]{\mathbb{P}} 0.$$

With $\mathcal{S} = \left\{\mathcal{X}\ \middle|\ \frac{\mathrm{d}\mathcal{X}}{\mathrm{d}t}(t) \in -\partial_C\mathcal{E}(\mathcal{X}(t))\right\}$.

Using results from Bianchi et al. [1]

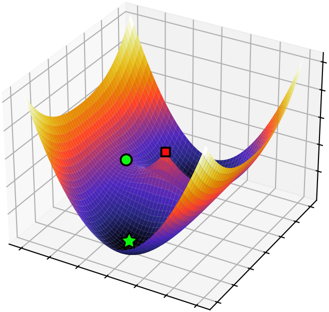
Convergence of Noised Trajectories

Noised SGD: $X^{(k+1)} = X^{(k)} - \alpha \nabla w_{\theta^{(k+1)}}(X^{(k)}) + \alpha \varepsilon^{(k+1)}$.

Convergence of Noised SGD [5]

$$\varlimsup_{k \longrightarrow +\infty} d(X_\alpha^{(k)}, \mathcal{Z}) \xrightarrow[\alpha \longrightarrow 0]{\mathbb{P}} 0.$$

With $\mathcal{Z} = \left\{ X \in \mathbb{R}^{n \times d} \mid 0 \in -\partial_C \mathcal{E}(X) \right\}$.



Using results from Bianchi et al. [1]

Convergence of Decreasing-Step Noised Trajectories

$$X^{(k+1)} = X^{(k)} - \alpha^{(k)} \nabla w_{\theta^{(k+1)}}(X^{(k)}) + \alpha \varepsilon^{(k+1)}.$$

Steps $\alpha^{(k)} \geq 0$ with $\sum_{k=0}^{+\infty} \alpha^{(k)} = +\infty$ and $\sum_{k=0}^{+\infty} (\alpha^{(k)})^2 = +\infty$.

Convergence of Decreasing-Step Noised SGD [5]

If $(X^{(k)})$ is a.s. bounded, then a.s.:

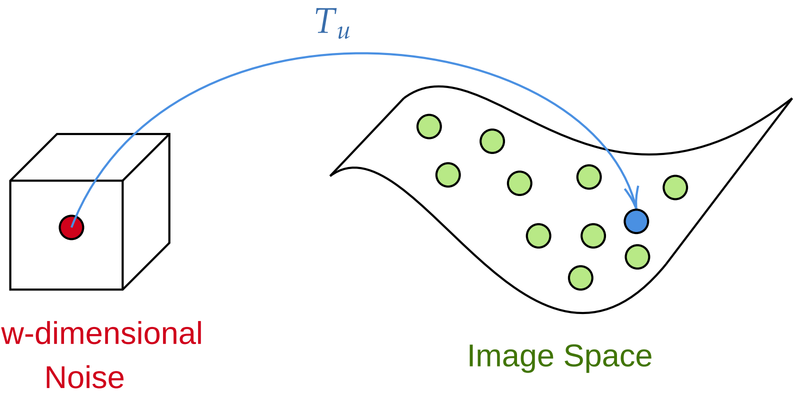
- $(\mathcal{E}(X^{(k)})_k$ converges.
- If $X^{(\varphi(k))} \xrightarrow[k \longrightarrow +\infty]{} X^\infty$, then $X^\infty \in \mathcal{Z}$.

With $\mathcal{Z} = \left\{ X \in \mathbb{R}^{n \times d} \mid 0 \in -\partial_C \mathcal{E}(X) \right\}$.

Using results from Davis et al. [2]

1. The Discrete Sliced Wasserstein Distance
2. Optimisation Properties
3. SGD Convergence
4. **SGD for Training SW Neural Networks**

Generative Modelling



Problem Statement

Goal: approximate $T_u\#\mathbb{x} \approx \mathbb{y}$.

Loss sample:

$$f(u, X, Y, \theta) = \mathrm{W}_2^2(\theta\#T_u\#\gamma_X, \theta\#\gamma_Y), \quad X \sim \mathbb{x}^{\otimes n},\ Y \sim \mathbb{y}^{\otimes n},\ \theta \sim \mathbb{\sigma}.$$

Population loss:

$$F(u) = \underset{X,Y,\theta}{\mathbb{E}}\left[\mathrm{W}_2^2(\theta\#T_u\#\gamma_X, \theta\#\gamma_Y)\right] = \underset{X,Y}{\mathbb{E}}\left[\mathrm{SW}_2^2(T_u\#\gamma_X, \gamma_Y)\right].$$

Convergence Results [3]

Under technical assumptions:

- Approximation of (Clarke) gradient flows
- Convergence in the parameters $u^{(t)}$ for a modified SGD scheme

Thank You

[1] Pascal Bianchi, Walid Hachem, and Sholom Schechtman.
Convergence of constant step stochastic gradient descent for non-smooth non-convex functions.
Set-Valued and Variational Analysis, 30(3):1117–1147, 2022.

[2] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee.
Stochastic subgradient method converges on tame functions.
Foundations of computational mathematics, 20(1):119–154, 2020.

[3] Eloi Tanguy.
Convergence of sgd for training neural networks with sliced Wasserstein losses.
Transactions on Machine Learning Research, October 2023.

[4] Eloi Tanguy, Rémi Flamary, and Julie Delon.
Reconstructing discrete measures from projections. consequences on the empirical sliced Wasserstein distance.
arXiv preprint arXiv:2304.12029, 2023.

[5] Eloi Tanguy, Rémi Flamary, and Julie Delon.
Properties of discrete sliced Wasserstein losses.