Orthogonal estimation of Wasserstein distances

Mark Rowland*, Jiri Hron*, Yunhao Tang*,
Krzysztof Choromanski, Tamas Sarlos, Adrian Weller

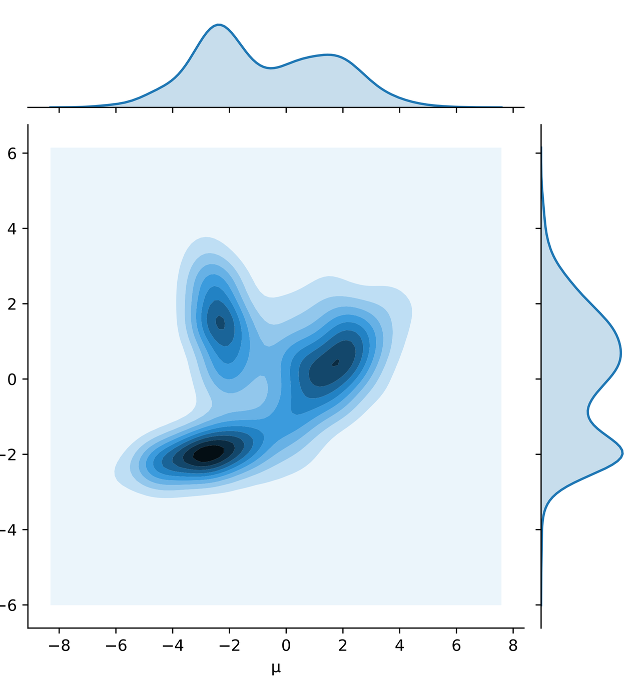
Wasserstein distances

- A class of metrics between probability distributions
- Naturally incorporate spatial information
- Applications from economics to machine learning

Def: *For a metric space* $(\mathcal{X}, d)$*, the* p-**Wasserstein distance** *between distributions* $\mu, \nu \in \mathcal{P}(\mathcal{X})$ *is*

$$\mathrm{W}_p(\mu,\nu) \coloneqq \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \int d(x,y)^p \,\mathrm{d}\gamma(x,y) \right)^{1/p},$$

where $\Gamma(\mu,\nu) \subseteq \mathcal{P}(\mathcal{X}\times\mathcal{X})$ *is the set of joint distributions with marginals* μ *and* ν*.*



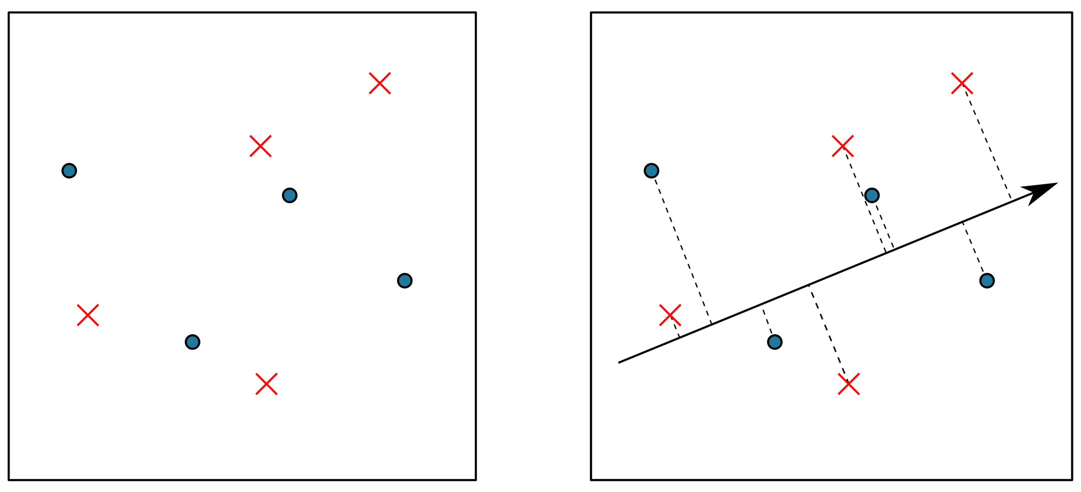
Source: commons.wikimedia.org/w/index.php?curid=64872543

Unfortunately, computation of $W_p(\mu,\nu)$ is often very **expensive or outright intractable**.

Sliced Wasserstein distance

Computational complexity improves if $\mathcal{X} = \mathbb{R}^d$, $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ and $d(x,y) = \|x-y\|_2$, as computation of W_p reduces to a matching problem with $\mathcal{O}(n^{5/2}\log n)$ complexity.

If $d = 1$, problem further reduces to sorting with complexity $\mathcal{O}(n \log n)$. Sliced Wasserstein distances take advantage of this **computational speed up**.



SW_p: Illustration of a single projection of μ,ν with $n=4$ and $d=2$

Def: *The* p-**sliced Wasserstein distance** *between* $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ *and* $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ *is*

$$\mathrm{SW}_p(\mu,\nu) \coloneqq \left[\mathbb{E}_v \left(\frac{1}{n}\sum_{i=1}^n |\langle v, x_i\rangle - \langle v, y_{\sigma_v(i)}\rangle|^p \right) \right]^{1/p},$$

$v \sim \mathrm{Unif}(S^{d-1})$*, and* $\sigma_v \colon [n] \to [n]$ *the bijective mapping with the property that*

$$\langle v, x_i\rangle < \langle v, x_j\rangle \Rightarrow \langle v, y_{\sigma_v(i)}\rangle \le \langle v, y_{\sigma_v(j)}\rangle .$$

Our contributions

- Analysis of an estimator of sliced Wasserstein distance based on **orthogonal coupling**
- Exploration of a new Wasserstein-like metric, **projected Wasserstein distance**

Projected Wasserstein distance

Idea: Use the coupling σ_v for $v \sim \mathrm{Unif}(S^{d-1})$ as SW_p, but assign cost in $(\mathbb{R}^d, \|\cdot\|_2)$ like W_p.

Def: *The* p-**projected Wasserstein distance** *between* $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ *and* $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ *is*

$$\mathrm{PW}_p(\mu,\nu) \coloneqq \left[\mathbb{E}_v \left(\frac{1}{n}\sum_{i=1}^n \|x_i - y_{\sigma_v(i)}\|_2^p \right) \right]^{1/p},$$

where v *and* σ_v *are as in the definition of* SW_p*.*

Properties:

For any two distributions $\mu,\nu \in \mathcal{P}_{(n)}(\mathbb{R}^d) \coloneqq \{\frac{1}{n}\sum_{i=1}^n \delta_{x_i} : \{x_i\}_{i=1}^n \subset \mathbb{R}^d\}$, $n \in \mathbb{N}$, and any $p \ge 1$:

- $\mathrm{PW}_p(\mu,\nu)$ is a metric
- $\mathrm{SW}_p(\mu,\nu) \le \mathrm{W}_p(\mu,\nu) \le \mathrm{PW}_p(\mu,\nu)$

PW_p shares many properties with W_p and SW_p:

- Helps with **theoretical analysis** of SW_p
- PW_p may be of **independent interest**

MC and orthogonal coupling

The computation of the expectation over $v \sim \mathrm{Unif}(S^{d-1})$ in SW (resp. PW) is often intractable $\Rightarrow$ estimate via **MC integration**:

$$\mathbb{E}_v[f_{\mu,\nu}(v)] \approx \frac{1}{m}\sum_{j=1}^m f_{\mu,\nu}(v_j),$$

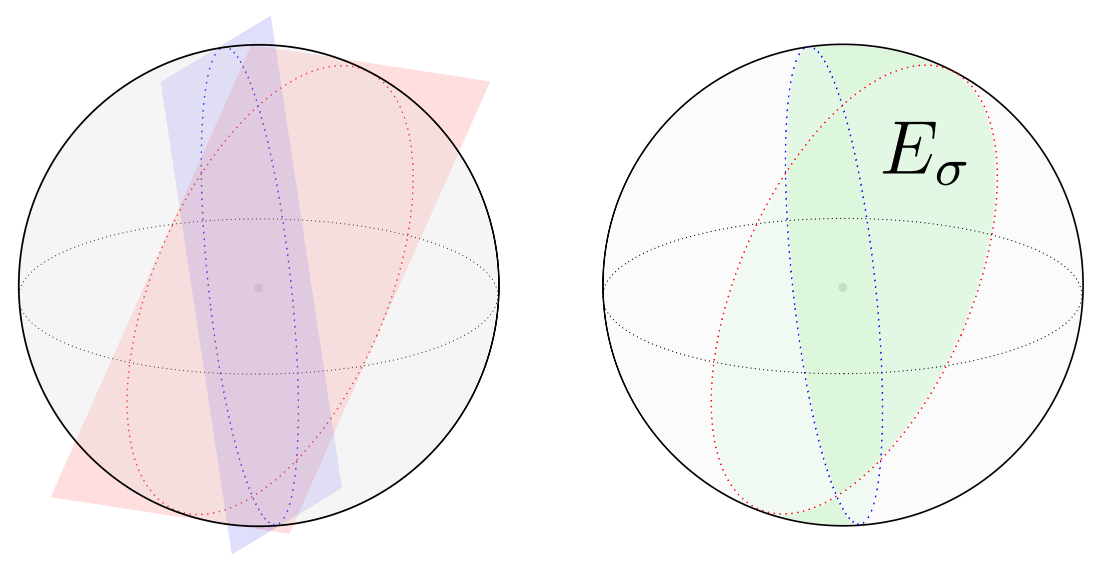
with $f_{\mu,\nu} \colon S^{d-1} \to \mathbb{R}$ defined as in SW (resp. PW).

Idea: Instead of i.i.d., sample $\{v_j\}_{j=1}^m$ uniformly at random from the space of orthogonal matrices.

Mean squared error analysis

MSE can be understood through a σ_v induced partition of $S^{d-1} = \bigcup_{\sigma\in\mathcal{S}_n} E_\sigma$, $E_\sigma \coloneqq \{v \colon \sigma_v = \sigma\}$.

Lem: E_σ *is a finite union of simply connected sets.*



S^{d-1} partition: σ_v changes whenever $\langle v, x_i - x_j\rangle = 0$ or $\langle v, y_i - y_j\rangle = 0$

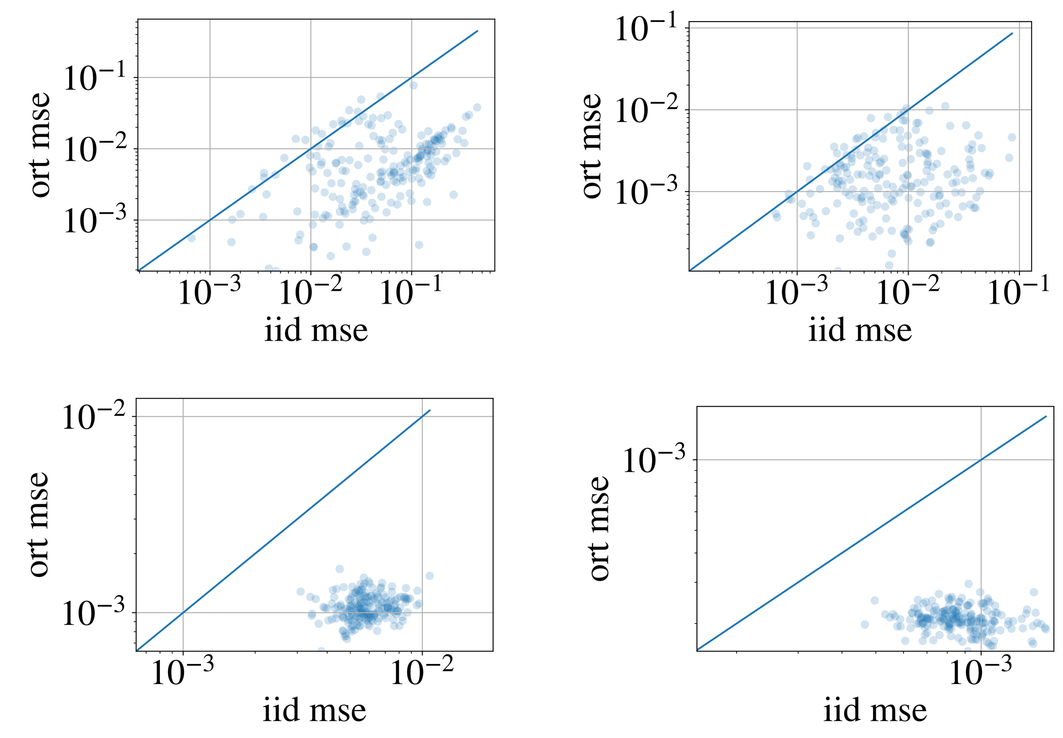
Prop: *An unbiased estimator for which*

$$\mathbb{P}(v_i \in E_\sigma, v_j \in E_\tau) > \mathbb{P}(v_i \in E_\sigma)\,\mathbb{P}(v_j \in E_\tau),$$

$i \neq j$, $\sigma \neq \tau$, *has MSE strictly lower than i.i.d.*

$\Rightarrow$ **Stratification** w.r.t. the σ_v induced partition leads to **improved MSE**. The number of possible $\sigma \in \mathcal{S}_n$ is $n!$ though, making direct **stratification computationally infeasible**.

Idea: View **orthogonal coupling** of the directions $\{v_j\}_{j=1}^m$ as an **approximation to stratification**.



MSE (SW): $d=2$ (left), $d=50$ (right); $n=2$ (top); $n=10$ (bottom)

Prop: *Let* $n=2, d=2$*. Then orthogonal coupling dominates i.i.d. in terms of MSE for the projected Wasserstein, but not sliced Wasserstein distance.*

Difference between PW_p and SW_p reveals why orthogonal coupling sometimes hurts SW_p estimation:

Prop: *Let* $\mathcal{F} \coloneqq \sigma(E_\sigma \colon \sigma \in \mathcal{S}_n)$*, and* $\{v_j\}_{j=1}^m$ *be orthogonally coupled. Then* $\{\mathbb{E}[f_{\mu,\nu}^{\mathrm{PW}}(v_j) \mid \mathcal{F}]\}_{j=1}^m$ *are pairwise independent, but the same is not true with* $f_{\mu,\nu}^{\mathrm{SW}}$*. Pairwise independence ensures stratification.*

Effect on downstream tasks

Sampling orthogonally coupled vectors

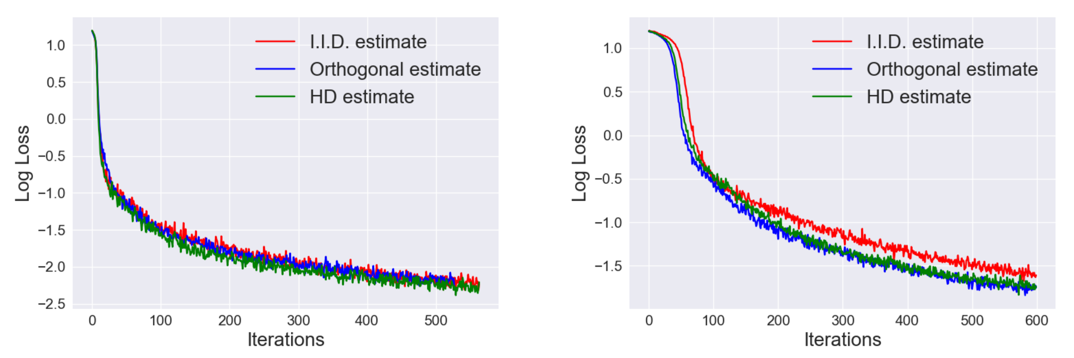
Exact: Sample i.i.d. and apply Gram-Schmidt

Approximate: Use $\prod_l^L H_l D_l$, H_l a scaled Hadamard matrix, D_l a diagonal Rademacher matrix

Sliced Wasserstein AE *(Kolouri et al.)*

Set-up: Encoder $h_\theta \colon \mathbb{R}^d \to \mathbb{R}^k$, decoder $g_\phi \colon \mathbb{R}^k \to \mathbb{R}^d$, empirical distribution P_X (MNIST), prior P_Z.

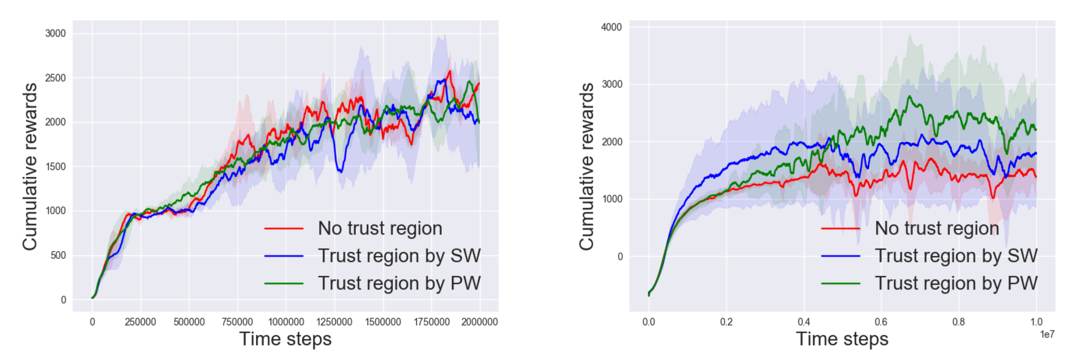
$$\mathbb{E}_{P_X}[\|g_\phi(h_\theta(X)) - X\|^2] + \mathrm{SW}_1((h_\theta)_\# P_X, P_Z).$$



SGD training: Orthogonality reduces gradient variance

Trust region policy optimisation *(Schulman et al.)*

Set-up: Policy $\pi_\theta \colon s_t \mapsto a_t$ and a fixed MDP. Maximise $J(\pi_\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^\infty \gamma^t r_t]$. Each step constrained by $D(\theta_t, \theta_{t+1}) \le \varepsilon$ with $D = \mathrm{SW}_1, \mathrm{PW}_1$.



Training curves: Hopper (left), HalfCheetah (right); 5 random seeds

Summary & future work

- Orthogonal coupling often improves MSE
- MSE improvement linked to stratified sampling
- Experimentally, reduced variance can help with downstream tasks but more research needed