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 $(X,d)$, the $p$-Wasserstein distance between distributions $\mu, \nu \in \mathcal{P}(X)$ is

$$W_p(\mu, \nu) = \left( \inf_{\gamma \in \Pi(\mu, \nu)} \int d(x,y) \, d\gamma(x,y) \right)^{1/p},$$

where $\Pi(\mu, \nu)$ is the set of joint distributions with marginals $\mu$ and $\nu$.

Sliced Wasserstein distance

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

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

Orthogonal coupling

Orthogonal estimation of Wasserstein distances

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

$$\text{PW}_p(\mu, \nu) := \left[ \mathbb{E}_v \left( \frac{1}{n} \sum_{i=1}^n \|v(x_i) - y_{\sigma(i)}\|_p^p \right) \right]^{1/p},$$

where $v$ and $\sigma$ are as defined in the definition of SW$_p$.

Properties:
- For any two distributions $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, $\mu \leq \text{PW}_p(\mu, \nu) \leq \text{PW}_p(\mu, \nu)$
- Helps with theoretical analysis of SW$_p$
- PW$_p$ may be of independent interest

MC and orthogonal coupling

The computation of the expectation over $v \sim \text{Unif}(S^{d-1})$ in SW (resp. PW) is often intractable.

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

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

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

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

where $\sigma$ is a bijection from $[n] \rightarrow [n]$. The bijective mapping with the property that

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

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

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

$$W_p(\mu, \nu) = \left( \inf_{\gamma \in \Pi(\mu, \nu)} \int d(x,y) \, d\gamma(x,y) \right)^{1/p},$$

where $\Pi(\mu, \nu)$ is the set of joint distributions with marginals $\mu$ and $\nu$.

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

$$\text{PW}_p^\perp(\mu, \nu) := \left[ \mathbb{E}_v \left( \frac{1}{n} \sum_{i=1}^n \|x_i - y_{\sigma(i)}\|_p^p \right) \right]^{1/p},$$

where $v$ and $\sigma$ are as defined in the definition of SW$_p$.

Properties:
- For any two distributions $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, $\mu \leq \text{PW}_p^\perp(\mu, \nu) \leq \text{PW}_p(\mu, \nu)$
- Helps with theoretical analysis of SW$_p$
- PW$_p^\perp$ may be of independent interest

Mean squared error analysis

MSE can be understood through a $\sigma_v$ induced partition of $S^{d-1} = \bigcup_{v \in \mathcal{S}^d} E_v$, $E_v = \{v : \beta_v = \sigma_v \}$. Lem: $E_v$ is a finite union of simply connected sets.

Effect on downstream tasks

Sampling orthogonally coupled vectors

Exact: Sample i.i.d. and apply Gram-Schmidt Approximate: Use $\prod_i H_i D_i$, $H_i$ a scaled Hadamard matrix, $D_i$ a diagonal Rademacher matrix

Sliced Wasserstein AE (Kolouri et al.)

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

$\mathbb{E}_X[\| g_{\sigma}(h_{\theta}(X)) - X \|^2] + \mathbb{E}_W[\| h_{\theta}(P_Z) - P_X \|^2].$

Trust region policy optimisation (Schulman et al.)

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

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