Supplementary Material to the Article:
When OT meets MoM: Robust estimation of Wasserstein Distance

A Technical Proofs

Current section details the proofs of the theoretical claims stated in the core article. We first recall a simple lemma on the difference between two median vectors.

Lemma 1. *Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors of $\mathbb{R}^d$. Then it holds*

$$\left|\text{median}(\boldsymbol{a}) - \text{median}(\boldsymbol{b})\right| \leq \|\boldsymbol{a} - \boldsymbol{b}\|_\infty.$$

Proof. It is direct to see that:

$$\boldsymbol{a} \preceq \boldsymbol{b} \preceq \boldsymbol{c} \quad \Rightarrow \quad \text{median}(\boldsymbol{a}) \leq \text{median}(\boldsymbol{b}) \leq \text{median}(\boldsymbol{c}).$$

Thus, for all $\boldsymbol{b}$ within the infinite ball of center $\boldsymbol{a}$ and radius ϵ it holds:

$$\text{median}(\boldsymbol{a}) - \epsilon = \text{median}(\boldsymbol{a} - \epsilon \mathbf{1}_d) \leq \text{median}(\boldsymbol{b}) \leq \text{median}(\boldsymbol{a} + \epsilon \mathbf{1}_d) = \text{median}(\boldsymbol{a}) + \epsilon.$$

Hence the conclusion. □

A.1 Proof of Proposition 4

We first show the consistency of $\mathcal{W}_{\text{MoU}}(\hat{\mu}_n, \hat{\nu}_m)$, that of $\mathcal{W}(\hat{\mu}_{\text{MoM}}, \mu)$ and $\mathcal{W}_{\text{MoU-diag}}(\hat{\mu}_n, \hat{\nu}_m)$ being then straightforward adaptations. Assume that $\tilde{\tau} = \tau_{\mathbf{X}} + \tau_{\mathbf{Y}} - \tau_{\mathbf{X}}\tau_{\mathbf{Y}} < 1/2$, and $K_{\mathbf{X}}, K_{\mathbf{Y}} > 0$ such that $2(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}} - \tau_{\mathbf{X}}\tau_{\mathbf{Y}}) < K_{\mathbf{X}}K_{\mathbf{Y}}/(nm)$. The latter condition implies that the blocks containing no outlier are in majority. Indeed, the number of contaminated blocks is upper bounded by:

$$n_{\mathcal{O}}K_{\mathbf{Y}} + n_{\mathcal{O}}K_{\mathbf{X}} - n_{\mathcal{O}}n_{\mathcal{O}} \leq (\tau_{\mathbf{X}} + \tau_{\mathbf{Y}} - \tau_{\mathbf{X}}\tau_{\mathbf{Y}})nm < K_{\mathbf{X}}K_{\mathbf{Y}}/2.$$

One may choose $K_{\mathbf{X}}$ and $K_{\mathbf{Y}}$ as small as possible such that the above condition is respected. Following this, it is a natural choice to set $K_{\mathbf{X}} = \lceil \sqrt{2\tilde{\tau}}\ n \rceil$ and $K_{\mathbf{Y}} = \lceil \sqrt{2\tilde{\tau}}\ m \rceil$.

Let $\mathcal{I}_{\mathbf{X}}$ (respectively $\mathcal{I}_{\mathbf{Y}}$) denote the set of indices of $\mathbf{X}$ blocks (respectively $\mathbf{Y}$ blocks) containing no outliers. Let $\mathcal{K}$ be a bounded subspace of $\mathbb{R}^d$, and assume that X, Y are valued in $\mathcal{X}, \mathcal{Y} \subset \mathcal{K}$. Finally, we denote by $\overline{\phi}_{\mathbf{X},k}$ and $\overline{\phi}_{\mathbf{Y},l}$ the quantities

$$\overline{\phi}_{\mathbf{X},k} = \frac{1}{B_{\mathbf{X}}} \sum_{i \in \mathcal{B}_k^{\mathbf{X}}} \phi(X_i), \qquad \text{and} \qquad \overline{\phi}_{\mathbf{Y},l} = \frac{1}{B_{\mathbf{Y}}} \sum_{j \in \mathcal{B}_l^{\mathbf{Y}}} \phi(Y_j).$$

Using the shortcut notation $\mathbb{E}_\mu[\phi] = \mathbb{E}_{X\sim\mu}[\phi(X)]$ and $\mathbb{E}_\nu[\phi] = \mathbb{E}_{Y\sim\nu}[\phi(Y)]$, first notice that:

$$\begin{aligned}
\mathcal{W}_{\text{MoU}}(\hat{\mu}_n, \hat{\nu}_m) &= \sup_{\phi \in \mathcal{B}_L} \ \text{MoU}_{\mathbf{XY}}[h_\phi], \\
&= \sup_{\phi \in \mathcal{B}_L} \ \underset{\substack{1\leq k\leq K_{\mathbf{X}} \\ 1\leq l\leq K_{\mathbf{Y}}}}{\text{med}} \left\{\overline{\phi}_{\mathbf{X},k} - \overline{\phi}_{\mathbf{Y},l}\right\}, \\
&= \sup_{\phi \in \mathcal{B}_L} \ \underset{\substack{1\leq k\leq K_{\mathbf{X}} \\ 1\leq l\leq K_{\mathbf{Y}}}}{\text{med}} \left\{\overline{\phi}_{\mathbf{X},k} - \mathbb{E}_\mu[\phi] + \mathbb{E}_\mu[\phi] - \mathbb{E}_\nu[\phi] + \mathbb{E}_\nu[\phi] - \overline{\phi}_{\mathbf{Y},l}\right\}, \\
&\leq \sup_{\phi \in \mathcal{B}_L} \ \underset{\substack{1\leq k\leq K_{\mathbf{X}} \\ 1\leq l\leq K_{\mathbf{Y}}}}{\text{med}} \left\{\overline{\phi}_{\mathbf{X},k} - \mathbb{E}_\mu[\phi] + \mathbb{E}_\nu[\phi] - \overline{\phi}_{\mathbf{Y},l}\right\} + \mathcal{W}(\mu, \nu). \qquad (1)
\end{aligned}$$

Conversely, it holds:

$$\begin{aligned}\mathcal{W}(\mu,\nu) &= \sup_{\phi\in\mathcal{B}_L}\left\{\mathbb{E}_\mu[\phi]-\mathbb{E}_\nu[\phi]\right\},\\ &\le \sup_{\phi\in\mathcal{B}_L}\left\{\mathbb{E}_\mu[\phi]-\overline{\phi}_{\mathcal{B}^{\mathbf{X}}_{\text{med}}}+\overline{\phi}_{\mathcal{B}^{\mathbf{Y}}_{\text{med}}}-\mathbb{E}_\nu[\phi]+\overline{\phi}_{\mathcal{B}^{\mathbf{X}}_{\text{med}}}-\overline{\phi}_{\mathcal{B}^{\mathbf{Y}}_{\text{med}}}\right\},\\ &\le \sup_{\phi\in\mathcal{B}_L}\ \operatorname*{med}_{\substack{1\le k\le K_{\mathbf{X}}\\ 1\le l\le K_{\mathbf{Y}}}}\left\{\mathbb{E}_\mu[\phi]-\overline{\phi}_{\mathbf{X},k}+\overline{\phi}_{\mathbf{Y},l}-\mathbb{E}_\nu[\phi]\right\}+\mathcal{W}_{\text{MoU}}(\hat{\mu}_n,\hat{\nu}_m),\end{aligned}\tag{2}$$

where $\mathcal{B}^{\mathbf{X}}_{\text{med}}$ and $\mathcal{B}^{\mathbf{Y}}_{\text{med}}$ are the median blocks of $\overline{\phi}_{\mathbf{X},k}-\overline{\phi}_{\mathbf{Y},l}$ for $1\le k\le K_{\mathbf{X}}$ and $1\le l\le K_{\mathbf{Y}}$. From Equations (1) and (2), we deduce that:

$$\begin{aligned}\left|\mathcal{W}_{\text{MoU}}(\hat{\mu}_n,\hat{\nu}_m)-\mathcal{W}(\mu,\nu)\right| &\le \sup_{\phi\in\mathcal{B}_L}\ \operatorname*{med}_{\substack{1\le k\le K_{\mathbf{X}}\\ 1\le l\le K_{\mathbf{Y}}}}\left\{\left|\overline{\phi}_{\mathbf{X},k}-\mathbb{E}_\mu[\phi]+\mathbb{E}_\nu[\phi]-\overline{\phi}_{\mathbf{Y},l}\right|\right\},\\ &\le \sup_{k\in\mathcal{I}_{\mathbf{X}},\ l\in\mathcal{I}_{\mathbf{Y}}}\ \sup_{\phi\in\mathcal{B}_L}\left|\overline{\phi}_{\mathbf{X},k}-\mathbb{E}_\mu[\phi]+\mathbb{E}_\nu[\phi]-\overline{\phi}_{\mathbf{Y},l}\right|,\\ &\le \sup_{k\in\mathcal{I}_{\mathbf{X}}}\ \sup_{\phi\in\mathcal{B}_L}\left|\overline{\phi}_{\mathbf{X},k}-\mathbb{E}_\mu[\phi]\right|+\sup_{l\in\mathcal{I}_{\mathbf{Y}}}\ \sup_{\phi\in\mathcal{B}_L}\left|\mathbb{E}_\nu[\phi]-\overline{\phi}_{\mathbf{Y},l}\right|,\end{aligned}\tag{3}$$

where we have used the fact that $\mathcal{I}_{\mathbf{X}}\times\mathcal{I}_{\mathbf{Y}}$ represents a majority of blocks, and the subadditivity of the supremum. By independence between samples $\mathbf{X}$ and $\mathbf{Y}$, and between the blocks, it holds:

$$\begin{aligned}&\mathbb{P}\left\{\left|\mathcal{W}_{\text{MoU}}(\hat{\mu}_n,\hat{\nu}_m)-\mathcal{W}(\mu,\nu)\right|\underset{\substack{n\to+\infty\\ m\to+\infty}}{\longrightarrow}0\right\}\\ \ge &\prod_{k\in\mathcal{I}_{\mathbf{X}}}\mathbb{P}\left\{\sup_{\phi\in\mathcal{B}_L}\left|\overline{\phi}_{\mathbf{X},k}-\mathbb{E}_\mu[\phi]\right|\underset{n\to+\infty}{\longrightarrow}0\right\}\cdot\prod_{l\in\mathcal{I}_{\mathbf{Y}}}\mathbb{P}\left\{\sup_{\phi\in\mathcal{B}_L}\left|\overline{\phi}_{\mathbf{Y},l}-\mathbb{E}[\phi]\right|\underset{m\to+\infty}{\longrightarrow}0\right\}.\end{aligned}$$

Now, the arguments to get the right-hand side equal to 1 are similar to those used in Lemma 3.1 and Proposition 3.2 in Sriperumbudur et al. (2012). We expose them explicitly for the sake of clarity.

Let $\mathcal{N}(\varepsilon,\mathcal{B}_L,L^1(\mu))$ be the *covering number* of $\mathcal{B}_L$ which is the minimal number of $L^1(\mu)$ balls of radius ε needed to cover $\mathcal{B}_L$. Let $\mathcal{H}(\varepsilon,\mathcal{B}_L,L^1(\mu))$ be the *entropy* of $\mathcal{B}_L$, defined as $\mathcal{H}(\varepsilon,\mathcal{B}_L,L^1(\mu))=\log\mathcal{N}(\varepsilon,\mathcal{B}_L,L^1(\mu))$. Let F be the minimal enveloppe function such that $F(x)=\sup_{\phi\in\mathcal{B}_L}|\phi(x)|$. We need to check that $\int Fd\mu$ and $\int Fd\nu$ are finite and that $(1/n)\mathcal{H}(\varepsilon,\mathcal{B}_L,L^1(\hat{\mu}_n))$ and $(1/m)\mathcal{H}(\varepsilon,\mathcal{B}_L,L^1(\hat{\nu}_m))$ go to zero when n and m go to infinity. Then, we can apply Theorem 3.7 in van de Geer (2000) which ensures the uniform (a.s.) convergence of empirical processes. For any $\phi\in\mathcal{B}_L$, one has

$$\phi(x)\le\sup_{x\in\mathcal{K}}|\phi(x)|\le\sup_{x,y\in\mathcal{K}}|\phi(x)-\phi(y)|\le\sup_{x,y\in\mathcal{K}}\|x-y\|=\text{diam}(\mathcal{K})<+\infty.\tag{4}$$

Therefore $F(x)$ is finite, and following Lemma 3.1. in Kolmogorov and Tihomirov (1961) we have

$$\mathcal{H}(\varepsilon,\mathcal{B}_L,\|\cdot\|_\infty)\le\mathcal{N}(\varepsilon/4,\mathcal{K},\|\cdot\|_2)\log\left(2\left\lceil\frac{2\text{diam}(\mathcal{K})}{\varepsilon}\right\rceil+1\right).$$

Since $\mathcal{H}(\varepsilon,\mathcal{B}_L,L^1(\hat{\mu}_n))\le\mathcal{H}(\varepsilon,\mathcal{B}_L,\|\cdot\|_\infty)$ and $\mathcal{H}(\varepsilon,\mathcal{B}_L,L^1(\hat{\nu}_m))\le\mathcal{H}(\varepsilon,\mathcal{B}_L,\|\cdot\|_\infty)$ then when, respectively, n and m go to infinity, we have

$$\frac{1}{n}\mathcal{H}(\varepsilon,\mathcal{B}_L,L^1(\hat{\mu}_n))\xrightarrow{\mu}0,\qquad\text{and}\qquad\frac{1}{m}\mathcal{H}(\varepsilon,\mathcal{B}_L,L^1(\hat{\nu}_m))\xrightarrow{\nu}0,$$

which leads to the desired result.

Adaptation to other estimators. The above proof can be adapted in a straightforward fashion to $\mathcal{W}(\hat{\mu}_{\text{MoM}},\mu)$ and $\mathcal{W}_{\text{MoU}-\text{diag}}(\hat{\mu}_n,\hat{\nu}_m)$. Indeed, it holds

$$\mathcal{W}(\hat{\mu}_{\text{MoM}},\mu)=\sup_{\phi\in\mathcal{B}_L}\ \operatorname*{med}_{1\le k\le K_{\mathbf{X}}}\left|\overline{\phi}_{\mathbf{X},k}-\mathbb{E}_\mu[\phi]\right|,$$

and

$$\left|\mathcal{W}_{\mathrm{MoU-diag}}(\hat{\mu}_n, \hat{\nu}_m) - \mathcal{W}(\mu,\nu)\right| \leq \sup_{\phi \in \mathcal{B}_L} \underset{\substack{1\leq k\leq K_{\mathbf{X}} \\ 1\leq l \leq K_{\mathbf{Y}}}}{\mathrm{med}} \left|\overline{\phi}_{\mathbf{X},k} - \mathbb{E}_\mu\left[\phi\right] + \mathbb{E}_\nu\left[\phi\right] - \overline{\phi}_{\mathbf{Y},k}\right|.$$

It is then direct to adapt the reasoning from Equation (3). □

A.2 Proof of Proposition 5

Let $\psi \in \mathcal{B}_L$. From Equation (4), we know that $-\mathrm{diam}(\mathcal{K}) \leq \psi(X) \leq \mathrm{diam}(\mathcal{K})$, so that $\psi(X)$ is in particular sub-Gaussian with parameter $\lambda = \mathrm{diam}(\mathcal{K})$. A direct application of Proposition 1 in Laforgue et al. (2020) then gives that for all $\delta \in]0, e^{-4n\sqrt{2\tau_{\mathbf{X}}}}]$ and $K_{\mathbf{X}} = \lceil \sqrt{2\tau_{\mathbf{X}}}n \rceil$, it holds with probability at least $1-\delta$:

$$\left|\mathrm{MoM}_{\mathbf{X}}[\psi] - \mathbb{E}_\mu\left[\psi\right]\right| \leq 4 \ \mathrm{diam}(\mathcal{K})\ \Gamma(\tau_{\mathbf{X}})\ \sqrt{\frac{\log(1/\delta)}{n}}, \tag{5}$$

with $\Gamma\colon \tau_{\mathbf{X}} \mapsto \sqrt{1+\sqrt{2\tau_{\mathbf{X}}}}/\sqrt{1-2\tau_{\mathbf{X}}}$. Using Lemma 1, observe also that $\forall (\phi,\psi) \in \mathcal{B}_L^2$ it holds:

$$\begin{aligned}\left|\mathrm{MoM}_{\mathbf{X}}[\phi] - \mathbb{E}_\mu\left[\phi\right]\right| &\leq \left|\mathrm{MoM}_{\mathbf{X}}[\phi] - \mathrm{MoM}_{\mathbf{X}}[\psi]\right| + \left|\mathbb{E}_\mu\left[\phi\right] - \mathbb{E}_\mu\left[\psi\right]\right| + \left|\mathrm{MoM}_{\mathbf{X}}[\psi] - \mathbb{E}_\mu\left[\psi\right]\right|,\\ &\leq 2\|\phi-\psi\|_\infty + \left|\mathrm{MoM}_{\mathbf{X}}[\psi] - \mathbb{E}_\mu\left[\psi\right]\right|. \end{aligned}\tag{6}$$

Now, let $\zeta > 0$, and $\psi_1, \ldots, \psi_{\mathcal{N}(\zeta,\mathcal{B}_L,\|\cdot\|_\infty)}$ be a ζ-coverage of $\mathcal{B}_L$ with respect to $\|\cdot\|_\infty$. We know from Sriperumbudur et al. (2012) that there exists $C_L > 0$ such that for all $\zeta > 0$ it holds:

$$\log(\mathcal{N}(\zeta, \mathcal{B}_L, \|\cdot\|_\infty)) \leq C_L^2 (1/\zeta)^d \tag{7}$$

From now on, we use $\mathcal{N} = \mathcal{N}(\zeta, \mathcal{B}_L, \|\cdot\|_\infty)$ for notation simplicity. Let ϕ be an arbitrary element of $\mathcal{B}_L$. By definition, there exists $i \leq \mathcal{N}$ such that $\|\phi - \psi_i\|_\infty \leq \zeta$. Equation (6) then gives:

$$\left|\mathrm{MoM}_{\mathbf{X}}[\phi] - \mathbb{E}_\mu\left[\phi\right]\right| \leq 2\zeta + \left|\mathrm{MoM}_{\mathbf{X}}[\psi_i] - \mathbb{E}_\mu\left[\psi_i\right]\right|. \tag{8}$$

Applying Equation (5) to every ψ_i, the union bound gives that with probability at least $1-\delta$ it holds:

$$\sup_{i\leq\mathcal{N}} \left|\mathrm{MoM}_{\mathbf{X}}[\psi_i] - \mathbb{E}_\mu\left[\psi_i\right]\right| \leq 4\ \mathrm{diam}(\mathcal{K})\ \Gamma(\tau_{\mathbf{X}})\ \sqrt{\frac{\log(\mathcal{N}/\delta)}{n}}.$$

Taking the supremum in both sides of Equation (8), it holds with probability at least $1-\delta$:

$$\sup_{\phi\in\mathcal{B}_L} \left|\mathrm{MoM}_{\mathbf{X}}[\phi] - \mathbb{E}_\mu\left[\phi\right]\right| \leq 2\zeta + 4\ \mathrm{diam}(\mathcal{K})\ \Gamma(\tau_{\mathbf{X}})\ \sqrt{\frac{C_L^2\zeta^{-d} + \log(1/\delta)}{n}}.$$

Choosing $\zeta \sim 1/n^{1/(d+2)}$ and breaking the square root finally gives that it holds with probability at least $1-\delta$:

$$\sup_{\phi\in\mathcal{B}_L} \left|\mathrm{MoM}_{\mathbf{X}}[\phi] - \mathbb{E}_\mu\left[\phi\right]\right| \leq \frac{C_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}} + C_2(\tau_{\mathbf{X}})\sqrt{\frac{\log(1/\delta)}{n}},$$

with $C_1(\tau_{\mathbf{X}}) = 2 + C_L C_2(\tau_{\mathbf{X}})$, and $C_2(\tau_{\mathbf{X}}) = 4\ \mathrm{diam}(\mathcal{K})\ \Gamma(\tau_{\mathbf{X}})$.

Adaptation to MoU. From Equation (4), we get that the kernel $h_\phi\colon (X,Y) \mapsto \phi(X) - \phi(Y)$ has finite essential supremum $\|h_\phi(X,Y)\|_\infty \leq \mathrm{diam}(\mathcal{K})$. Using Proposition 4 in Laforgue et al. (2020) with the same reasoning as above leads to the desired result, multiplying constants by factor 2. □

A.3 Proof of Theorem 7

Since $n^{\frac{1}{d+2}+\frac{1-\beta}{2}} \geq C_1(\tau_{\mathbf{X}})/(2C_2(\tau_{\mathbf{X}})(2\tau_{\mathbf{X}})^{\frac{1}{4}})$, then for all $\delta \in]0, e^{-4n\sqrt{2\tau_{\mathbf{X}}}}]$, it holds:

$$\frac{C_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}} \leq C_2(\tau_{\mathbf{X}})\sqrt{\frac{4n\sqrt{2\tau_{\mathbf{X}}}}{n^\beta}} \leq C_2(\tau_{\mathbf{X}})\sqrt{\frac{\log(1/\delta)}{n^\beta}}.$$

One then has:

$$\mathcal{W}(\hat{\mu}_{\mathrm{MoM}}, \mu) \geq 0 \geq \frac{C_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}} - C_2(\tau_{\mathbf{X}})\sqrt{\frac{\log(1/\delta)}{n^\beta}}.$$

Combining with the first results of Proposition 4, for all $\delta \in]0, e^{-4n\sqrt{2\tau_{\mathbf{X}}}}]$, it holds with probability at least $1-\delta$:

$$\left|\mathcal{W}(\hat{\mu}_{\mathrm{MoM}}, \mu) - \frac{C_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}}\right| \leq C_2(\tau_{\mathbf{X}})\sqrt{\frac{\log(1/\delta)}{n^\beta}}.$$

Reverting the inequation gives that it holds

$$\mathbb{P}\left\{\left|\mathcal{W}(\hat{\mu}_{\mathrm{MoM}}, \mu) - \frac{C_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}}\right| > t\right\} \leq e^{-n^\beta t^2/{C_2}^2(\tau_{\mathbf{X}})}, \tag{9}$$

for all t such that

$$t \geq (32\ \tau_{\mathbf{X}})^{1/4} C_2(\tau_{\mathbf{X}})\ \sqrt{n^{1-\beta}} = \frac{(32\ \tau_{\mathbf{X}})^{1/4}}{\sqrt{\tau_{\mathbf{X}}}} C_2(\tau_{\mathbf{X}})\sqrt{n^{1-\beta}\ \frac{n_{\mathcal{O}}}{n}}. \tag{10}$$

One may finally use that for a nonnegative random variable it holds:

$$\begin{aligned}
\mathbb{E}\left|\mathcal{W}(\hat{\mu}_{\mathrm{MoM}}, \mu) - \frac{C_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}}\right| &= \int_0^\infty \mathbb{P}\left\{\left|\mathcal{W}(\hat{\mu}_{\mathrm{MoM}}, \mu) - \frac{C_1(\tau_{\mathbf{X}})}{n^{1/(d+2)}}\right| > t\right\} dt, \\
&\leq \int_0^{\frac{(32\ \tau_{\mathbf{X}})^{1/4}}{\sqrt{\tau_{\mathbf{X}}}} C_{\mathcal{O}} C_2(\tau_{\mathbf{X}})\sqrt{n^{\alpha_{\mathcal{O}}-\beta}}} 1 dt + \int_0^\infty e^{-n^\beta t^2/{C_2}^2(\tau_{\mathbf{X}})} dt, \\
&\leq \frac{(32\ \tau_{\mathbf{X}})^{1/4}}{\sqrt{\tau_{\mathbf{X}}}}\ \frac{C_{\mathcal{O}} C_2(\tau_{\mathbf{X}})}{n^{(\beta-\alpha_{\mathcal{O}})/2}} + \frac{\sqrt{\pi}\ C_2(\tau_{\mathbf{X}})}{2\ n^{\beta/2}}. \\
&= 2\ (2/\tau_{\mathbf{X}})^{1/4}\ \frac{C_{\mathcal{O}} C_2(\tau_{\mathbf{X}})}{n^{(\beta-\alpha_{\mathcal{O}})/2}} + \frac{\sqrt{\pi}\ C_2(\tau_{\mathbf{X}})}{2\ n^{\beta/2}}.
\end{aligned} \tag{11}$$

Where the second line holds thanks to Assumption 6.

Adaptation to MoU. The adaptation is straightforward, up to Equation (10), that now writes:

$$\begin{aligned}
t &\geq 2 \times (32(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}}))^{1/4} C_2(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}})\ \sqrt{n^{1-\beta}}, \\
&= 2 \times \frac{(32(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}}))^{1/4}}{\sqrt{\tau_{\mathbf{X}} + \tau_{\mathbf{Y}}}} C_2(\tau_{\mathbf{X}} + \tau_{\mathbf{Y}})\sqrt{n^{1-\beta}\ \left(\frac{n_{\mathcal{O}}}{n} + \frac{m_{\mathcal{O}}}{m}\right)}.
\end{aligned}$$

Using Assumption 6 on both samples $\mathbf{X}$ and $\mathbf{Y}$, it leads to the desired results. ☐

B Additional material of the numerical part

In this part, we introduce algorithms and additional experiments that could not be contained in the paper due to space constraints.

B.1 Additional algorithms

Here, algorithms to compute $\mathcal{W}_{\text{MoU-diag}}(\mu_n, \nu_n)$ and $\mathcal{W}_{\text{MoU}}(\mu_n, \nu_n)$ are displayed.

Algorithm 1 Computation of $\mathcal{W}_{\text{MoU-diag}}(\mu_n, \nu_n)$.

Initialization: η, the learning rate. c, the clipping parameter. w_0 the initial weigths.

for $t = 0, \ldots, n_{\text{iter}}$ **do**

  Sample $K = K_{\mathbf{X}} \wedge K_{\mathbf{Y}}$ disjoint blocks $\mathcal{B}^{\mathbf{XY}}_{1,1}, \mathcal{B}^{\mathbf{XY}}_{2,2}, \ldots, \mathcal{B}^{\mathbf{XY}}_{k,k}, \ldots \mathcal{B}^{\mathbf{XY}}_{K,K}$ from a sampling scheme

  Find the median blocks $\mathcal{B}^{\mathbf{XY}}_{med}$

$$G_w \longleftarrow \lfloor K/n \rfloor \sum_{(i,j) \in \mathcal{B}^{\mathbf{XY}}_{med}} \nabla_w \left[\phi_w(X_i) - \phi_w(Y_j)\right]$$

  7.1 $w \leftarrow w + \eta \times \text{RMSProp}(w, G_w)$

  7.2 $w \leftarrow \text{clip}(w, -c, c)$

end for

Output: w, $\widetilde{\mathcal{W}}_{\text{MoU-diag}}$, ϕ_w.

Algorithm 2 Computation of $\mathcal{W}_{\text{MoU}}(\mu_n, \nu_n)$.

Initialization: η, the learning rate. c, the clipping parameter. w_0 the initial weigths.

for $t = 0, \ldots, n_{\text{iter}}$ **do**

  Sample $K_{\mathbf{X}} \times K_{\mathbf{Y}}$ disjoint blocks $\mathcal{B}^{\mathbf{XY}}_{1,1}, \ldots, \mathcal{B}^{\mathbf{XY}}_{k,l}, \ldots \mathcal{B}^{\mathbf{XY}}_{K_{\mathbf{X}},K_{\mathbf{Y}}}$ from a sampling scheme

  Find the median blocks $\mathcal{B}^{\mathbf{XY}}_{med}$

$$G_w \longleftarrow \lfloor K_{\mathbf{X}}/n \rfloor \times \lfloor K_{\mathbf{Y}}/m \rfloor \sum_{(i,j) \in \mathcal{B}^{\mathbf{XY}}_{med}} \nabla_w \left[\phi_w(X_i) - \phi_w(Y_j)\right]$$

  7.1 $w \leftarrow w + \eta \times \text{RMSProp}(w, G_w)$

  7.2 $w \leftarrow \text{clip}(w, -c, c)$

end for

Output: w, $\widetilde{\mathcal{W}}_{\text{MoU}}$, ϕ_w.

B.2 Additional experiments

In this part, numerical results for $\widetilde{\mathcal{W}}_{\text{MoU}}$ and $\widetilde{\mathcal{W}}_{\text{MoM}}$, related to the Section 4.2 of the paper, are displayed. Results of both experiments, depicted in Figures 1 and 2, are quite similar, which can be explained by the relative simplicity of the problem.

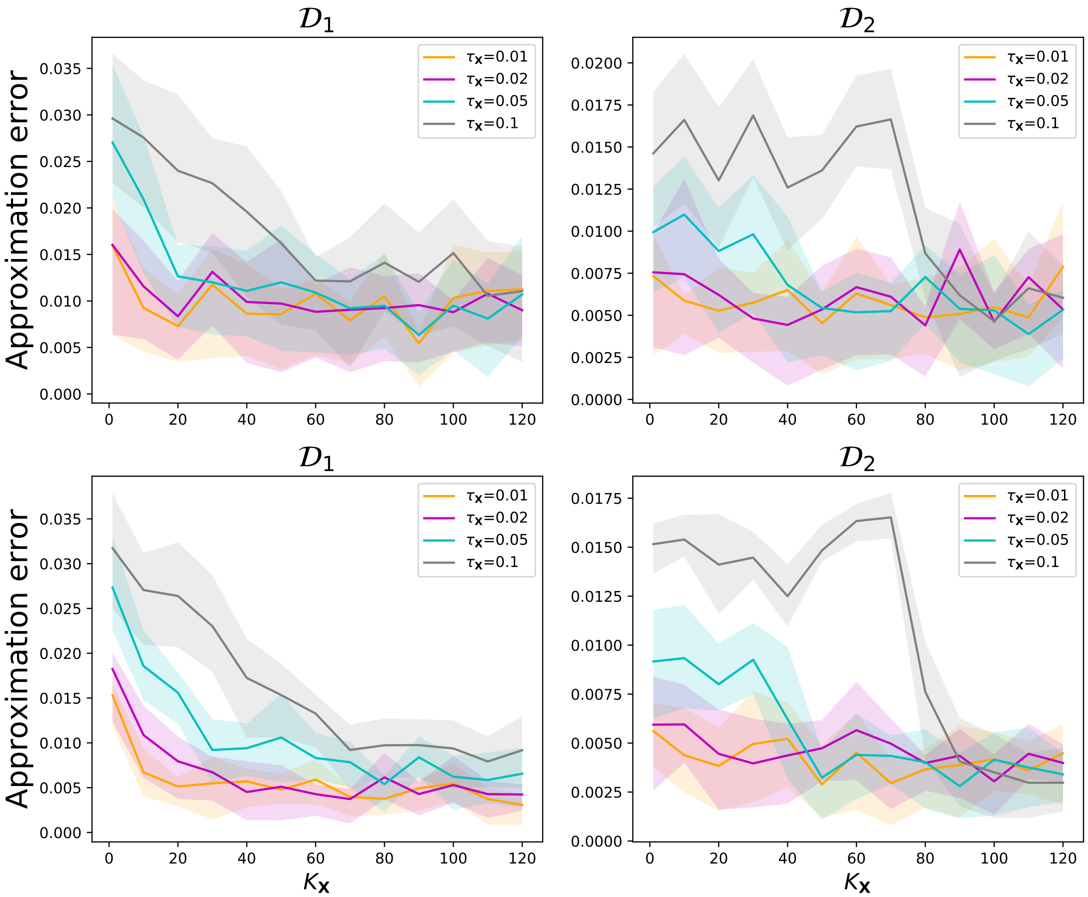

Figure 1: $\widetilde{\mathcal{W}}_{\text{MoU}}$ (top) and $\widetilde{\mathcal{W}}_{\text{MoM}}$ (bottom) over $K_{\mathbf{X}}$ for different fractions of anomalies τ_X on $\mathcal{D}_1$ (left) and $\mathcal{D}_2$ (right).

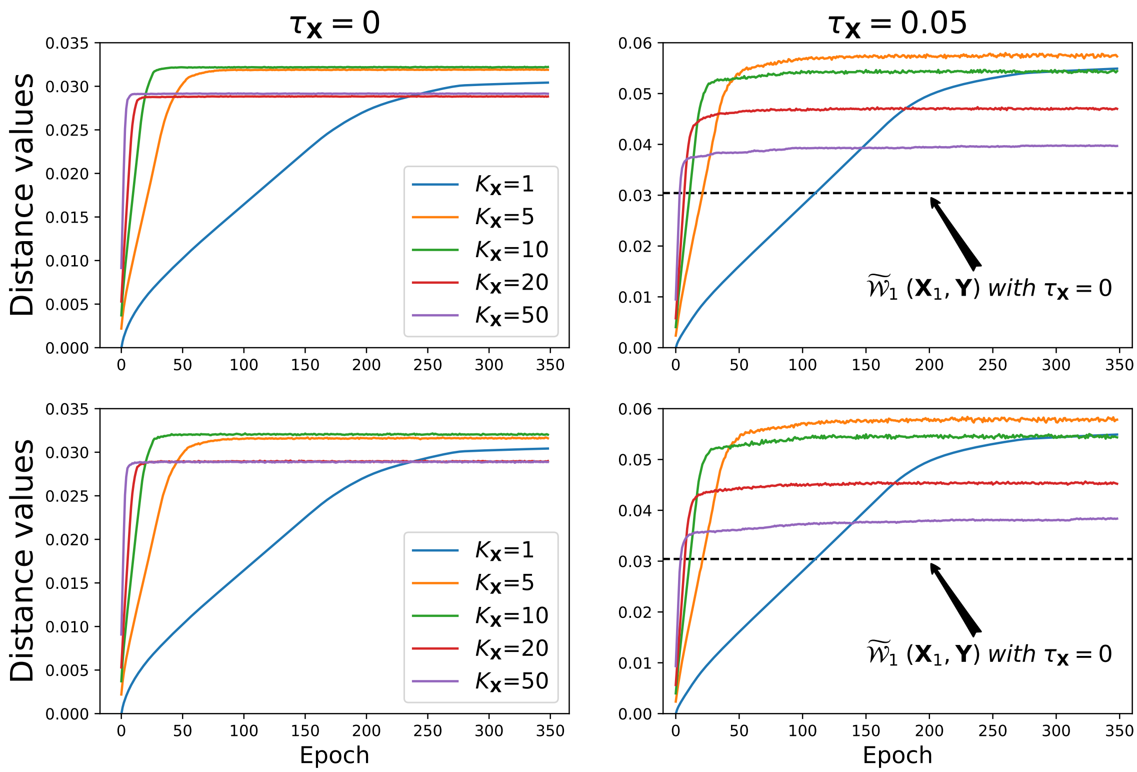

Figure 2: Convergence of $\widetilde{\mathcal{W}}_{\mathrm{MoU}}$ (top) and $\widetilde{\mathcal{W}}_{\mathrm{MoM}}$ (bottom) without anomalies (left) and with 5% anomalies (right) for different $K_{\mathbf{X}}$.

References

Kolmogorov, A. N. and Tihomirov, V. M. (1961). On the empirical estimation of integral probability metrics. *American Mathematical Society Translations 2*, 17:277–364.

Laforgue, P., Staerman, G., and Clémençon, S. (2020). How robust is the median-of-means? concentration bounds in presence of outliers. arxiv.org/abs/2006.05240.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. (2012). On the empirical estimation of integral probability metrics. *Electron. J. Statist.*, 6:1550–1599.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.