MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means

Matthieu Lerasle, Zoltán Szabó, Gaspar Massiot, Eric Moulines

MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means

Matthieu Lerasle [1] **Zoltán Szabó** [2] **Gaspar Massiot** [3] **Eric Moulines** [2]

Abstract

Mean embeddings provide an extremely flexible and powerful tool in machine learning and statistics to represent probability distributions and define a semi-metric (MMD, maximum mean discrepancy; also called N-distance or energy distance), with numerous successful applications. The representation is constructed as the expectation of the feature map defined by a kernel. As a mean, its classical empirical estimator, however, can be arbitrary severely affected even by a single outlier in case of unbounded features. To the best of our knowledge, unfortunately even the consistency of the existing few techniques trying to alleviate this serious sensitivity bottleneck is unknown. In this paper, we show how the recently emerged principle of median-of-means can be used to design minimax-optimal estimators for kernel mean embedding and MMD, with finite-sample strong outlier-robustness guarantees.

1. Introduction

Kernel methods form the backbone of a tremendous number of successful applications in machine learning thanks to their power in capturing complex relations (Aronszajn, 1950; Shawe-Taylor & Cristianini, 2004; Schölkopf & Smola, 2002; Berlinet & Thomas-Agnan, 2004; Steinwart & Christmann, 2008). The main idea behind these techniques is to map the data points to a feature space (RKHS, reproducing kernel Hilbert space) determined by the kernel, and apply linear methods in the feature space, without the need to explicitly compute the map.

One crucial component contributing to this flexibility and efficiency (beyond the solid theoretical foundations) is the versatility of domains where kernels exist; examples include trees (Collins & Duffy, 2001; Kashima & Koyanagi, 2002), time series (Cuturi, 2011), strings (Lodhi et al., 2002), mixture models, hidden Markov models or linear dynamical systems (Jebara et al., 2004), sets (Haussler, 1999; Gärtner et al., 2002), fuzzy domains (Guevara et al., 2017), distributions (Hein & Bousquet, 2005; Martins et al., 2009; Muandet et al., 2011), groups (Cuturi et al., 2005) such as specific constructions on permutations (Jiao & Vert, 2016), or graphs (Vishwanathan et al., 2010; Kondor & Pan, 2016).

Given a kernel-enriched domain $(\mathcal{X}, K)$ one can represent probability distributions on $\mathcal{X}$ as a mean

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_K, \quad \varphi(x) := K(\cdot, x),$$

which is a point in the RKHS determined by K. This representation called mean embedding (Berlinet & Thomas-Agnan, 2004; Smola et al., 2007) induces (Smola et al., 2007; Gretton et al., 2012) a semi-metric[1] on distributions named as maximum mean discrepancy (MMD)

$$\mathrm{MMD}_K(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_K}. \tag{1}$$

Specializing the kernel, classical integral transforms widely used in probability theory and statistics can be recovered by $\mu_{\mathbb{P}}$; for example $K(x, y) = e^{\langle x, y \rangle}$ gives the well-known moment-generating function, the $K(x, y) = e^{\gamma \|x-y\|_2^2}$ ($\gamma > 0$) choice results in the Weierstrass transform. As it has been shown (Sejdinovic et al., 2013) energy distance (Baringhaus & Franz, 2004; Székely & Rizzo, 2004; 2005)—also known as N-distance (Zinger et al., 1992; Klebanov, 2005) in the statistical literature—coincides with MMD.

Mean embedding and maximum mean discrepancy have been applied successfully, in kernel Bayesian inference (Song et al., 2011; Fukumizu et al., 2013), approximate Bayesian computation (Park et al., 2016), model criticism (Lloyd et al., 2014; Kim et al., 2016), two-sample (Baringhaus & Franz, 2004; Székely & Rizzo, 2004; 2005; Harchaoui et al., 2007; Gretton et al., 2012), independence (Gretton et al., 2008; Pfister et al., 2017) and goodness-of-fit testing (Jitkrittum et al., 2017; Balasubramanian et al.,

[1]Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud; CNRS, Université Paris Saclay, France [2]Center for Applied Mathematics (CMAP), École Polytechnique, Palaiseau, France [3]ONERA - The French Aerospace Lab, Chemin de la Hunière - BP 80100, 91123 Palaiseau Cedex, France. Correspondence to: Matthieu Lerasle <matthieu.lerasle@polytechnique.edu>, Zoltán Szabó <zoltan.szabo@polytechnique.edu>.

[1]Fukumizu et al. (2008); Sriperumbudur et al. (2010) provide conditions when MMD_K is a metric, i.e. μ is injective.

2017), domain adaptation (Zhang et al., 2013) and generalization (Blanchard et al., 2017), probabilistic programming (Schölkopf et al., 2015), post selection inference (Yamada et al., 2016), distribution classification (Muandet et al., 2011; Zaheer et al., 2017) and regression (Szabó et al., 2016; Law et al., 2018), causal discovery (Mooij et al., 2016; Pfister et al., 2017) or topological data analysis (Kusano et al., 2016), among many others; Muandet et al. (2017) provide an in-depth review on the topic.

Crucial to the success of these applications is the efficient and robust approximation of the mean embedding and MMD. As a mean, the most natural approach to estimate $\mu_{\mathbb{P}}$ is the empirical average. Plugging this estimate to Eq. (1) produces directly an approximation of MMD, which can also be made unbiased (by a small correction) or approximated recursively. These are the V-statistic, U-statistic and online approaches (Gretton et al., 2012). Kernel mean shrinkage estimators (Muandet et al., 2016) represent an other successful direction: they improve the efficiency of the mean embedding estimation by taking into account the Stein phenomenon. Minimax results have recently been established: the optimal rate of mean embedding estimation given N samples from $\mathbb{P}$ is $N^{-1/2}$ (Tolstikhin et al., 2017) for discrete measures and the class of measures with infinitely differentiable density when K is a continuous, shift-invariant kernel on $\mathcal{X} = \mathbb{R}^d$. For MMD, using N_1 and N_2 samples from $\mathbb{P}$ and $\mathbb{Q}$, it is $N_1^{-1/2} + N_2^{-1/2}$ (Ilya Tolstikhin & Schölkopf, 2016) in case of radial universal kernels defined on $\mathcal{X} = \mathbb{R}^d$.

A critical property of an estimator is its robustness w.r.t. contaminated data, outliers which are omnipresent in currently available massive and heterogenous datasets. To the best of our knowledge, systematically *designing outlier-robust mean embedding and MMD estimators* has hardly been touched in the literature; this is the focus of the current paper. The issue is particularly serious in case of unbounded kernels when for example even a single outlier can ruin completely a classical empirical average based estimator. Examples for unbounded kernels are the exponential kernel (see the example above about moment-generating functions), polynomial kernel, string or graph kernels.

Existing related techniques comprise robust kernel density estimation (KDE) (Kim & Scott, 2012): the authors elegantly combine ideas from the KDE and M-estimator literature to arrive at a robust KDE estimate of density functions. They assume that the underlying smoothing kernels[2] are shift-invariant on $\mathcal{X} = \mathbb{R}^d$ and reproducing, and interpret KDE as a weighted mean in $\mathcal{H}_K$. The idea has been (i) adapted to construct outlier-robust covariance operators in RKHSs in the context of kernel canonical correlation analysis (Alam et al., 2017), and (ii) relaxed to general Hilbert spaces (Sinova et al., 2018). Unfortunately, the consistency of the investigated empirical M-estimators is unknown, except for finite-dimensional feature maps (Sinova et al., 2018).

To achieve our goal, we leverage the idea of Median-Of-meaNs (MON). Intuitively, MONs replace the linear operation of expectation with the median of averages taken over non-overlapping blocks of the data, in order to get a robust estimate thanks to the median step. MONs date back to Jerrum et al. (1986); Alon et al. (1999); Nemirovski & Yudin (1983) for the estimation of the mean of real-valued random variables. Their concentration properties have been recently studied by Devroye et al. (2016); Minsker & Strawn (2017); the idea of systematically investigating estimators from a deviation point of view goes back to Catoni (2012). The point is that such a study can then be used to extend from the estimation of the mean of real-valued random variables to more general prediction problems in learning theory by minmax aggregation of pairwise comparisons (Audibert & Catoni, 2011).

Independently of Audibert & Catoni (2011), there have been several attempts to extend the usage of MON estimators from $\mathbb{R}$ to more general settings. For example, Minsker (2015); Minsker & Strawn (2017) consider the problem of estimating the mean of a Banach-valued random variable using "geometrical" MONs. The estimators in (Minsker, 2015; Minsker & Strawn, 2017) are computationally tractable but the deviation bounds are slightly suboptimal compared to those one can prove for the empirical mean under subgaussian assumptions. Lugosi & Mendelson (2016; 2017) proposed to combine the classical MON estimators on $\mathbb{R}$ in a "tournament" procedure that can be seen as a Le Cam type aggregation of tests (Le Cam, 1973). The remarkable achievement in (Lugosi & Mendelson, 2016; 2017) is that they were able to obtain optimal deviation bounds for the resulting estimator using the powerful so-called "small-ball method" of Mendelson (2015). Catoni & Giulini (2017) have recently improved the constants in (Lugosi & Mendelson, 2016; 2017), reaching almost optimal constants in the deviation bounds for the estimation of the mean in $\mathbb{R}^d$ and least-squares regression under a slightly stronger moment assumption.

In this paper, we extend the results of Lugosi & Mendelson (2017) to mean embedding and MMD estimation in RKHS. In order to attain this goal, we modify the aggregation step of Lugosi & Mendelson (2016; 2017) using the minmax formulation of Audibert & Catoni (2011). As a result we get a conceptually simpler aggregation, which simultaneously serves as a novel application of the small-ball method. A further statistically appealing property of our estimator is

[2] Smoothing kernels extensively studied in the non-parametric statistical literature (Györfi et al., 2002) are assumed to be non-negative functions integrating to one.

that it satisfies rate-optimal subgaussian deviation bounds under minimal assumptions (assuming that the underlying covariance operators are of trace class).

Our **contributions** can be summarized as follows:

- We design novel mean embedding and MMD estimators based on the MON principle.
- We establish their finite-sample and outlier-robustness properties. The obtained MONK estimators are consistent, their convergence speed match the discussed minimax rates, and they are tolerant to excessive contamination.

Section 2 contains definitions and problem formulation. Our results are given in Section 3. Section 4 is about proofs.

2. Definitions & Problem Formulation

In this section, we formally introduce the goal of our paper.

Notations: $\mathbb{Z}^+$ is the set of positive integers. $[M] := \{1, \dots, M\}$, $u_S := (u_m)_{m\in S}$, $S \subseteq [M]$. For a set S, $|S|$ denotes its cardinality. $\mathbb{E}$ stands for expectation. $\mathrm{med}_{q\in[Q]}\{z_q\}$ is the median of the $(z_q)_{q\in[Q]}$ numbers. Let $\mathcal{X}$ be a separable topological space endowed with the Borel σ-field, $x_{1:N}$ denotes a sequence of i.i.d. random variables on $\mathcal{X}$ with law $\mathbb{P}$ (shortly, $x_{1:N} \sim \mathbb{P}$). $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous (reproducing) kernel on $\mathcal{X}$, $\mathcal{H}_K$ is the reproducing kernel Hilbert space associated to K; $\langle \cdot, \cdot \rangle_K := \langle \cdot, \cdot \rangle_{\mathcal{H}_K}$, $\|\cdot\|_K := \|\cdot\|_{\mathcal{H}_K}$.[3] The reproducing property of the kernel means that evaluation of functions in $\mathcal{H}_K$ can represented by inner products

$$f(x) = \langle f, K(\cdot, x)\rangle_K \quad (\forall x \in \mathcal{X}, f \in \mathcal{H}_K).$$

The mean embedding of a probability measure $\mathbb{P}$ is defined as

$$\mu_{\mathbb{P}} = \int_{\mathcal{X}} K(\cdot, x)\mathrm{d}\mathbb{P}(x) \in \mathcal{H}_K, \tag{2}$$

where the integral is meant in Bochner sense; $\mu_{\mathbb{P}}$ exists iff

$$\int_{\mathcal{X}} \|K(\cdot, x)\|_K \,\mathrm{d}\mathbb{P}(x) = \int_{\mathcal{X}} \sqrt{K(x, x)}\mathrm{d}\mathbb{P}(x) < \infty.$$

It is well-known that the mean embedding has mean-reproducing property

$$\mathbb{P}f := \mathbb{E}_{x\sim\mathbb{P}} f(x) = \langle f, \mu_{\mathbb{P}}\rangle_K \quad (\forall f \in \mathcal{H}_K), \tag{3}$$

and

$$\mu_{\mathbb{P}} = \mathrm{argmin}_{f\in\mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot, x)\|_K^2 \,\mathrm{d}\mathbb{P}(x). \tag{4}$$

[3] $\mathcal{H}_K$ is separable by the separability of $\mathcal{X}$ and the continuity of K (Steinwart & Christmann, 2008, Lemma 4.33).

The solution of this problem can be obtained by solving the following minmax optimization

$$\mu_{\mathbb{P}} = \mathrm{argmin}_{f\in\mathcal{H}_K} \sup_{g\in\mathcal{H}_K} J(f, g), \tag{5}$$
$$J(f, g) = \mathbb{E}_{x\sim\mathbb{P}}\left[\|f - K(\cdot, x)\|_K^2 - \|g - K(\cdot, x)\|_K^2\right].$$

The equivalence of (4) and (5) is obvious since the expectation is linear. Nevertheless, this equivalence is essential in the construction of our estimators because we will below replace the expectation by a non-linear estimator of this quantity. More precisely, the unknown expectations are computed by using the Median-of-meaN estimator (MON). Given a partition of the dataset into blocks, the MON estimator is the median of the empirical means over each block. MON estimators are naturally robust thanks to the median step.

More precisely, the procedure goes as follows. For any map $f : \mathcal{X} \to \mathbb{R}$ and any non-empty subset $S \subseteq [N]$, denote by $\mathbb{P}_S := \frac{1}{|S|}\sum_{i\in S} \delta_{x_i}$ the empirical measure associated to the subset x_S and $\mathbb{P}_S f = \frac{1}{|S|}\sum_{i\in S} f(x_i)$; we will use the shorthand $\mu_S := \mu_{\mathbb{P}_S}$. Assume that $N \in \mathbb{Z}^+$ is divisible by $Q \in \mathbb{Z}^+$ and let $(S_q)_{q\in[Q]}$ denote a partition of $[N]$ into subsets with the same cardinality $|S_q| = N/Q$ ($\forall q \in [Q]$). The Median Of meaN (MON) is defined as

$$\mathrm{MON}_Q[f] := \mathrm{med}_{q\in[Q]}\left\{\mathbb{P}_{S_q} f\right\} = \mathrm{med}_{q\in[Q]}\left\{\left\langle f, \mu_{S_q}\right\rangle_K\right\},$$

where the second equality is a consequence of the mean-reproducing property of $\mu_{\mathbb{P}}$ [Eq. (3)]. We define the MON-based estimator associated to kernel K (MONK) as

$$\hat{\mu}_{\mathbb{P},Q} = \hat{\mu}_{\mathbb{P},Q}(x_{1:N}) \in \mathrm{argmin}_{f\in\mathcal{H}_K} \sup_{g\in\mathcal{H}_K} \tilde{J}(f, g), \tag{6}$$
$$\tilde{J}(f, g) = \mathrm{MON}_Q\left[\|f - K(\cdot, x)\|_K^2 - \|g - K(\cdot, x)\|_K^2\right].$$

One can use the mean embedding [Eq. (2)] to get a semi-metric on probability measures: the maximum mean discrepancy (MMD) of $\mathbb{P}$ and $\mathbb{Q}$ is

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) := \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_K = \sup_{f\in B_K} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\rangle_K,$$

where $B_K = \{f \in \mathcal{H}_K : \|f\|_K \le 1\}$ is the closed unit ball around the origin in $\mathcal{H}_K$. The 2nd equality shows that MMD is a specific integral probability metric (Müller, 1997; Zolotarev, 1983). Assume that we have access to $x_{1:N} \sim \mathbb{P}$, $y_{1:N} \sim \mathbb{Q}$ samples.[4] Denote by $\mathbb{P}_{S,x} := \frac{1}{|S|}\sum_{i\in S} \delta_{x_i}$ the empirical measure associated to the subset x_S ($\mathbb{P}_{S,y}$ is defined similarly for y), $\mu_{S_q,\mathbb{P}} := \mu_{\mathbb{P}_{S_q,x}}$, $\mu_{S_q,\mathbb{Q}} := \mu_{\mathbb{P}_{S_q,y}}$.

[4] The size of the two samples is assumed to be the same for simplicity.

We propose the following MON-based MMD estimator

$$\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]} \left\{ \left\langle f, \mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}} \right\rangle_K \right\}.$$

Our **goal** is to lay down the foundations and study the finite-sample behaviour, establish the consistency and outlier-robustness properties of the $\hat{\mu}_{\mathbb{P},Q}$ and $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$ MONK estimators.

A **few more notations** will be needed throughout the paper. $S_1 \backslash S_2$ is the difference of set S_1 and S_2. For any linear operator $A : \mathcal{H}_K \to \mathcal{H}_K$, denote by $\|A\| := \sup_{0\neq f\in\mathcal{H}_K} \|Af\|_K / \|f\|_K$ the operator norm of A. Let $\mathcal{L}(\mathcal{H}_K) = \{A : \mathcal{H}_K \to \mathcal{H}_K \text{ linear operator} : \|A\| < \infty\}$ be the space of bounded linear operators. For any $A \in \mathcal{L}(\mathcal{H}_K)$, let $A^* \in \mathcal{L}(\mathcal{H}_K)$ denote the adjoint of A, that is the operator such that $\langle Af, g\rangle_K = \langle f, A^*g\rangle_K$ for all $f, g \in \mathcal{H}_K$. An operator $A \in \mathcal{L}(\mathcal{H}_K)$ is called non-negative if $\langle Af, f\rangle_K \geq 0$ for all $f \in \mathcal{H}_K$. By the separability of $\mathcal{H}_K$, there exists a countable orthonormal basis (ONB) $(e_i)_{i\in I}$ in $\mathcal{H}_K$. An operator $A \in \mathcal{L}(\mathcal{H}_K)$ is called (i) trace-class if $\|A\|_1 := \sum_{i\in I} \left\langle (A^*A)^{1/2} e_i, e_i \right\rangle_K < \infty$ and in this case $\mathrm{Tr}(A) := \sum_{i\in I} \langle Ae_i, e_i\rangle_K < \infty$,[5] (ii) Hilbert-Schmidt if $\|A\|_2^2 := \mathrm{Tr}(A^*A) = \sum_{i\in I} \langle Ae_i, Ae_i\rangle_K < \infty$. Denote by $\mathcal{L}_1(\mathcal{H}_K) := \{A \in \mathcal{L}(\mathcal{H}_K) : \|A\|_1 < \infty\}$ and $\mathcal{L}_2(\mathcal{H}_K) := \{A \in \mathcal{L}(\mathcal{H}_K) : \|A\|_2 < \infty\}$ the class of trace-class and (Hilbert) space of Hilbert-Schmidt operators on $\mathcal{H}_K$, respectively.[6] The tensor product of $a, b \in \mathcal{H}_K$ is

$$(a \otimes b)(c) = a \langle b, c\rangle_K, \quad (\forall c \in \mathcal{H}_K).$$

$a\otimes b \in \mathcal{L}_2(\mathcal{H}_K)$, $\mathcal{L}_2(\mathcal{H}_K) \cong \mathcal{H}_K \otimes \mathcal{H}_K$ where $\otimes$ is the tensor product of Hilbert spaces and $\|a \otimes b\|_2 = \|a\|_K \|b\|_K$. Whenever

$$\int_{\mathcal{X}} \|K(\cdot,x) \otimes K(\cdot,x)\|_2 \,\mathrm{d}\mathbb{P}(x) = \int_{\mathcal{X}} K(x,x)\mathrm{d}\mathbb{P}(x) < \infty,$$

let $\Sigma_{\mathbb{P}}$ denote the covariance operator

$$\Sigma_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}} \left([K(\cdot,x) - \mu_{\mathbb{P}}] \otimes [K(\cdot,x) - \mu_{\mathbb{P}}] \right) \in \mathcal{L}_2(\mathcal{H}_K),$$

where the expectation (integral) is again meant in Bochner sense. $\Sigma_{\mathbb{P}}$ is non-negative, self-adjoint, moreover it has covariance-reproducing property

$$\langle f, \Sigma_{\mathbb{P}} f\rangle_K = \mathbb{E}_{x\sim\mathbb{P}}[f(x) - \mathbb{P}f]^2. \tag{7}$$

It is known that $\|A\| \leq \|A\|_2 \leq \|A\|_1$.

[5] If A is a non-negative and self-adjoint, then A is trace class iff $\mathrm{Tr}(A) < \infty$; this will hold for the covariance operator ($\Sigma_{\mathbb{P}}$).

[6] One can show that these definitions are independent of the choice of the ONB $(e_i)_{i\in I}$.

3. Results

Below we present our main results on the MONK estimators.

We allow that from the $x_{1:N}$ samples N_c is *arbitrarily* corrupted ($(x_{n_j})_{j=1}^{N_c}$); the resulting dataset is used for estimation.[7] The measure of contamination can be (almost) half of the block size, in other words, $\exists \delta \in \left(0, \frac{1}{2}\right]$ such that $N_c \leq Q(\frac{1}{2} - \delta)$.[8]

Theorem 1 (Consistency & outlier-robustness of $\hat{\mu}_{\mathbb{P},Q}$)**.** *Assume that $\Sigma_{\mathbb{P}} \in \mathcal{L}_1(\mathcal{H}_K)$, and let $c_1 = 2\left(1+\sqrt{2}\right)$. Then, with probability at least $1 - e^{-\frac{Q\delta^2}{18}}$*

$$\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq c_1 \max\left(\sqrt{\frac{3\,\|\Sigma_{\mathbb{P}}\|\,Q}{\delta N}}, \frac{12}{\delta}\sqrt{\frac{Tr\,(\Sigma_{\mathbb{P}})}{N}} \right).$$

Theorem 2 (Consistency & outlier-robustness of $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})$)**.** *Assume that $\Sigma_{\mathbb{P}}$ and $\Sigma_{\mathbb{Q}} \in \mathcal{L}_1(\mathcal{H}_K)$. Then, with probability at least $1 - e^{-\frac{Q\delta^2}{18}}$*

$$\left| \widehat{MMD}_Q(\mathbb{P},\mathbb{Q}) - MMD_Q(\mathbb{P},\mathbb{Q}) \right| \leq$$
$$\leq 2\max\left(\sqrt{\frac{3(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)Q}{\delta N}}, \frac{12}{\delta}\sqrt{\frac{Tr\,(\Sigma_{\mathbb{P}})+Tr\,(\Sigma_{\mathbb{Q}})}{N}} \right).$$

Remarks:

- These finite-sample guarantees show that the MONK estimators
  - have optimal $N^{-1/2}$-rate—by recalling Ilya Tolstikhin & Schölkopf (2016); Tolstikhin et al. (2017)'s discussed results—, and
  - they are robust to outliers, providing consistent estimates with high probability even under excessive contamination.
- Theorem 2 also implies that fixing the trace of the covariance operators of $\mathbb{P}$ and $\mathbb{Q}$, the MON-based MMD estimator can separate $\mathbb{P}$ and $\mathbb{Q}$ at the rate of $N^{-1/2}$.

4. Proofs

In this section we provide the proofs of our results. Theorem 1 (Theorem 2) is in the focus of Section 4.1 (Section 4.2); few lemmas are relegated to the supplement.

4.1. Proof of Theorem 1

The structure of the proof is as follows:

1. We show that $\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq (1+\sqrt{2}) r_{Q,N}$, where

$$r_{Q,N} = \sup_{f\in B_K} \mathrm{MOM}_Q\Big[\underbrace{\langle f, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K}_{f(x)-\mathbb{P}f} \Big].$$

[7] In MMD estimation $\{(x_{n_j}, y_{n_j})\}_{j=1}^{N_c}$ can be contaminated.

[8] $N_c = 0$, or equivalently the $\delta = \frac{1}{2}$ case, means clean data.

In other words, the analysis can be reduced to B_K.

2. Then $r_{Q,N}$ is bounded using empirical processes.

Step-1: Since $\mathcal{H}_K$ is a Euclidean space, for any $f \in \mathcal{H}_K$

$$\|f - K(\cdot,x)\|_K^2 - \|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2 = \\ = \|f - \mu_{\mathbb{P}}\|_K^2 - 2\langle f - \mu_{\mathbb{P}}, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K . \quad (8)$$

Hence, by denoting $e = \hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}$, $\tilde{g} = g - \mu_{\mathbb{P}}$ we get

$$\begin{aligned}
&\|e\|_K^2 - 2r_{Q,N}\|e\|_K \\
&\overset{(a)}{\leq} \|e\|_K^2 - 2\mathrm{MON}_Q\left[\left\langle \frac{e}{\|e\|_K}, K(\cdot,x) - \mu_{\mathbb{P}}\right\rangle_K\right]\|e\|_K \\
&\overset{(b)}{\leq} \mathrm{MON}_Q\left[\|e\|_K^2 - 2\left\langle \frac{e}{\|e\|_K}, K(\cdot,x) - \mu_{\mathbb{P}}\right\rangle_K\|e\|_K\right] \\
&\overset{(c)}{\leq} \mathrm{MON}_Q\left[\|\hat{\mu}_{\mathbb{P},Q} - K(\cdot,x)\|_K^2 - \|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2\right] \\
&\overset{(d)}{\leq} \sup_{g\in\mathcal{H}_K} \mathrm{MON}_Q\left[\|\hat{\mu}_{\mathbb{P},Q} - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\right] \\
&\overset{(e)}{\leq} \sup_{g\in\mathcal{H}_K} \mathrm{MON}_Q\left[\|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\right] \\
&\overset{(f)}{=} \sup_{g\in\mathcal{H}_K}\left\{2\mathrm{MOM}_Q\Big[\underbrace{\langle\tilde{g}, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K}_{\|\tilde{g}\|_K\left\langle\frac{\tilde{g}}{\|\tilde{g}\|_K}, K(\cdot,x)-\mu_{\mathbb{P}}\right\rangle_K}\Big] - \|\tilde{g}\|_K^2\right\} \\
&\overset{(g)}{=} \sup_{g\in\mathcal{H}_K}\left\{2\|\tilde{g}\|_K r_{Q,N} - \|\tilde{g}\|_K^2\right\} \overset{(h)}{\leq} r_{Q,N}^2, \qquad (9)
\end{aligned}$$

where we used in (a) the definition of $r_{Q,N}$, (b) the linearity[9] of $\mathrm{MON}_Q[\cdot]$, (c) Eq. (8), (d) $\sup_g$, (e) the definition of $\hat{\mu}_{\mathbb{P},Q}$, (f) Eq. (8) and the linearity of $\mathrm{MON}_Q[\cdot]$, (g) the definition of $r_{Q,N}$. In step (h), by denoting $a = \|\tilde{g}\|_K$, $r = r_{Q,N}$, the argument of the sup takes the form $2ar - a^2$; $2ar - a^2 \leq r^2 \Leftrightarrow 0 \leq r^2 - 2ar + a^2 = (r-a)^2$.

In Eq. (9), we obtained an equation $a^2 - 2ra \leq r^2$ where $a := \|e\|_K \geq 0$. Hence $r^2 + 2ra - a^2 \geq 0$, $r_{1,2} = \left[-2a \pm \sqrt{4a^2 + 4a^2}\right]/2 = \left(-1 \pm \sqrt{2}\right)a$, thus by the non-negativity of a, $r \geq (-1+\sqrt{2})a$, i.e., $a \leq \frac{r}{\sqrt{2}-1} = (\sqrt{2}+1)r$. In other words, we arrived at

$$\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq \left(1+\sqrt{2}\right) r_{Q,N}. \qquad (10)$$

It remains to upper bound $r_{Q,N}$.

Step-2: Our goal is to provide a probabilistic bound on

$$r_{Q,N} = \sup_{f\in B_K} \mathrm{MON}_Q\left[\langle f, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K\right] \\ = \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]}\{\underbrace{\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K}_{=:r(f,q)}\}.$$

[9] $\mathrm{MON}_Q[c_1 + c_2 f] = c_1 + c_2\mathrm{MON}_Q[f]$ for any $c_1, c_2 \in \mathbb{R}$.

The N_c corrupted samples can affect (at most) N_c of the $(B_q)_{q\in[Q]}$ blocks. Let $U := [Q]\backslash C$ stand for the indices of the uncorrupted sets, where $C := \{q \in [Q] : \exists n_j \text{ s.t. } n_j \in S_q,\ j \in [N_c]\}$ contains the indices of the corrupted sets. If

$$\forall f \in B_K : \underbrace{|\{q \in U : r(f,q) \geq \epsilon\}|}_{\sum_{q\in U}\mathbb{I}_{r(f,q)\geq\epsilon}} + N_c \leq \frac{Q}{2}, \qquad (11)$$

then for $\forall f \in B_K$, $\operatorname{med}_{q\in[Q]}\{r(f,q)\} \leq \epsilon$, i.e. $\sup_{f\in B_K}\operatorname{med}_{q\in[Q]}\{r(f,q)\} \leq \epsilon$. Thus, our task boils down to controlling the event in (11) by appropriately choosing ϵ.

- **Controlling** $r(f,q)$: For any $f \in B_K$ the random variables $\langle f, k(\cdot,x_i) - \mu_{\mathbb{P}}\rangle_{\mathcal{H}_K} = f(x_i) - \mathbb{P}f$ are independent, have zero mean, and

$$\mathbb{E}_{x_i\sim\mathbb{P}}\langle f, k(\cdot,x_i) - \mu_{\mathbb{P}}\rangle_K^2 = \langle f, \Sigma_{\mathbb{P}} f\rangle_K \leq \\ \leq \|f\|_K \|\Sigma_{\mathbb{P}} f\|_K \leq \|f\|_K^2 \|\Sigma_{\mathbb{P}}\| = \|\Sigma_{\mathbb{P}}\| \quad (12)$$

  using the reproducing property of the kernel and the covariance operator, the Cauchy-Bunyakovsky-Schwarz (CBS) inequality and $\|f\|_{\mathcal{H}_K} = 1$.
  For a zero-mean random variable z by the Chebyshev's inequality $\mathbb{P}(z > a) \leq \mathbb{P}(|z| > a) \leq \mathbb{E}(z^2)/a^2$, which implies $\mathbb{P}\left(z > \sqrt{\mathbb{E}(z^2)/\alpha}\right) \leq \alpha$ by a $\alpha = \mathbb{E}(z^2)/a^2$ substitution. With $z := r(f,q)$ ($q \in U$), using $\mathbb{E}[z^2] = \mathbb{E}\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K^2 = \frac{Q}{N}\mathbb{E}_{x_i\sim\mathbb{P}}\langle f, k(\cdot,x_i) - \mu_{\mathbb{P}}\rangle_K^2$ and Eq. (12) one gets that for all $f \in B_K$, $\alpha \in (0,1)$ and $q \in U$:

$$\mathbb{P}\left(r(f,q) > \sqrt{\frac{\|\Sigma_{\mathbb{P}}\|\,Q}{\alpha N}}\right) \leq \alpha.$$

  This means $\mathbb{P}\left(r(f,q) > \frac{\epsilon}{2}\right) \leq \alpha$ with $\epsilon \geq 2\sqrt{\frac{\|\Sigma_{\mathbb{P}}\|Q}{\alpha N}}$.
- **Reduction to** ϕ: As a result

$$\sum_{q\in U}\mathbb{P}\left(r(f,q) \geq \frac{\epsilon}{2}\right) \leq |U|\alpha \Leftrightarrow \\ \sum_{q\in U}\mathbb{I}_{r(f,q)\geq\epsilon} \leq |U|\alpha + \\ + \sum_{q\in U}\Big[\mathbb{I}_{r(f,q)\geq\epsilon} - \underbrace{\mathbb{P}\left(r(f,q) \geq \frac{\epsilon}{2}\right)}_{\mathbb{E}\left[\mathbb{I}_{r(f,q)\geq\frac{\epsilon}{2}}\right]}\Big] =: A.$$

  Let us introduce the $\phi(t) = (t-1)\mathbb{I}_{1\leq t\leq 2} + \mathbb{I}_{t\geq 2}$ notation. ϕ is 1-Lipschitz and satisfies $\mathbb{I}_{2\leq t} \leq \phi(t) \leq \mathbb{I}_{1\leq t}$ for any $t \in \mathbb{R}$. Hence, we can upper bound A as

$$A \leq |U|\alpha + \sum_{q\in U}\left[\phi\left(\frac{2r(f,q)}{\epsilon}\right) - \mathbb{E}\phi\left(\frac{2r(f,q)}{\epsilon}\right)\right]$$

by noticing that $\epsilon \leq r(f,q) \Leftrightarrow 2 \leq 2r(f,q)/\epsilon$ and $\epsilon/2 \leq r(f,q) \Leftrightarrow 1 \leq 2r(f,q)/\epsilon$, and by using the $\mathbb{I}_{2\leq t} \leq \phi(t)$ and the $\phi(t) \leq \mathbb{I}_{1\leq t}$ bound, respectively. Taking supremum over B_K we arrive at

$$\begin{aligned}
&\sup_{f\in B_K} \sum_{q\in U} \mathbb{I}_{r(f,q)\geq\epsilon} \leq \\
&\leq |U|\alpha + \sup_{f\in B_K} \sum_{q\in U} \left[\phi\left(\frac{2r(f,q)}{\epsilon}\right) - \mathbb{E}\phi\left(\frac{2r(f,q)}{\epsilon}\right)\right] \\
&\leq |U|\alpha + \underbrace{\sup_{f\in B_K} \sum_{q\in U} \phi\left(\frac{2r(f,q)}{\epsilon}\right)}_{=:Z},
\end{aligned}$$

where at the end we exploited the non-negativity of ϕ.

- **Concentration of Z around its mean:** Notice that Z is a function of x_V, the samples in the uncorrupted blocks; $V = \cup_{q\in U} S_q$. By the bounded difference property of Z (Lemma 4) for any $\beta > 0$, the McDiarmid inequality (Lemma 6; we choose $\tau := Q\beta^2/2$ to get linear scaling in Q on the r.h.s.) implies that

$$\mathbb{P}\left(Z < \mathbb{E}_{x_V}[Z] + Q\beta\right) \geq 1 - e^{-\frac{Q\beta^2}{2}}.$$

- **Bounding $\mathbb{E}_{x_V}[Z]$:** Let $M = N/Q$ denote the number of elements in S_q-s. The $\mathcal{G} = \{g_f : f \in B_K\}$ class with $g_f : \mathcal{X}^M \to \mathbb{R}$ and $\mathbb{P}_M := \frac{1}{M}\sum_{m=1}^M \delta_{u_m}$ defined as

$$g_f(u_{1:M}) = \phi\left(\frac{\langle f, \mu_{\mathbb{P}_M} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right)$$

is uniformly bounded separable Carathéodory (Lemma 5), hence the symmetrization technique (Steinwart & Christmann, 2008, Prop. 7.10), (Ledoux & Talagrand, 1991) gives

$$\mathbb{E}_{x_V}[Z] \leq 2\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q\in U} e_q \phi\left(\frac{2r(f,q)}{\epsilon}\right)\right|,$$

where $\mathbf{e} = (e_q)_{q\in U} \in \mathbb{R}^{|U|}$ with i.i.d. Rademacher entries [$\mathbb{P}(e_q = \pm 1) = \frac{1}{2}$ $(\forall q)$].

- **Discarding ϕ:** Since $\phi(0) = 0$ and ϕ is 1-Lipschitz, by Talagrand's concentration inequality of Rademacher processes (Ledoux & Talagrand, 1991), (Koltchinskii, 2011, Theorem 2.3) one gets

$$\begin{aligned}
&\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q\in U} e_q \phi\left(\frac{2r(f,q)}{\epsilon}\right)\right| \leq \\
&\leq 2\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q\in U} e_q \frac{2r(f,q)}{\epsilon}\right|.
\end{aligned}$$

- **Switching from $|U|$ to N terms**: Applying an other symmetrization [(a)], the CBS inequality, $f \in B_K$, and the Jensen inequality

$$\begin{aligned}
&\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q=1}^{Q} e_q \frac{r(f,q)}{\epsilon}\right| \leq \\
&\overset{(a)}{\leq} \frac{2Q}{\epsilon N}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left[\sup_{f\in B_K} \Big|\underbrace{\sum_{n\in V} e'_n \langle f, K(\cdot,x_n) - \mu_{\mathbb{P}}\rangle_K}_{=\left\langle f, \sum_{n\in V} e'_n [K(\cdot,x_n)-\mu_{\mathbb{P}}]\right\rangle_K}\Big|\right] \\
&\leq \frac{2Q}{\epsilon N}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left[\sup_{f\in B_K} \underbrace{\|f\|_K}_{=1} \left\|\sum_{n\in V} e'_n [K(\cdot,x_n) - \mu_{\mathbb{P}}]\right\|_K\right] \\
&= \frac{2Q}{\epsilon N}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left\|\sum_{n\in V} e'_n [K(\cdot,x_n) - \mu_{\mathbb{P}}]\right\|_K \\
&\leq \frac{2Q}{\epsilon N}\sqrt{\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left\|\sum_{n\in V} e'_n [K(\cdot,x_n) - \mu_{\mathbb{P}}]\right\|_K^2} \\
&\overset{(b)}{=} \frac{2Q\sqrt{|V|\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon N}.
\end{aligned}$$

In (a), we proceeded as follows:

$$\begin{aligned}
&\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q\in U} e_q \frac{r(f,q)}{\epsilon}\right| = \\
&= \mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}} \sup_{f\in B_K} \left|\sum_{q\in U} e_q \frac{\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right| \\
&\overset{(c)}{\leq} \frac{2Q}{N\epsilon}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}}\mathbb{E}_{\mathbf{e}'} \sup_{f\in B_K} \left|\sum_{n\in V} e'_n e''_n \langle f, K(\cdot,x_n) - \mu_{\mathbb{P}}\rangle_K\right| \\
&= \frac{2Q}{N\epsilon}\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'} \sup_{f\in B_K} \left|\sum_{n\in V} e'_n \langle f, K(\cdot,x_n) - \mu_{\mathbb{P}}\rangle_K\right|,
\end{aligned}$$

where in (c) we applied symmetrization, $\mathbf{e}' = (e'_n)_{n\in V} \in \mathbb{R}^{|V|}$ with i.i.d. Rademacher entries, $e''_n = e_q$ if $n \in S_q$ $(q \in U)$, and we used that $(e'_n e''_n \langle f, K(\cdot,x_n) - \mu_{\mathbb{P}}\rangle_K)_{n\in V} \overset{\text{distr}}{=} (e'_n \langle f, K(\cdot,x_n) - \mu_{\mathbb{P}}\rangle_K)_{n\in V}$.
In step (b), we had

$$\begin{aligned}
&\mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'}\left\|\sum\nolimits_{n\in V} e'_n [K(\cdot,x_n) - \mu_{\mathbb{P}}]\right\|_K^2 = \\
&= \mathbb{E}_{x_V}\mathbb{E}_{\mathbf{e}'} \sum_{n\in V} [e'_n]^2 \langle K(\cdot,x_n) - \mu_{\mathbb{P}}, K(\cdot,x_n) - \mu_{\mathbb{P}}\rangle_K \\
&= |V|\mathbb{E}_{x\sim\mathbb{P}} \langle K(\cdot,x) - \mu_{\mathbb{P}}, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K \\
&= |V|\mathbb{E}_{x\sim\mathbb{P}} \operatorname{Tr}\left([K(\cdot,x) - \mu_{\mathbb{P}}] \otimes [K(\cdot,x) - \mu_{\mathbb{P}}]\right) \\
&= |V| \operatorname{Tr}(\Sigma_{\mathbb{P}})
\end{aligned}$$

exploiting the independence of e'_n-s and $[e'_n]^2 = 1$.

Until this point we showed that for all $\alpha \in (0,1)$, $\beta > 0$, if $\epsilon \geq 2\sqrt{\frac{\|\Sigma_{\mathbb{P}}\|Q}{\alpha N}}$ then

$$\sup_{f \in B_K} \sum_{q=1}^{Q} \mathbb{I}_{r(f,q) \geq \epsilon} \leq |U|\alpha + Q\beta + \frac{8Q\sqrt{|V|\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon N}$$

with probability at least $1 - e^{-\frac{Q\beta^2}{2}}$. Thus, to ensure that $\sup_{f \in B_K} \sum_{q=1}^{Q} \mathbb{I}_{r(f,q) \geq \epsilon} + N_c \leq Q/2$ it is sufficient to choose $(\alpha, \beta, \epsilon)$ such that

$$|U|\alpha + Q\beta + \frac{8Q\sqrt{|V|\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon N} + N_c \leq \frac{Q}{2},$$

and in this case $\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq (1+\sqrt{2})\epsilon$. Applying the $|U| \leq Q$ and $|V| \leq N$ bounds, we want to have

$$Q\alpha + Q\beta + \frac{8Q\sqrt{\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon\sqrt{N}} + N_c \leq \frac{Q}{2}. \tag{13}$$

Choosing $\alpha = \beta = \frac{\delta}{3}$ in Eq. (13), the sum of the first two terms is $Q\frac{2\delta}{3}$; $\epsilon \geq \max\left(2\sqrt{\frac{3\|\Sigma_{\mathbb{P}}\|Q}{\delta N}}, \frac{24}{\delta}\sqrt{\frac{\operatorname{Tr}(\Sigma_{\mathbb{P}})}{N}}\right)$ gives $\leq Q\frac{\delta}{3}$ for the third term. Theorem 1 follows since $N_c \leq Q(\frac{1}{2} - \delta)$.

4.2. Proof of Theorem 2

The reasoning is similar to Theorem 1; we detail the differences below. The high-level structure of the proof is as follows:

- First we prove that

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}_Q(\mathbb{P},\mathbb{Q})\right| \leq r_{Q,N},$$

  where

$$r_{Q,N} = \sup_{f \in B_K} \left| \operatorname*{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \right\rangle_K \right\} \right|.$$

- Then $r_{Q,N}$ is bounded.

Step-1:

- $\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}_Q(\mathbb{P},\mathbb{Q}) \leq r_{Q,N}$: By the subadditivity of supremum $[\sup_f(a_f+b_f) \leq \sup_f a_f + \sup_f b_f]$ one gets

$$\begin{aligned}
&\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) = \\
&= \sup_{f \in B_K} \operatorname*{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) \mp \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \right\rangle_K \right\} \\
&\leq \sup_{f \in B_K} \operatorname*{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \right\rangle_K \right\} \\
&\quad + \sup_{f \in B_K} \left\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \right\rangle_K \\
&\leq \underbrace{\sup_{f \in B_K} \left| \operatorname*{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \right\rangle_K \right\} \right|}_{= r_{Q,N}} \\
&\quad + \mathrm{MMD}_Q(\mathbb{P},\mathbb{Q}).
\end{aligned}$$

- $\mathrm{MMD}_Q(\mathbb{P},\mathbb{Q}) - \widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) \leq r_{Q,N}$: Let

$$\begin{aligned}
a_f &:= \left\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \right\rangle_K, \\
b_f &:= \operatorname*{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) - \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) \right\rangle_K \right\}.
\end{aligned}$$

  Then

$$\begin{aligned}
a_f - b_f &= \left\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \right\rangle_K + \\
&\quad + \operatorname{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \right\rangle_K \right\} \\
&= \operatorname{med}_{q \in [Q]} \left\{ \left\langle f, \mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}} \right\rangle_K \right\}
\end{aligned}$$

  by $\operatorname{med}_{q \in [Q]}\{-z_q\} = -\operatorname{med}_{q \in [Q]}\{z_q\}$. Applying the $\sup_f(a_f - b_f) \geq \sup_f a_f - \sup_f b_f$ inequality (it follows from the subadditivity of sup):

$$\begin{aligned}
&\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) \geq \mathrm{MMD}_Q(\mathbb{P},\mathbb{Q}) - \\
&\quad - \sup_{f \in B_K} \underbrace{\operatorname*{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) - \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) \right\rangle_K \right\}}_{- \operatorname{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \right\rangle_K \right\}} \\
&\geq \mathrm{MMD}_Q(\mathbb{P},\mathbb{Q}) - \\
&\quad - \underbrace{\sup_{f \in B_K} \left| \operatorname*{med}_{q \in [Q]} \left\{ \left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \right\rangle_K \right\} \right|}_{r_{Q,N}}.
\end{aligned}$$

Step-2: Our goal is to control

$$\begin{aligned}
r_{Q,N} &= \sup_{f \in B_K} \left| \operatorname{med}_{q \in [Q]} \left\{ r(f,q) \right\} \right|, \text{ where} \\
r(f,q) &:= \left\langle f, \left(\mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}}\right) - \left(\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\right) \right\rangle_K.
\end{aligned}$$

The relevant quantities which change compared to the proof of Theorem 1 are as follows.

- **Median rephrasing**:

$$\begin{aligned}
&\sup_{f \in B_K} \left| \operatorname*{med}_{q \in [Q]} \{r(f,q)\} \right| \leq \epsilon && \Leftrightarrow \\
&\forall f \in B_K : -\epsilon \leq \operatorname{med}_{q \in [Q]} \{r(f,q)\} \leq \epsilon && \Leftarrow \\
&\forall f \in B_K : |\{q : r(f,q) \leq -\epsilon\}| \leq Q/2, \text{ and} \\
&\qquad\qquad |\{q : r(f,q) \geq \epsilon\}| \leq Q/2 && \Leftarrow \\
&\forall f \in B_K : |\{q : |r(f,q)| \geq \epsilon\}| \leq Q/2.
\end{aligned}$$

  Thus,

$$\forall f \in B_K : |\{q \in U : |r(f,q)| \geq \epsilon\}| + N_c \leq \frac{Q}{2},$$

  implies $\sup_{f \in B_K} \left| \operatorname{med}_{q \in [Q]} \{r(f,q)\} \right| \leq \epsilon$.

- **Controlling** $|r(f,q)|$: For any $f \in B_K$ the random variables $[f(x_i) - f(y_i)] - [\mathbb{P}f - \mathbb{Q}f]$ are independent, zero-mean and

$$\begin{aligned}
&\mathbb{E}_{(x,y) \sim \mathbb{P} \otimes \mathbb{Q}} ([f(x) - \mathbb{P}f] - [f(y) - \mathbb{Q}f])^2 = \\
&\quad = \mathbb{E}_{x \sim \mathbb{P}} [f(x) - \mathbb{P}f]^2 + \mathbb{E}_{y \sim \mathbb{Q}} [f(y) - \mathbb{Q}f]^2 \\
&\quad \leq \|\Sigma_{\mathbb{P}}\| + \|\Sigma_{\mathbb{Q}}\|,
\end{aligned}$$

where $\mathbb{P} \otimes \mathbb{Q}$ is the product measure. The Chebyshev argument with $z = |r(f,q)|$ implies that $\forall \alpha \in (0,1)$

$$(\mathbb{P} \otimes \mathbb{Q}) \left(|r(f,q)| > \sqrt{\frac{(\|\Sigma_{\mathbb{P}}\| + \|\Sigma_{\mathbb{Q}}\|)\, Q}{\alpha N}} \right) \leq \alpha.$$

This means $(\mathbb{P} \otimes \mathbb{Q})\left(|r(f,q)| > \epsilon/2\right) \leq \alpha$ with $\epsilon \geq 2\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)Q}{\alpha N}}$.

- **Switching from $|U|$ to N terms**: With $(xy)_V = \{(x_i, y_i) : i \in V\}$, in '(b)' with $\tilde{x}_n := K(\cdot, x_n) - \mu_{\mathbb{P}}$, $\tilde{y}_n := K(\cdot, y_n) - \mu_{\mathbb{Q}}$ we arrive at

$$\begin{aligned}
&\mathbb{E}_{(xy)_V} \mathbb{E}_{\mathbf{e}'} \left\| \sum_{n \in V} e'_n \left(\tilde{x}_n - \tilde{y}_n \right) \right\|_K^2 = \\
&= \mathbb{E}_{(xy)_V} \mathbb{E}_{\mathbf{e}'} \sum_{n \in V} \left[e'_n \right]^2 \langle \tilde{x}_n - \tilde{y}_n, \tilde{x}_n - \tilde{y}_n \rangle_K \\
&= |V| \mathbb{E}_{(xy) \sim \mathbb{P}} \left\| [K(\cdot, x) - \mu_{\mathbb{P}}] - [K(\cdot, y) - \mu_{\mathbb{Q}}] \right\|_K \\
&= |V| \mathbb{E}_{(xy) \sim \mathbb{P}} \operatorname{Tr} \left([K(\cdot, x) - \mu_{\mathbb{P}}] - [K(\cdot, y) - \mu_{\mathbb{Q}}] \otimes \right. \\
&\qquad \left. \otimes [K(\cdot, x) - \mu_{\mathbb{P}}] - [K(\cdot, y) - \mu_{\mathbb{Q}}] \right) \\
&= |V| \left[\operatorname{Tr}\left(\Sigma_{\mathbb{P}}\right) + \operatorname{Tr}\left(\Sigma_{\mathbb{Q}}\right) \right].
\end{aligned}$$

- These results imply

$$Q\alpha + Q\beta + \frac{8Q\sqrt{\operatorname{Tr}\left(\Sigma_{\mathbb{P}}\right) + \operatorname{Tr}\left(\Sigma_{\mathbb{Q}}\right)}}{\epsilon\sqrt{N}} + N_c \leq Q/2.$$

$$\epsilon \geq \max\left(2\sqrt{\frac{3(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)Q}{\delta N}}, \frac{24}{\delta}\sqrt{\frac{\operatorname{Tr}\left(\Sigma_{\mathbb{P}}\right)+\operatorname{Tr}\left(\Sigma_{\mathbb{Q}}\right)}{N}} \right),$$

$\alpha = \beta = \frac{\delta}{3}$ choice finishes the proof of Theorem 2.

References

Alam, Md. Ashad, Fukumizu, Kenji, and Wang, Yu-Ping. Influence function and robust variant of kernel canonical correlation analysis. Technical report, 2017. (https://arxiv.org/abs/1705.04194).

Alon, Noga, Matias, Yossi, and Szegedy, Mario. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1, part 2): 137–147, 1999.

Aronszajn, Nachman. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Audibert, Jean-Yves and Catoni, Olivier. Robust linear least squares regression. *The Annals of Statistics*, 39(5): 2766–2794, 2011.

Balasubramanian, Krishnakumar, Li, Tong, and Yuan, Ming. On the optimality of kernel-embedding based goodness-of-fit tests. Technical report, 2017. (https://arxiv.org/abs/1709.08148).

Baringhaus, Ludwig and Franz, C. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88: 190–206, 2004.

Berlinet, Alain and Thomas-Agnan, Christine. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Blanchard, Gilles, Deshmukh, Aniket Anand, Dogan, Urun, Lee, Gyemin, and Scott, Clayton. Domain generalization by marginal transfer learning. Technical report, 2017. (https://arxiv.org/abs/1711.07910).

Catoni, Olivier. Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 48(4):1148–1185, 2012.

Catoni, Olivier and Giulini, Ilaria. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. Technical report, 2017. (https://arxiv.org/abs/1712.02747).

Collins, Michael and Duffy, Nigel. Convolution kernels for natural language. In *NIPS*, pp. 625–632, 2001.

Cuturi, Marco. Fast global alignment kernels. In *ICML*, pp. 929–936, 2011.

Cuturi, Marco, Fukumizu, Kenji, and Vert, Jean-Philippe. Semigroup kernels on measures. *Journal of Machine Learning Research*, 6:1169–1198, 2005.

Devroye, Luc, Lerasle, Matthieu, Lugosi, Gábor, and Oliveira, Roberto I. Sub-Gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

Fukumizu, Kenji, Gretton, Arthur, Sun, Xiaohai, and Schölkopf, Bernhard. Kernel measures of conditional dependence. In *NIPS*, pp. 498–496, 2008.

Fukumizu, Kenji, Song, Le, and Gretton, Arthur. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14:3753–3783, 2013.

Gärtner, Thomas, Flach, Peter A., Kowalczyk, Adam, and Smola, Alexander. Multi-instance kernels. In *ICML*, pp. 179–186, 2002.

Gretton, Arthur, Fukumizu, Kenji, Teo, Choon Hui, Song, Le, Schölkopf, Bernhard, and Smola, Alexander J. A kernel statistical test of independence. In *NIPS*, pp. 585–592, 2008.

Gretton, Arthur, Borgwardt, Karsten M., Rasch, Malte J., Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Guevara, Jorge, Hirata, Roberto, and Canu, Stéphane. Cross product kernels for fuzzy set similarity. In *FUZZ-IEEE*, pp. 1–6, 2017.

Györfi, László, Kohler, Michael, Krzyzak, Adam, and Walk, Harro. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New-york, 2002.

Harchaoui, Zaid, Bach, Francis, and Moulines, Eric. Testing for homogeneity with kernel Fisher discriminant analysis. In *NIPS*, pp. 609–616, 2007.

Haussler, David. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999. (http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Hein, Matthias and Bousquet, Olivier. Hilbertian metrics and positive definite kernels on probability measures. In *AISTATS*, pp. 136–143, 2005.

Ilya Tolstikhin, Bharath K. Sriperumbudur and Schölkopf, Bernhard. Minimax estimation of maximal mean discrepancy with radial kernels. In *NIPS*, pp. 1930–1938, 2016.

Jebara, Tony, Kondor, Risi, and Howard, Andrew. Probability product kernels. *Journal of Machine Learning Research*, 5:819–844, 2004.

Jerrum, Mark R., G.Valiant, Leslie, and V.Vazirani, Vijay. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43 (2-3):169–188, 1986.

Jiao, Yunlong and Vert, Jean-Philippe. The Kendall and Mallows kernels for permutations. In *ICML (PMLR)*, volume 37, pp. 2982–2990, 2016.

Jitkrittum, Wittawat, Xu, Wenkai, Szabó, Zoltán, Fukumizu, Kenji, and Gretton, Arthur. A linear-time kernel goodness-of-fit test. In *NIPS*, pp. 261–270, 2017.

Kashima, Hisashi and Koyanagi, Teruo. Kernels for semi-structured data. In *ICML*, pp. 291–298, 2002.

Kim, Been, Khanna, Rajiv, and Koyejo, Oluwasanmi O. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*, pp. 2280–2288, 2016.

Kim, JooSeuk and Scott, Clayton D. Robust kernel density estimation. *Journal of Machine Learning Research*, 13: 2529–2565, 2012.

Klebanov, Lev. *N-Distances and Their Applications*. Charles University, Prague, 2005.

Koltchinskii, Vladimir. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

Kondor, Risi and Pan, Horace. The multiscale Laplacian graph kernel. In *NIPS*, pp. 2982–2990, 2016.

Kusano, Genki, Fukumizu, Kenji, and Hiraoka, Yasuaki. Persistence weighted Gaussian kernel for topological data analysis. In *ICML*, pp. 2004–2013, 2016.

Law, Ho Chung Leon, Sutherland, Dougal J., Sejdinovic, Dino, and Flaxman, Seth. Bayesian approaches to distribution regression. In *AISTATS*, 2018.

Le Cam, Lucien. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53, 1973.

Ledoux, Michel and Talagrand, Michel. *Probability in Banach spaces*. Springer-Verlag, 1991.

Lloyd, James Robert, Duvenaud, David, Grosse, Roger, Tenenbaum, Joshua B., and Ghahramani, Zoubin. Automatic construction and natural-language description of nonparametric regression models. In *AAAI Conference on Artificial Intelligence*, pp. 1242–1250, 2014.

Lodhi, Huma, Saunders, Craig, Shawe-Taylor, John, Cristianini, Nello, and Watkins, Chris. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

Lugosi, Gábor and Mendelson, Shahar. Risk minimization by median-of-means tournaments. Technical report, 2016. (https://arxiv.org/abs/1608.00757).

Lugosi, Gábor and Mendelson, Shahar. Sub-gaussian estimators of the mean of a random vector. Technical report, 2017. (https://arxiv.org/abs/1702.00482).

Martins, André F. T., Smith, Noah A., Xing, Eric P., Aguiar, Pedro M. Q., and Figueiredo, Mário A. T. Nonextensive information theoretic kernels on measures. *The Journal of Machine Learning Research*, 10:935–975, 2009.

Mendelson, Shahar. Learning without concentration. *Journal of the ACM*, 62(3):21:1–21:25, 2015.

Minsker, Stanislav. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

Minsker, Stanislav and Strawn, Nate. Distributed statistical estimation and rates of convergence in normal approximation. Technical report, 2017. (https://arxiv.org/abs/1704.02658).

Mooij, Joris M., Peters, Jonas, Janzing, Dominik, Zscheischler, Jakob, and Schölkopf, Bernhard. Distinguishing cause from effect using observational data: Methods and benchmarks. *Journal of Machine Learning Research*, 17: 1–102, 2016.

Muandet, Krikamol, Fukumizu, Kenji, Dinuzzo, Francesco, and Schölkopf, Bernhard. Learning from distributions via support measure machines. In *NIPS*, pp. 10–18, 2011.

Muandet, Krikamol, Sriperumbudur, Bharath K., Fukumizu, Kenji, Gretton, Arthur, and Schölkopf, Bernhard. Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17:1–41, 2016.

Muandet, Krikamol, Fukumizu, Kenji, Sriperumbudur, Bharath, and Schölkopf, Bernhard. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

Müller, Alfred. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

Nemirovski, Arkadi S. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. John Wiley & Sons Ltd., 1983.

Park, Mijung, Jitkrittum, Wittawat, and Sejdinovic, Dino. K2-ABC: Approximate Bayesian computation with kernel embeddings. In *AISTATS (PMLR)*, volume 51, pp. 51:398–407, 2016.

Pfister, Niklas, Bühlmann, Peter, Schölkopf, Bernhard, and Peters, Jonas. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2017. ISSN 1467-9868.

Schölkopf, Bernhard and Smola, Alexander J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

Schölkopf, Bernhard, Muandet, Krikamol, Fukumizu, Kenji, Harmeling, Stefan, and Peters, Jonas. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25(4): 755–766, 2015.

Sejdinovic, Dino, Sriperumbudur, Bharath K., Gretton, Arthur, and Fukumizu, Kenji. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013.

Shawe-Taylor, John and Cristianini, Nello. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Sinova, Beatriz, Gonzlez-Rodríguez, Gil, and Aelst, Stefan Van. M-estimators of location for functional data. *Bernoulli*, 24:2328–2357, 2018.

Smola, Alexander, Gretton, Arthur, Song, Le, and Schöolkopf, Bernhard. A Hilbert space embedding for distributions. In *ALT*, pp. 13–31, 2007.

Song, Le, Gretton, Arthur, Bickson, Danny, Low, Yucheng, and Guestrin, Carlos. Kernel belief propagation. In *AISTATS*, pp. 707–715, 2011.

Sriperumbudur, Bharath K., Gretton, Arthur, Fukumizu, Kenji, Schölkopf, Bernhard, and Lanckriet, Gert R.G. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11: 1517–1561, 2010.

Steinwart, Ingo and Christmann, Andreas. *Support Vector Machines*. Springer, 2008.

Szabó, Zoltán, Sriperumbudur, Bharath, Póczos, Barnabás, and Gretton, Arthur. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17 (152):1–40, 2016.

Székely, Gábor J. and Rizzo, Maria L. Testing for equal distributions in high dimension. *InterStat*, 5, 2004.

Székely, Gábor J. and Rizzo, Maria L. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80, 2005.

Tolstikhin, Ilya, Sriperumbudur, Bharath K., and Muandet, Krikamol. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18:1–47, 2017.

Vishwanathan, S.V. N., Schraudolph, Nicol N., Kondor, Risi, and Borgwardt, Karsten M. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

Yamada, Makoto, Umezu, Yuta, Fukumizu, Kenji, and Takeuchi, Ichiro. Post selection inference with kernels. Technical report, 2016. (https://arxiv.org/abs/1610.03725).

Zaheer, Manzil, Kottur, Satwik, Ravanbakhsh, Siamak, Póczos, Barnabás, Salakhutdinov, Ruslan R., and Smola, Alexander J. Deep sets. In *NIPS*, pp. 3394–3404, 2017.

Zhang, Kun, Schölkopf, Bernhard, Muandet, Krikamol, and Wang, Zhikun. Domain adaptation under target and conditional shift. *Journal of Machine Learning Research*, 28 (3):819–827, 2013.

Zinger, A. A., Kakosyan, A. V., and Klebanov, L. B. A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Soviet Mathematics*, 1992.

Zolotarev, V. M. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.

Supplement

The supplement contains a few technical lemmas used during the proofs of our results (Section A), and the McDiarmid inequality for self-containedness (Section B).

A. Technical Lemmas

Lemma 3 (Supremum)**.**

$$\left|\sup_f a_f - \sup_f b_f\right| \leq \sup_f |a_f - b_f|.$$

Proof. By using the subadditive property of sup:

$$\begin{aligned}\sup_f b_f &= \sup_f (b_f - a_f + a_f) \leq \sup_f (b_f - a_f) + \sup_f a_f \\ &\leq \sup_f |a_f - b_f| + \sup_f a_f, \\ \sup_f a_f &= \sup_f (a_f - b_f + b_f) \leq \sup_f (a_f - b_f) + \sup_f b_f \\ &\leq \sup_f |a_f - b_f| + \sup_f b_f.\end{aligned}$$

These two inequalities imply $\pm\left[\sup_f a_f - \sup_f b_f\right] \leq \sup_f |a_f - b_f|$, from which the statement follows. □

Lemma 4 (Bounded difference property of Z)**.** *Let $N \in \mathbb{Z}^+$, $(S_q)_{q\in[Q]}$ be a partition of $[N]$, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel, μ be the mean embedding associated to K, $x_{1:N}$ be i.i.d. random variables on $\mathcal{X}$, $Z(x_V) = \sup_{f\in B_K} \sum_{q\in U} \phi\left(\frac{2\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\rangle_K}{\epsilon}\right)$, where $U \subseteq [Q]$, $V = \cup_{q\in U} S_q$. Let x'_{V_i} be x_V except for the $i \in V$-th coordinate; x_i is changed to x'_i. Then*

$$\sup_{x_V \in \mathcal{X}^{|V|}, x'_i \in \mathcal{X}} \left|Z(x_V) - Z\left(x'_{V_i}\right)\right| \leq 2, \forall i \in V.$$

Proof. Since $(S_q)_{q\in[Q]}$ is a partition of $[Q]$, $(S_q)_{q\in U}$ forms a partition of V and there exists a unique $r \in U$ such that $i \in S_r$. Let

$$\begin{aligned}Y_q &:= Y_q(f, x_V) = \phi\left(\frac{2\left\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\right\rangle_K}{\epsilon}\right), \\ Y'_r &:= Y_r(f, x'_{V_i}),\ q \in U.\end{aligned}$$

In this case

$$\begin{aligned}&\left|Z(x_V) - Z\left(x'_{V_i}\right)\right| = \\ &= \left|\sup_{f\in B_K} \sum_{q\in U} Y_q - \sup_{f\in B_K} \left(\sum_{q\in U\setminus\{r\}} Y_q + Y'_r\right)\right| \\ &\overset{(a)}{\leq} \sup_{f\in B_K} |Y_r - Y'_r| \overset{(b)}{\leq} \sup_{f\in B_K} \Big(\underbrace{|Y_r|}_{\leq 1} + \underbrace{|Y'_r|}_{\leq 1}\Big) \leq 2,\end{aligned}$$

where in (a) we used Lemma 3, (b) the triangle inequality and the boundedness of ϕ [$|\phi(t)| \leq 1$ for all t].

□

Lemma 5 (Uniformly bounded separable Carathéodory family)**.** *Let $\epsilon > 0$, $N \in \mathbb{Z}^+$, $Q \in \mathbb{Z}^+$, $M = N/Q \in \mathbb{Z}^+$, $\phi(t) = (t-1)\mathbb{I}_{1\leq t\leq 2} + \mathbb{I}_{t\geq 2}$, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous kernel on the separable topological domain $\mathcal{X}$, μ is the mean embedding associated to K, $\mathbb{P}_M := \frac{1}{M}\sum_{m=1}^M \delta_{u_m}$, $\mathcal{G} = \{g_f : f \in B_K\}$, where $g_f : \mathcal{X}^M \to \mathbb{R}$ is defined as*

$$g_f(u_{1:M}) = \phi\left(\frac{2\left\langle f, \mu_{\mathbb{P}_M} - \mu_{\mathbb{P}}\right\rangle_K}{\epsilon}\right).$$

Then $\mathcal{G}$ is a uniformly bounded separable Carathéodory family:

1. *$\sup_{f\in B_K} \|g_f\|_\infty < \infty$ where $\|g\|_\infty = \sup_{u_{1:M}\in\mathcal{X}^M} |g(u_{1:M})|$.*
2. *$u_{1:M} \mapsto g_f(u_{1:M})$ is measurable for all $f \in B_K$.*
3. *$f \mapsto g_f(u_{1:M})$ is continuous for all $u_{1:M} \in \mathcal{X}^M$.*
4. *B_K is separable.*

Proof.

1. $|\phi(t)| \leq 1$ for any t, hence $\|g_f\|_\infty \leq 1$ for all $f \in B_K$.
2. Any $f \in B_K$ is continuous since $\mathcal{H}_K \subset C(\mathcal{X}) = \{h : \mathcal{X} \to \mathbb{R}$ continuous$\}$, so $u_{1:M} \mapsto (f(u_1), \ldots, f(u_M))$ is continuous. ϕ is Lipschitz, specifically continuous. The continuity of these two maps imply that of $u_{1:M} \mapsto g_f(u_{1:M})$, specifically it is Borel-measurable.
3. The statement follows by the continuity of $f \mapsto \langle f, h\rangle_K$ ($h = \mu_{\mathbb{P}_M} - \mu_{\mathbb{P}}$) and that of ϕ.
4. B_K is separable since $\mathcal{H}_K$ is so by assumption.

□

B. External Lemma

Below we state the McDiarmid inequality for self-containedness.

Lemma 6 (McDiarmid inequality)**.** *Let $x_{1:N}$ be $\mathcal{X}$-valued independent random variables. Assume that $f : \mathcal{X}^N \to \mathbb{R}$ satisfies the bounded difference property:*

$$\sup_{u_1,\ldots,u_N,u'_r\in\mathcal{X}} |f(u_{1:N}) - f(u'_{1:N})| \leq c, \quad \forall n \in [N],$$

where $u'_{1:N} = (u_1, \ldots, u_{n-1}, u'_n, u_{n+1}, \ldots, u_N)$. Then for any $\tau > 0$

$$\mathbb{P}\left(f(x_{1:N}) < \mathbb{E}_{x_{1:N}}[f(x_{1:N})] + c\sqrt{\frac{\tau N}{2}}\right) \geq 1 - e^{-\tau}.$$