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

Matthieu Lerasle, Zoltán Szabó, Timothée Mathieu, Guillaume Lecué

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

Matthieu Lerasle[†] **Zoltán Szabó**[‡] **Timothée Mathieu**[*] **Guillaume Lecué**[**]

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 estimators for kernel mean embedding and MMD with excessive resistance properties to outliers, and optimal sub-Gaussian deviation bounds under mild assumptions.

1 INTRODUCTION

Kernel methods [3] form the backbone of a tremendous number of successful applications in machine learning thanks to their power in capturing complex relations [61, 67]. 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.

[†]Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud; CNRS, Université Paris Saclay, France; [‡]CMAP, École Polytechnique, Palaiseau, France; [*]Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, France; [**]CREST ENSAE ParisTech, France. Correspondence to: Matthieu Lerasle <matthieu.lerasle@math.u-psud.fr>, Zoltán Szabó <zoltan.szabo@polytechnique.edu>.

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 [11, 30], time series [12], strings [44], mixture models, hidden Markov models or linear dynamical systems [26], sets [24, 18], fuzzy domains [21], distributions [25, 47, 52], groups [13] such as specific constructions on permutations [28], or graphs [74, 36].

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 [7, 64] induces a semi-metric[1] on distributions called maximum mean discrepancy (MMD) [64, 19]

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

Specializing the kernel, classical integral transforms widely used in probability theory and statistics can be recovered by $\mu_{\mathbb{P}}$; for example, if $\mathcal{X}$ equipped with the scalar product $\langle \cdot, \cdot \rangle$ is a Hilbert space, the kernel $K(x, y) = \mathrm{e}^{\langle x, y \rangle}$ gives the moment-generating function, $K(x, y) = \mathrm{e}^{\gamma \|x-y\|_2^2}$ $(\gamma > 0)$ the Weierstrass transform. As it has been shown [62] energy distance [6, 69, 70]—also known as N-distance [78, 33] in the statistical literature—coincides with MMD.

Mean embedding and maximum mean discrepancy have been applied successfully, in kernel Bayesian inference [65, 17], approximate Bayesian computation [57], model criticism [43, 31], two-sample [6, 69, 70, 23, 19], independence [20, 58] and goodness-of-fit testing [29, 5], domain adaptation [77] and generalization [8], probabilistic programming [60], post selection inference [75], distribution classification [52, 76] and regression [68, 38], causal discovery [51, 58] or topological data analysis [37], among many others; [53] provide an in-depth review on the topic.

[1] [16, 66] provide conditions when MMD is a metric, i.e. μ is injective.

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 [19]. Kernel mean shrinkage estimators [54] 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 $N^{-1/2}$ [71] 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 $N_1^{-1/2} + N_2^{-1/2}$ [72] in case of radial universal kernels defined on $\mathcal{X} = \mathbb{R}^d$.

A critical property of an estimator is its resistance 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) [32]: 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 [1], and (ii) relaxed to general Hilbert spaces [63]. Unfortunately, the consistency of the investigated empirical M-estimators is unknown, except for finite-dimensional feature maps [63].

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 [27, 2, 56] for the estimation of the mean of real-valued random variables. Their concentration properties have been recently studied by [14, 50] following the approach of [9] for M-estimators. These studies focusing on the estimation of the mean of real-valued random variables are important as they can be used to tackle more general prediction problems in learning theory via the classical empirical risk minimization approach [73] or by more sophisticated approach such as the minmax procedure [4].

In parallel to the minmax approach, there have been several attempts to extend the usage of MON estimators from $\mathbb{R}$ to more general settings. For example, [49, 50] consider the problem of estimating the mean of a Banach-valued random variable using "geometrical" MONs. The estimators in [49, 50] are computationally tractable but the deviation bounds are suboptimal compared to those one can prove for the empirical mean under sub-Gaussian assumptions. In regression problems, [45, 40] proposed to combine the classical MON estimators on $\mathbb{R}$ in a "test" procedure that can be seen as a Le Cam test estimator [39]. The achievement in [45, 40] is that they were able to obtain optimal deviation bounds for the resulting estimator using the powerful so-called small-ball method of [35, 48]. This approach was then extended to mean estimation $\mathbb{R}^d$ in [46] providing the first rate-optimal sub-Gaussian deviation bounds under minimal L^2-assumptions. The constants of [45, 40, 46] have been improved in [10] for the estimation of the mean in $\mathbb{R}^d$ under L^4-moment assumption and in least-squares regression under L^4/L^2-condition that is stronger than the small-ball assumption used in [45, 40]. Unfortunately, these estimators are computationally intractable; there risk bounds however serve as an important baseline for computable estimators such as the minmax MON estimators in regression [41].

From statistical point of view, our goal is to prove optimal sub-Gaussian deviation bounds for MON-based mean estimators in RKHS-s which hold under *minimal* stochastic assumptions, requiring the finiteness of second-order moments only (see later the trace-class assumption on the covariance operators) and with the potential presence of corrupted (and even adversarial) data. We extend the results of [46, 41] to mean embedding and MMD estimation in RKHS. In order to attain this goal, we modify the aggregation step of [45, 46, 40] using the minmax formulation of [4, 41]. As a result we get conceptually simpler estimators with computationally tractable algorithms.

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

1. We design novel mean embedding and MMD estimators based on the MON principle.
2. We establish their finite-sample and outlier-robustness properties. The obtained MONK es-

[2]Smoothing kernels extensively studied in the non-parametric statistical literature [22] are assumed to be non-negative functions integrating to one.

timators (i) obey optimal sub-Gaussian deviation bounds under mild trace-class conditions, (i) their convergence speed match the discussed minimax rates, and (iii) thanks to the usage of medians they are also robust to contamination.

Section 2 contains definitions and problem formulation. Our main results are given in Section 3. Implementation of the MONK estimators is the focus of Section 4, with numerical illustrations in Section 5.

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 be represented by inner products $f(x) = \langle f, K(\cdot,x)\rangle_K$ for all $x \in \mathcal{X}$, $f \in \mathcal{H}_K$. 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 $\mathbb{P}f := \mathbb{E}_{x\sim\mathbb{P}} f(x) = \langle f, \mu_{\mathbb{P}}\rangle_K$ for all $f \in \mathcal{H}_K$, and it is the unique solution of the problem:

$$\mu_{\mathbb{P}} = \mathrm{argmin}_{f\in\mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot,x)\|_K^2 \,\mathrm{d}\mathbb{P}(x)\,. \tag{3}$$

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

$$\mu_{\mathbb{P}} = \mathrm{argmin}_{f\in\mathcal{H}_K} \sup_{g\in\mathcal{H}_K} J(f,g), \tag{4}$$

where

$$J(f,g) = \mathbb{E}_{x\sim\mathbb{P}}\left[\|f - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\right].$$

The equivalence of (3) and (4) 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 $h : \mathcal{X} \to \mathbb{R}$ and any non-empty subset $S \subseteq [N]$, denote by $\mathbb{P}_S := |S|^{-1}\sum_{i\in S}\delta_{x_i}$ the empirical measure associated to the subset x_S and $\mathbb{P}_S h = |S|^{-1}\sum_{i\in S} h(x_i)$; 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

$$\begin{aligned}\mathrm{MON}_Q[h] &= \mathrm{med}_{q\in[Q]}\left\{\mathbb{P}_{S_q} h\right\} \\ &= \mathrm{med}_{q\in[Q]}\left\{\langle h, \mu_{S_q}\rangle_K\right\},\end{aligned}$$

where assuming that $h \in \mathcal{H}_K$ the second equality is a consequence of the mean-reproducing property of $\mu_{\mathbb{P}}$. Specifically, in case of $Q = 1$ the MON operation reduces to the classical mean: $\mathrm{MON}_1[h] = N^{-1}\sum_{n=1}^N h(x_n)$.

We define the minmax MON-based estimator associated to kernel K (MONK) as

$$\hat{\mu}_{\mathbb{P},Q} = \hat{\mu}_{\mathbb{P},Q}(x_{1:N}) \in \mathrm{argmin}_{f\in\mathcal{H}_K} \sup_{g\in\mathcal{H}_K} \tilde{J}(f,g), \tag{5}$$

where for all $f, g \in \mathcal{H}_K$

$$\begin{aligned}&\tilde{J}(f,g) = \\ &\quad = \mathrm{MON}_Q\left[x \mapsto \|f - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\right].\end{aligned}$$

When $Q = 1$, since $\mathrm{MON}_1[h]$ is the empirical mean, we obtain the classical empirical mean based estimator: $\hat{\mu}_{\mathbb{P},1} = \frac{1}{N}\sum_{n=1}^N K(\cdot,x_n)$.

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 second equality shows that MMD is a specific integral probability metric [55, 79]. Assume that we have access to $x_{1:N} \sim \mathbb{P}$, $y_{1:N} \sim \mathbb{Q}$ samples, where we assumed the size of the two samples to be the same for simplicity. 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), $\mu_{S_q,\mathbb{P}} := \mu_{\mathbb{P}_{S_q,x}}$, $\mu_{S_q,\mathbb{Q}} := \mu_{\mathbb{P}_{S_q,y}}$. We propose the

[3] $\mathcal{H}_K$ is separable by the separability of $\mathcal{X}$ and the continuity of K [67, Lemma 4.33].

following MON-based MMD estimator

$$\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \operatorname*{med}_{q\in[Q]} \left\{ \left\langle f, \mu_{S_q,\mathbb{P}} - \mu_{S_q,\mathbb{Q}} \right\rangle_K \right\}. \tag{6}$$

Again, with the $Q=1$ choice, the classical V-statistic based MMD estimator [19] is recovered:

$$\widehat{\mathrm{MMD}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in B_K} \left[\frac{1}{N}\sum_{n\in[N]} f(x_n) - \frac{1}{N}\sum_{n\in[N]} f(y_n) \right]$$
$$= \sqrt{\frac{1}{N^2}\sum_{i,j\in[N]} \left(K^x_{ij} + K^y_{ij} - 2K^{xy}_{ij}\right)}, \tag{7}$$

where $K^x_{ij} = K(x_i,x_j), K^y_{ij} = K(y_i,y_j)$ and $K^{xy}_{ij} = K(x_i,y_j)$ for all $i,j\in[N]$. Changing in Eq. (7) $\sum_{i,j\in[N]}$ to $\sum_{i,j\in[N],i\neq j}$ in case of the K^x_{ij} and K^y_{ij} terms gives the (unbiased) U-statistic based MMD estimator

$$\frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\neq j}} \left(K^x_{ij} + K^y_{ij}\right) - \frac{2}{N^2}\sum_{i,j\in[N]} K^{xy}_{ij}. \tag{8}$$

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

A **few additional 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 $\mathcal{H}_K$. $A\in\mathcal{L}(\mathcal{H}_K)$ is called 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 $\mathrm{Tr}(A) := \sum_{i\in I}\langle Ae_i,e_i\rangle_K < \infty$. If A is non-negative and self-adjoint, then A is trace class iff $\mathrm{Tr}(A)<\infty$; this will hold for the covariance operator ($\Sigma_{\mathbb{P}}$, see Eq. (9)). $A\in\mathcal{L}(\mathcal{H}_K)$ is called Hilbert-Schmidt if $\|A\|_2^2 := \mathrm{Tr}(A^*A) = \sum_{i\in I}\langle Ae_i,Ae_i\rangle_K < \infty$. One can show that the definitions of trace-class and Hilbert-Schmidt operators are independent of the particular choice of the ONB $(e_i)_{i\in I}$. 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. The tensor product of $a,b\in\mathcal{H}_K$ is $(a\otimes b)(c) = a\langle b,c\rangle_K$, $(\forall c\in\mathcal{H}_K)$. $a\otimes b\in\mathcal{L}_2(\mathcal{H}_K)$, $\mathcal{L}_2(\mathcal{H}_K)\cong\mathcal{H}_K\otimes\mathcal{H}_K$ 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

$$\Sigma_{\mathbb{P}} = \mathbb{E}_{x\sim\mathbb{P}}\left([K(\cdot,x)-\mu_{\mathbb{P}}]\otimes[K(\cdot,x)-\mu_{\mathbb{P}}]\right)\in\mathcal{L}_2(\mathcal{H}_K), \tag{9}$$

where the expectation (integral) is again meant in Bochner sense. $\Sigma_{\mathbb{P}}$ is non-negative, self-adjoint, moreover it has covariance-reproducing property $\langle f,\Sigma_{\mathbb{P}}f\rangle_K = \mathbb{E}_{x\sim\mathbb{P}}[f(x)-\mathbb{P}f]^2$. It is known that $\|A\|\leq\|A\|_2\leq\|A\|_1$.

3 MAIN RESULTS

Below we present our main results on the MONK estimators, followed by a discussion. We allow that N_c elements$((x_{n_j})_{j=1}^{N_c}$) of the sample $x_{1:N}$ are arbitrarily corrupted (In MMD estimation $\{(x_{n_j},y_{n_j})\}_{j=1}^{N_c}$ can be contaminated). The number of corrupted samples can be (almost) half of the number of blocks, in other words, there exists $\delta\in(0,1/2]$ such that $N_c\leq Q(1/2-\delta)$. If the data are free from contaminations, then $N_c=0$ and $\delta=1/2$. Using these notations, we can prove the following optimal sub-Gaussian deviation bounds on the MONK estimators.

Theorem 1 (Consistency & outlier-robustness of $\hat{\mu}_{\mathbb{P},Q}$)**.** *Assume that $\Sigma_{\mathbb{P}}\in\mathcal{L}_1(\mathcal{H}_K)$. Then, for any $\eta\in(0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q\in(N_c/(1-\delta),N/2)$, with probability at least $1-\eta$,*

$$\|\hat{\mu}_{\mathbb{P},Q}-\mu_{\mathbb{P}}\|_K \leq \frac{12\left(1+\sqrt{2}\right)}{\delta}\max\left(\sqrt{\frac{6\|\Sigma_{\mathbb{P}}\|\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}(\Sigma_{\mathbb{P}})}{N}}\right).$$

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, for any $\eta\in(0,1)$ such that $Q = 72\delta^{-2}\ln(1/\eta)$ satisfies $Q\in(N_c/(1-\delta),N/2)$, with probability at least $1-\eta$,*

$$\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})\right| \leq \frac{12}{\delta}\max\left(\sqrt{\frac{(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)\ln(1/\eta)}{\delta N}}, 2\sqrt{\frac{\mathrm{Tr}(\Sigma_{\mathbb{P}})+\mathrm{Tr}(\Sigma_{\mathbb{Q}})}{N}}\right).$$

Proof (sketch). The technical challenge is to get the optimal deviation bounds under the (mild) trace-class assumption. The reasonings for the mean embedding

and MMD follow a similar high-level idea; here we focus on the former. First we show that the analysis can be reduced to the unit ball in $\mathcal{H}_K$ by proving that

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

where

$$\begin{aligned} r_{Q,N} &= \sup_{f \in B_K} \mathrm{MON}_Q\big[x \mapsto \langle f, K(\cdot,x) - \mu_{\mathbb{P}} \rangle_K\big] \\ &= \sup_{f \in B_K} \operatorname*{med}_{q \in [Q]} \{r(f,q)\} \end{aligned}$$

with $r(f,q) = \left\langle f, \mu_{S_q} - \mu_{\mathbb{P}} \right\rangle_K$. The Chebyshev inequality with a Lipschitz argument allows us to control the probability of the event $\{r_{Q,N} \leq \epsilon\}$ using the variable $Z = \sup_{f \in B_K} \sum_{q \in U} [\phi(2r(f,q)/\epsilon) - \mathbb{E}\phi(2r(f,q)/\epsilon)]$, where U stands for the indices of the uncorrupted blocks and $\phi(t) = (t-1)\mathbb{I}_{1 \leq t \leq 2} + \mathbb{I}_{t \geq 2}$. The bounded difference property of the Z supremum of empirical processes guarantees its concentration around the expectation by using the McDiarmid inequality. The symmetrization technique combined with the Talagrand's contraction principle of Rademacher processes (thanks to the Lipschitz property of ϕ), followed by an other symmetrization leads to the deviation bound. Details are provided in Section A.1-A.2 (for Theorem 1-2) in the supplementary material. □

Remarks:

- Dependence on N: These finite-sample guarantees show that the MONK estimators
  - have optimal $N^{-1/2}$-rate—by recalling [72, 71]'s discussed results—, and
  - they are robust to outliers, providing consistent estimates with high probability even under arbitrary adversarial contamination (affecting less than half of the samples).
- Dependence on δ: Recall that larger δ corresponds to less outliers, i.e., cleaner data in which case the bounds above become tighter. In other words, making use of medians the MONK estimators show robustness to outliers; this property is a nice byproduct of our optimal sub-Gaussian deviation bound. Whether this robustness to outliers is optimal in the studied setting is an open question.
- Dependence on Σ: It is worth contrasting the rates obtained in Theorem 1 and that of the tournament procedures [46] derived for the finite-dimensional case. The latter paper elegantly resolved a long-lasting open question concerning the optimal dependency in terms of Σ. Theorem 1 proves the same dependency in the infinite-dimensional case, while giving rise to computionally tractable algorithms (Section 4).
- Separation rate: 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 COMPUTING THE MONK ESTIMATOR

This section is dedicated to the computation[4] of the analyzed MONK estimators; particularly we will focus on the MMD estimator given in Eq. (6). Numerical illustrations are provided in Section 5. Recall that the MONK estimator for MMD [Eq. (6)] is given by

$$\begin{aligned} &\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) \qquad (10) \\ &= \sup_{f \in B_K} \operatorname*{med}_{q \in [Q]} \left\{ \frac{1}{|S_q|} \sum_{j \in S_q} f(x_j) - \frac{1}{|S_q|} \sum_{j \in S_q} f(y_j) \right\}. \end{aligned}$$

By the representer theorem [59], the optimal f can be expressed as

$$f(\mathbf{a},\mathbf{b}) = \sum_{n \in [N]} a_n K(\cdot, x_n) + \sum_{n \in [N]} b_n K(\cdot, y_n), \quad (11)$$

where $\mathbf{a} = (a_n)_{n \in [N]} \in \mathbb{R}^N$ and $\mathbf{b} = (b_n)_{n \in [N]} \in \mathbb{R}^N$. Denote $\mathbf{c} = [\mathbf{a};\mathbf{b}] \in \mathbb{R}^{2N}$, $\mathbf{K} = [\mathbf{K}_{xx}, \mathbf{K}_{xy}; \mathbf{K}_{yx}, \mathbf{K}_{yy}] \in \mathbb{R}^{2N \times 2N}$, $\mathbf{K}_{xx} = [K(x_i,x_j)]_{i,j \in [N]} \in \mathbb{R}^{N \times N}$, $\mathbf{K}_{xy} = [K(x_i,y_j)]_{i,j \in [N]} = \mathbf{K}_{yx}^* \in \mathbb{R}^{N \times N}$, $\mathbf{K}_{yy} = [K(y_i,y_j)]_{i,j \in [N]} \in \mathbb{R}^{N \times N}$. With these notations, the optimisation problem (10) can be rewritten as

$$\max_{\mathbf{c} \in \mathbb{R}^{2N} : \mathbf{c}^* \mathbf{K} \mathbf{c} \leq 1} \operatorname*{med}_{q \in [Q]} \Big\{ |S_q|^{-1} [\mathbf{1}_q; -\mathbf{1}_q]^* \mathbf{K} \mathbf{c} \Big\}, \quad (12)$$

where $\mathbf{1}_q \in \mathbb{R}^N$ is indicator vector of the block S_q. To enable efficient optimization we follow a block-coordinate descent (BCD)-type scheme: choose the $q_{\mathrm{m}} \in [N]$ index for which the median is attained in (12), and solve

$$\max_{\mathbf{c} \in \mathbb{R}^{2N} : \mathbf{c}^* \mathbf{K} \mathbf{c} \leq 1} |S_{q_{\mathrm{m}}}|^{-1} [\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]^* \mathbf{K} \mathbf{c}. \quad (13)$$

This optimization problem can be solved analytically:

$$\mathbf{c} = \frac{[\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]}{\|\mathbf{L}^* [\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]\|_2} \ .$$

where $\mathbf{L}$ is the Cholesky factor of $\mathbf{K}$ ($\mathbf{K} = \mathbf{L}\mathbf{L}^*$). The observations are shuffled after each iteration. The pseudo-code of the final MONK BCD estimator is summarized in Algorithm 1.

[4]The Python code reproducing our numerical experiments is enclosed in the supplement. It will also be publicly released upon acceptance of the manuscript.

Table 1: Computational complexity of MMD estimators. N: sample number, Q: number of blocks, T: number of iterations.

Method	Complexity
U-Stat	$\mathcal{O}\left(N^2\right)$
MONK BCD	$\mathcal{O}\left(N^3 + T\left[N^2 + Q\log(Q)\right]\right)$
MONK BCD-Fast	$\mathcal{O}\left(\frac{N^3}{Q^2} + T\left[\frac{N^2}{Q} + Q\log(Q)\right]\right)$

Notice that computing $\mathbf{L}$ in MONK BCD costs $O(N^3)$, which can be prohibitive for large sample size. In order to alleviate this bottleneck we also consider an approximate version of MONK BCD (referred to as MONK BCD-Fast), where the $\sum_{n\in[N]}$ summation after plugging (11) to (10) is replaced with $\sum_{n\in S_q}$:

$$\max_{\mathbf{c}=[\mathbf{a},\mathbf{b}]\in\mathbb{R}^{2N}:\mathbf{c}^*\mathbf{K}\mathbf{c}\leq 1} \operatorname*{med}_{q\in[Q]} \left\{ \frac{1}{|S_q|}\sum_{j,n\in S_q}[a_n K(x_j,x_n)+ b_n K(x_j,y_n)] - \frac{1}{|S_q|}\sum_{j,n\in S_q}[a_n K(y_j,x_n)+b_n K(y_j,y_n)] \right\}.$$

This modification allows local computations restricted to blocks and improved running time. The samples are shuffled periodically (e.g., at every 10th iterations) to renew the blocks. The resulting method is presented in Algorithm 2. The computational complexity of the different MMD estimators are summarized in Table 1.

5 NUMERICAL ILLUSTRATIONS

In this section, we demonstrate the performance of the proposed MONK estimators. We exemplify the idea on the MMD estimator [Eq. (6)] with the BCD optimization schemes (MONK BCD and MONK BCD-Fast) discussed in Section 4. Our baseline is the classical U-statistic based MMD estimator [Eq. (8); referred to as U-Stat in the sequel].

The primary goal in the first set of experiments is to understand and demonstrate various aspects of the estimators for $(K,\mathbb{P},\mathbb{Q})$ triplets [53, Table 3.3] when analytical expression is available for MMD. This is the case for polynomial and RBF kernels (K), with Gaussian distributions ($\mathbb{P}$, $\mathbb{Q}$). Notice that in the first (second) case the features are unbounded (bounded). Our second numerical example illustrates the applicability of the studied MONK estimators in biological context, in discriminating DNA subsequences with string kernel.

Experiment-1: We used the quadratic and the RBF kernel with bandwith $\sigma = 1$ for demonstration purposes and investigated the estimation error compared to the true MMD value: $|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})|$. The errors are aggregates over 100 Monte-Carlo simulations, summarized in the median and quartile values. The number of samples (N) was chosen from $\{200, 400, \ldots, 2000\}$.

We considered three different experimental settings for $(\mathbb{P},\mathbb{Q})$ and the absence/presence of outliers:

1. Gaussian distributions with no outliers: In this case $\mathbb{P} = \mathcal{N}\left(\mu_1,\sigma_1^2\right)$ and $\mathbb{Q} = \mathcal{N}\left(\mu_2,\sigma_2^2\right)$ were normal where $(\mu_1,\sigma_1) \neq (\mu_2,\sigma_2)$, μ_1, σ_1, μ_2, σ_2 were randomly chosen from the $[0,1]$ interval, and then their values were fixed. The estimators had access to $(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and $(y_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$.
2. Gaussian distributions with outliers: This setting is a corrupted version of the first one. Particularly, the dataset consisted of $(x_n)_{n=1}^{N-5} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, $(y_n)_{n=1}^{N-5} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$, while the remaining 5-5 samples were set to $x_{N-4} = \ldots = x_N = 2000$, $y_{N-4} = \cdots = y_N = 4000$.
3. Pareto distribution without outliers: In this case $\mathbb{P} = \mathbb{Q} = Pareto(3)$ hence $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) = 0$ and the estimators worked based on $(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ and $(y_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{Q}$.

The 3 experiments were constructed to understand different aspects of the estimators: how a few outliers can ruin classical estimators (as we move from Experiment-1 to Experiment-2); in Experiment-3 the heavyness of the tail of a Pareto distribution makes the task non-trivial.

Our results on the three datasets with various Q choices are summarized in Fig. 1. As we can see from Fig. 1a and 1b in the outlier-free case, the MONK estimators are slower than the U-statistic based one; the accuracy is of the same order for both kernels. As demonstrated by Fig. 1c in the corrupted setup even a small number of outliers can completely ruin traditional MMD estimators for unbounded features while the MONK estimators are naturally robust to outliers with suitable choice of Q[5] ; this is precisely the setting the MONK estimators were designed for. In case of bounded kernels [Fig. 1d], by construction, traditional MMD estimators are naturally resistant to outliers; the MONK BCD-Fast method achieves comparable performance. In the final Pareto experiment [Fig. 1e-1f] where the distribution produces "natural outliers", again MONK estimators are more robust with respect to corruption than the one relying on U-statistics in the case of polynomial kernel. These experiments illustrate the power of the studied MONK schemes: these estimators achieve comparable performance in case of

[5] In this case of unknown N_c, one can choose Q adaptively by the Lepski method (see for example [14]) at the price of increasing the computational effort.

Algorithm 1 MONK BCD estimator for MMD

Input: Aggregated Gram matrix: $\mathbf{K}$ with Cholesky factor $\mathbf{L}$ ($\mathbf{K} = \mathbf{L}\mathbf{L}^*$).
for all $t = 1, \dots, T$ **do**
Generate a random permutation of $[N]$: σ.
Shuffle the samples according to σ: $S_q = \left\{\sigma\left((q-1)\frac{N}{Q}+1\right), \dots, \sigma\left(q\frac{N}{Q}\right)\right\}$, for $\forall q \in [Q]$.
Find the block attaining the median (q_{m}): $\frac{1}{|S_{q_{\mathrm{m}}}|}[\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]^*\mathbf{K}\mathbf{c} = \underset{q\in[Q]}{\mathrm{med}}\left(\frac{1}{|S_q|}[\mathbf{1}_q; -\mathbf{1}_q]^*\mathbf{K}\mathbf{c}\right)$.
Compute the coefficient vector: $\mathbf{c} = \frac{[\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]}{\|\mathbf{L}^*[\mathbf{1}_{q_{\mathrm{m}}}; -\mathbf{1}_{q_{\mathrm{m}}}]\|_2}$.
Output: $\underset{q\in[Q]}{\mathrm{med}}\left(\frac{1}{|S_q|}[\mathbf{1}_q; -\mathbf{1}_q]^*\mathbf{K}\mathbf{c}\right)$

Algorithm 2 MONK BCD-Fast estimator for MMD

Input: Aggregated Gram matrix: $\mathbf{K}$ with Cholesky factor $\mathbf{L}$ ($\mathbf{K} = \mathbf{L}\mathbf{L}^*$). Incides at which we shuffle: J.
for all $t = 1, \dots, T$ **do**
if $t \in J$ **then**
Generate a random permutation of $[N]$: σ.
Shuffle the samples according to σ: $S_q = \left\{\sigma\left((q-1)\frac{N}{Q}+1\right), \dots, \sigma\left(q\frac{N}{Q}\right)\right\}$, for $\forall q \in [Q]$.
Compute the Gram matrices and the Cholesky factors on each block $\mathbf{K}_q$ and $\mathbf{L}_q$ for $q \in [Q]$.
Find the block[a] attaining the median (q_{m}): $\frac{1}{|S_{q_{\mathrm{m}}}|}[\mathbb{1}_{q_{\mathrm{m}}}; -\mathbb{1}_{q_{\mathrm{m}}}]^*\mathbf{K}_{q_{\mathrm{m}}}\mathbf{c}_{q_{\mathrm{m}}} = \underset{q\in[Q]}{\mathrm{med}}\left(\frac{1}{|S_q|}[\mathbb{1}_q; -\mathbb{1}_q]^*\mathbf{K}_q\mathbf{c}_q\right)$.
Update the coefficient vector: $\mathbf{c}_{q_{\mathrm{m}}} = \frac{[\mathbb{1}_{q_{\mathrm{m}}}; -\mathbb{1}_{q_{\mathrm{m}}}]}{\left\|\mathbf{L}^*_{q_{\mathrm{m}}}[\mathbb{1}_{q_{\mathrm{m}}}; -\mathbb{1}_{q_{\mathrm{m}}}]\right\|_2}$.
Output: $\underset{q\in[Q]}{\mathrm{med}}\left(\frac{1}{|S_q|}[\mathbb{1}_q; -\mathbb{1}_q]^*\mathbf{K}_q\mathbf{c}_q\right)$

[a] $\mathbb{1}_q \in \mathbb{R}^{|S_q|}$ denotes the vector of ones of size $|S_q|$.

bounded features, while for unbounded features they can efficiently cope with the presence of outliers.

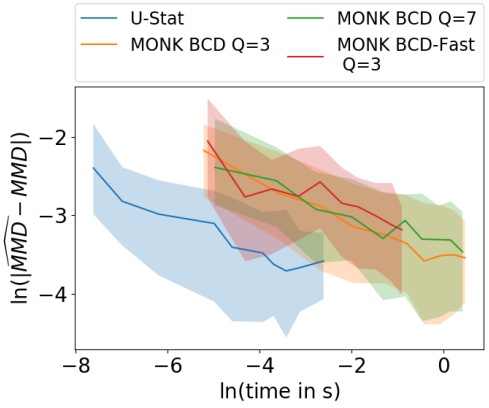
Experiment-2 (discrimination of DNA subsequences): In order to demonstrate the applicability of our estimators in biological context, we chose a DNA benchmark from the UCI repository [15], the Molecular Biology (Splice-junction Gene Sequences) Data Set. The dataset consists of 3190 instances of 60-character-long DNA subsequences. The problem is to recognize, given a sequence of DNA, the boundaries between exons (the parts of the DNA sequence retained after splicing) and introns (the parts of the DNA sequence that are spliced out). This task consists of two subproblems, identifying the exon/intron boundaries (referred to as EI sites) and the intron/exon boundaries (IE sites).[6] We took 1532 of these samples by selecting $N = 766$ instances from both the EI and the IE classes (the class of those being neither EI nor IE is more heterogeneous and thus we dumped it from the study), and investigated the discriminability of the EI and IE categories. We represented the DNA sequences as strings ($\mathcal{X}$), chose the Subsequent String Kernel to compute MMD, and tuned the kernel parameter by using 5-fold cross-validation to maximize the distance between intra-class MMD and inter-class MMD. Our results obtained from 50 Monte Carlo simulations (each time with a random 70% percent of the samples) summarized in Figure 2 show that the estimated inter-class $\widehat{\mathrm{MMD}}(\mathrm{EI}, \mathrm{IE})$ distances are significantly bigger than the two intra-class $\widehat{\mathrm{MMD}}(\mathrm{EI}, \mathrm{EI})$ and $\widehat{\mathrm{MMD}}(\mathrm{IE}, \mathrm{IE})$ ones. The MONK BCD and the U-Stat techniques perform similarly in terms of both the precision of the estimates and running time. The MONK BCD-Fast method accelerates the computation by its block-wise operation for the computationally heavy string kernel, while offering somewhat less distinctive discriminability values. These results illustrate the applicability of our estimators in biology.

[6] In the biological community, IE borders are referred to as "acceptors" while EI borders are referred to as "donors".

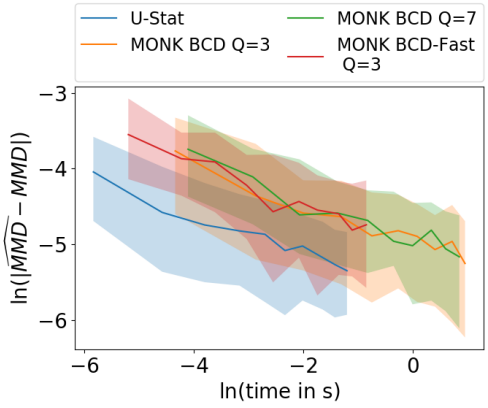
References

[1] Md. Ashad Alam, Kenji Fukumizu, and Yu-Ping Wang. Influence function and robust variant of kernel canonical correlation analysis. *Neurocomputing (to appear)*, 304:12–29, 2018.

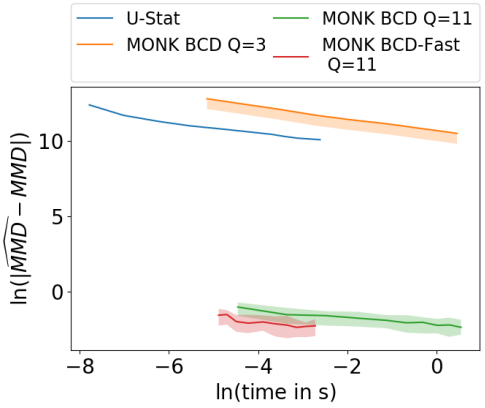
[2] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1, part 2):137–147, 1999.



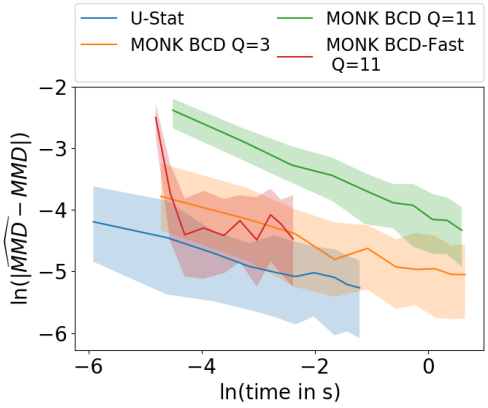
(a) Gaussian distribution, $N_c = 0$ (no outlier), quadratic kernel



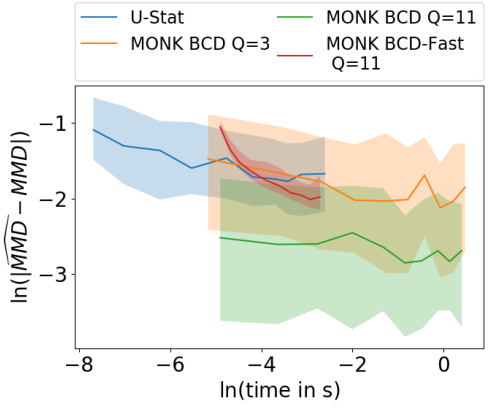
(b) Gaussian distribution, $N_c = 0$ (no outlier), RBF kernel



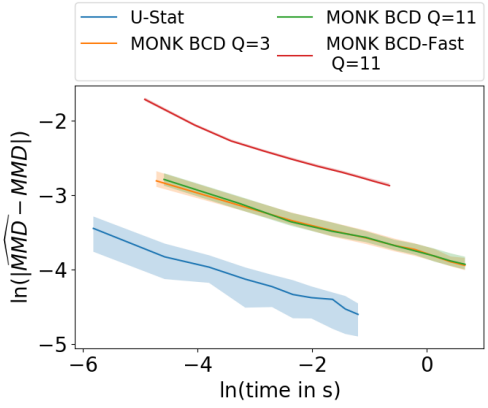
(c) Gaussian distribution $N_c = 5$ outliers, quadratic kernel



(d) Gaussian distribution, $N_c = 5$ outliers, RBF kernel



(e) Pareto distribution, quadratic kernel



(f) Pareto distribution, RBF kernel

Figure 1: Performance of the MMD estimators: median and quartiles of $\ln(|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}(\mathbb{P},\mathbb{Q})|)$. Rows from top to bottom: Experiment-1 – Experiment-3. Left: quadratic kernel, right: RBF kernel.

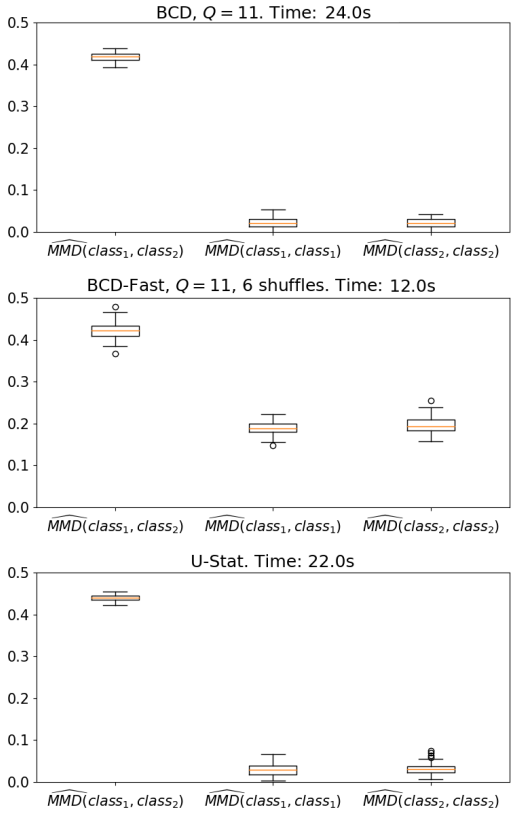


Figure 2: Intra-class and inter-class MMD estimates computed by the three algorithms; the median and quartiles are represented by the boxplots.

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

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

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

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

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

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

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

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

[11] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In *NIPS*, pages 625–632, 2001.

[12] Marco Cuturi. Fast global alignment kernels. In *ICML*, pages 929–936, 2011.

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

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

[15] Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. (`http://archive.ics.uci.edu/ml`).

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

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

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

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

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

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

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

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

[24] David Haussler. 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`).

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

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

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

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

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

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

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

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

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

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

[35] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, (23):12991–13008, 2015.

[36] Risi Kondor and Horace Pan. The multiscale Laplacian graph kernel. In *NIPS*, pages 2982–2990, 2016.

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

[38] Ho Chung Leon Law, Dougal J. Sutherland, Dino Sejdinovic, and Seth Flaxman. Bayesian approaches to distribution regression. *PMLR (AISTATS)*, 84:1167–1176, 2018.

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

[40] Guillaume Lecué and Matthieu Lerasle. Learning from MOM's principles: Le cam's approach. Technical report, CNRS, ENSAE, Université Paris Sud Orsay, 2017. (`https://arxiv.org/abs/1701.01961`).

[41] Guillaume Lecué and Matthieu Lerasle. Robust machine learning via median of means: theory and practice. Technical report, CNRS, ENSAE, Université Paris Sud Orsay, 2017. (`https://arxiv.org/abs/1711.10306`).

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

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

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

[45] Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *To appear in Journal of the European Mathematical Society*, 2016. (`https://arxiv.org/abs/1608.00757`).

[46] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *To appear in Annals of Statistics*, 2017. (`https://arxiv.org/absf/1702.00482`).

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

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

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

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

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

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

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

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

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

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

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

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

[59] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *COLT*, pages 416–426, 2001.

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

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

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

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

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

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

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

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

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

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

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

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

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

[73] Vladimir N. Vapnik. *The nature of statistical learning theory*. Statistics for Engineering and Information Science. Springer-Verlag, New York, second edition, 2000.

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

[75] Makoto Yamada, Yuta Umezu, Kenji Fukumizu, and Ichiro Takeuchi. Post selection inference with kernels. In *AISTATS (PMLR)*, volume 84, pages 152–160, 2018.

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

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

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

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

Supplement

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

A PROOFS OF THEOREM 1 AND THEOREM 2

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

$$\begin{aligned}&\|f - K(\cdot,x)\|_K^2 - \|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2 \\ &= \|f - \mu_{\mathbb{P}}\|_K^2 - 2\langle f - \mu_{\mathbb{P}}, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K . \end{aligned} \tag{14}$$

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

$$\begin{aligned}&\|e\|_K^2 - 2 r_{Q,N} \|e\|_K \\ &\overset{(a)}{\leq} \|e\|_K^2 - 2\mathrm{MON}_Q\left[\left\langle \frac{e}{\|e\|_K}, K(\cdot,x) - \mu_{\mathbb{P}}\right\rangle_K \|e\|_K\right] \\ &\overset{(b)}{\leq} \mathrm{MON}_Q\left[\|e\|_K^2 - 2\left\langle \frac{e}{\|e\|_K}, K(\cdot,x) - \mu_{\mathbb{P}}\right\rangle_K \|e\|_K\right] \\ &\overset{(c)}{\leq} \mathrm{MON}_Q\left[\|\hat{\mu}_{\mathbb{P},Q} - K(\cdot,x)\|_K^2 - \|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2\right] \\ &\overset{(d)}{\leq} \sup_{g \in \mathcal{H}_K} \mathrm{MON}_Q\left[\|\hat{\mu}_{\mathbb{P},Q} - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\right] \\ &\overset{(e)}{\leq} \sup_{g \in \mathcal{H}_K} \mathrm{MON}_Q\left[\|\mu_{\mathbb{P}} - K(\cdot,x)\|_K^2 - \|g - K(\cdot,x)\|_K^2\right] \\ &\overset{(f)}{=} \sup_{g \in \mathcal{H}_K}\Big\{2\mathrm{MON}_Q\Big[\underbrace{\langle \tilde{g}, K(\cdot,x) - \mu_{\mathbb{P}}\rangle_K}_{\|\tilde{g}\|_K \left\langle \frac{\tilde{g}}{\|\tilde{g}\|_K}, K(\cdot,x) - \mu_{\mathbb{P}}\right\rangle_K}\Big] - \|\tilde{g}\|_K^2\Big\} \\ &\overset{(g)}{=} \sup_{g \in \mathcal{H}_K}\Big\{2\|\tilde{g}\|_K r_{Q,N} - \|\tilde{g}\|_K^2\Big\} \overset{(h)}{\leq} r_{Q,N}^2, \end{aligned} \tag{15}$$

where we used in (a) the definition of $r_{Q,N}$, (b) the linearity[7] of $\mathrm{MON}_Q[\cdot]$, (c) Eq. (14), (d) $\sup_g$, (e) the definition of $\hat{\mu}_{\mathbb{P},Q}$, (f) Eq. (14) 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. (15), 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

$$\|\hat{\mu}_{\mathbb{P},Q} - \mu_{\mathbb{P}}\|_K \leq \left(1 + \sqrt{2}\right) r_{Q,N}. \tag{16}$$

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

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

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

The N_c corrupted samples can affect (at most) N_c of the $(S_q)_{q \in [Q]}$ 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

$$\forall f \in B_K : \underbrace{|\{q \in U : r(f,q) \geq \epsilon\}|}_{\sum_{q \in U} \mathbb{I}_{r(f,q) \geq \epsilon}} + N_c \leq \frac{Q}{2}, \tag{17}$$

then for $\forall f \in B_K$, $\mathrm{med}_{q \in [Q]}\{r(f,q)\} \leq \epsilon$, i.e. $\sup_{f \in B_K} \mathrm{med}_{q \in [Q]}\{r(f,q)\} \leq \epsilon$. Thus, our task boils down to controlling the event in (17) 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

$$\begin{aligned}&\mathbb{E}_{x_i \sim \mathbb{P}} \langle f, k(\cdot,x_i) - \mu_{\mathbb{P}}\rangle_K^2 = \langle f, \Sigma_{\mathbb{P}} f\rangle_K \\ &\leq \|f\|_K \|\Sigma_{\mathbb{P}} f\|_K \leq \|f\|_K^2 \|\Sigma_{\mathbb{P}}\| = \|\Sigma_{\mathbb{P}}\| \end{aligned} \tag{18}$$

using the reproducing property of the kernel and the covariance operator, the Cauchy-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 $\mathbb{E}[z^2] = \mathbb{E}\left\langle f, \mu_{S_q} - \mu_{\mathbb{P}}\right\rangle_K^2 = \frac{Q}{N}\mathbb{E}_{x_i \sim \mathbb{P}} \langle f, k(\cdot,x_i) - \mu_{\mathbb{P}}\rangle_K^2$ and Eq. (18) 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 $\epsilon \geq 2\sqrt{\frac{\|\Sigma_{\mathbb{P}}\| Q}{\alpha N}}$.

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

- **Reduction to ϕ**: As a result

$$\sum_{q\in U}\mathbb{P}\left(r(f,q)\geq\frac{\epsilon}{2}\right)\leq|U|\alpha$$

happens if and only if

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

Let us introduce $\phi:t\in\mathbb{R}\to(t-1)\mathbb{I}_{1\leq t\leq 2}+\mathbb{I}_{t\geq 2}$. 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

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

- **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 $\tau:=Q\beta^2/8$ to get linear scaling in Q on the r.h.s.) implies that

$$\mathbb{P}\left(Z<\mathbb{E}_{x_V}[Z]+Q\beta\right)\geq 1-e^{-\frac{Q\beta^2}{8}}.$$

- **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 [67, Prop. 7.10], [42] 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 contraction principle of Rademacher processes [42], [34, Theorem 2.3] one gets

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

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

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

In (a), we proceed as follows:

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

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 $(e'_ne''_n\left\langle f,K(\cdot,x_n)-\mu_{\mathbb{P}}\right\rangle_K)_{n\in V}\overset{\text{distr}}{=}(e'_n\left\langle f,K(\cdot,x_n)-\mu_{\mathbb{P}}\right\rangle_K)_{n\in V}$.

In step (b), we had

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

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 $1-e^{-\frac{Q\beta^2}{8}}$. 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 (α,β,ϵ) 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

$$Q\alpha+Q\beta+\frac{8Q\sqrt{\operatorname{Tr}(\Sigma_{\mathbb{P}})}}{\epsilon\sqrt{N}}+N_c\leq\frac{Q}{2}. \tag{19}$$

Choosing $\alpha=\beta=\frac{\delta}{3}$ in Eq. (19), 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. Since $N_c\leq Q(\frac{1}{2}-\delta)$, we got

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

with probability at least $1-\mathrm{e}^{-\frac{Q\delta^2}{72}}$. With an $\eta=\mathrm{e}^{-\frac{Q\delta^2}{72}}$, and hence $Q=\frac{72\ln\left(\frac{1}{\eta}\right)}{\delta^2}$ reparameterization Theorem 1 follows.

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

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

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

- $\mathrm{MMD}_Q(\mathbb{P},\mathbb{Q})-\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})\leq r_{Q,N}$: Let $a_f:=\left\langle f,\mu_{\mathbb{P}}-\mu_{\mathbb{Q}}\right\rangle_K$ and $b_f:=\operatorname*{med}_{q\in[Q]}\left\{\left\langle f,(\mu_{\mathbb{P}}-\mu_{\mathbb{Q}})-(\mu_{S_q,\mathbb{P}}-\mu_{S_q,\mathbb{Q}})\right\rangle_K\right\}$. Then

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

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):

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

Step-2: Our goal is to control

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

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

- **Median rephrasing**:

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

  Thus, $\forall f\in B_K:|\{q\in U:|r(f,q)|\ge\epsilon\}|+N_c\le\frac{Q}{2}$, implies $\sup_{f\in B_K}\left|\operatorname{med}_{q\in[Q]}\{r(f,q)\}\right|\le\epsilon$.
- **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

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

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

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

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

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

- These results imply

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

  $\epsilon\ \ge\ \max\left(2\sqrt{\frac{3(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|)Q}{\delta N}},\frac{24}{\delta}\sqrt{\frac{\operatorname{Tr}\left(\Sigma_{\mathbb{P}}\right)+\operatorname{Tr}\left(\Sigma_{\mathbb{Q}}\right)}{N}}\right)$, $\alpha=\beta=\frac{\delta}{3}$ choice gives that

$$\begin{aligned}
&\left|\widehat{\mathrm{MMD}}_Q(\mathbb{P},\mathbb{Q})-\mathrm{MMD}(\mathbb{P},\mathbb{Q})\right|\\
&\le 2\max\left(\sqrt{\frac{3\left(\|\Sigma_{\mathbb{P}}\|+\|\Sigma_{\mathbb{Q}}\|\right)Q}{\delta N}},\frac{12}{\delta}\sqrt{\frac{\operatorname{Tr}\left(\Sigma_{\mathbb{P}}\right)+\operatorname{Tr}\left(\Sigma_{\mathbb{Q}}\right)}{N}}\right)
\end{aligned}$$

  with probability at least $1-e^{-\frac{Q\delta^2}{72}}$. $\eta=e^{-\frac{Q\delta^2}{72}}$, i.e. $Q=\frac{72\ln\left(\frac{1}{\eta}\right)}{\delta^2}$ reparameterization finishes the proof of Theorem 2.

B TECHNICAL LEMMAS

Lemma 3 (Supremum)**.**

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

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 $\mathcal{X}$, $Z(x_V)=\sup_{f\in B_K}\sum_{q\in U}\left[\phi\left(\frac{2\left\langle f,\mu_{S_q}-\mu_{\mathbb{P}}\right\rangle_K}{\epsilon}\right)-\mathbb{E}\phi\left(\frac{2\left\langle f,\mu_{S_q}-\mu_{\mathbb{P}}\right\rangle_K}{\epsilon}\right)\right]$, 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*

$$\sup_{x_V\in\mathcal{X}^{|V|},x'_i\in\mathcal{X}}\left|Z\left(x_V\right)-Z\left(x'_{V_i}\right)\right|\le 4,\ \forall i\in V.$$

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

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

In this case

$$\begin{aligned}
&\left|Z\left(x_V\right)-Z\left(x'_{V_i}\right)\right|\\
&=\left|\sup_{f\in B_K}\sum_{q\in U}Y_q-\sup_{f\in B_K}\left(\sum_{q\in U\setminus\{r\}}Y_q+Y'_r\right)\right|\\
&\overset{(a)}{\le}\sup_{f\in B_K}|Y_r-Y'_r|\overset{(b)}{\le}\sup_{f\in B_K}\Big(\underbrace{|Y_r|}_{\le 2}+\underbrace{|Y'_r|}_{\le 2}\Big)\le 4,
\end{aligned}$$

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

□

Lemma 5 (Uniformly bounded separable Carathéodory family)**.** *Let $\epsilon>0$, $N\in\mathbb{Z}^+$, $Q\in\mathbb{Z}^+$, $M=N/Q\in\mathbb{Z}^+$, $\phi(t)=(t-1)\mathbb{I}_{1\le t\le 2}+\mathbb{I}_{t\ge 2}$, $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ 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: (i) $\sup_{f\in B_K}\|g_f\|_\infty<\infty$ where $\|g\|_\infty=$

$\sup_{u_{1:M}\in\mathcal{X}^M}|g(u_{1:M})|$, *(ii)* $u_{1:M}\mapsto g_f(u_{1:M})$ *is measurable for all* $f\in B_K$, *(iii)* $f\mapsto g_f(u_{1:M})$ *is continuous for all* $u_{1:M}\in\mathcal{X}^M$, *(iv)* B_K *is separable.*

Proof.

(i) $|\phi(t)|\leq 1$ for any t, hence $\|g_f\|_\infty\leq 1$ for all $f\in B_K$.

(ii) Any $f\in B_K$ is continuous since $\mathcal{H}_K\subset C(\mathcal{X})=\{h:\mathcal{X}\to\mathbb{R}\text{ continuous}\}$, so $u_{1:M}\mapsto(f(u_1),\dots,f(u_M))$ 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.

(ii) 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.

(iv) B_K is separable since $\mathcal{H}_K$ is so by assumption.

□

C 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*

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

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

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