MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means
Matthieu Lerasle, Zoltán Szabó, Timothée Mathieu, Guillaume Lecué

To cite this version:
Matthieu Lerasle, Zoltán Szabó, Timothée Mathieu, Guillaume Lecué. MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means. International Conference on Machine Learning (ICML), Jun 2019, Long Beach, United States. hal-01705881v5

HAL Id: hal-01705881
https://hal.archives-ouvertes.fr/hal-01705881v5
Submitted on 15 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means

Matthieu Lerasle 1 2 Zoltán Szabó 3 Timothée Mathieu 1 Guillaume Lecué 4

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 (Aronszajn, 1950) form the backbone of a tremendous number of successful applications in machine learning thanks to their power in capturing complex relations (Schölkopf & Smola, 2002; 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; Martíns 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; Konidor & Pan, 2016).

Given a kernel-enriched domain $(\mathcal{X}, K)$ one can represent probability distributions on $\mathcal{X}$ as a mean
\[
\mu_\mathcal{P} = \int_{\mathcal{X}} \varphi(x) d\mathcal{P}(x) \in \mathcal{H}_K, \quad \varphi(x) := K(\cdot, x),
\]
which is a point in the RKHS determined by $K$. This representation called mean embedding (Berlinet & Thomas-Agnan, 2004; Smola et al., 2007) induces a semi-metric on distributions called maximum mean discrepancy (MMD) (Smola et al., 2007; Gretton et al., 2012)
\[
\text{MMD}(\mathcal{P}, \mathcal{Q}) = \| \mu_\mathcal{P} - \mu_\mathcal{Q} \|_{\mathcal{H}_K}. \tag{1}
\]

With appropriate choice of the kernel, classical integral transforms widely used in probability theory and statistics can be recovered by $\mu_\mathcal{P}$; for example, if $\mathcal{X}$ equipped with the scalar product $\langle \cdot, \cdot \rangle$ is a Hilbert space, the kernel $K(x, y) = e^{\langle x, y \rangle}$ gives the moment-generating function, $K(x, y) = e^{\gamma \|x - y\|^2}$ ($\gamma > 0$) 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; Harachou et al., 2007; Gretton et al., 2012) or its differential private variant (Raj et al., 2018), independence (Gretton et al., 2008; Pfister et al., 2017) and goodness-of-fit testing (Jitkrittum et al., 2017; Balasubramian et al., 2017), domain

1Laboratoire de Mathématiques d’Orsay, Univ. Paris-Sud, France 2CNRS, Université Paris Saclay, France 3CMAP, École Polytechnique, Palaiseau, France 4CREST ENSAE ParisTech, France. Correspondence to: Matthieu Lerasle <matthieu.lerasle@math.u-psud.fr>, Zoltán Szabó <zoltan.szabo@polytechnique.edu>.


---

1Fukumizu et al. (2008); Sriperumbudur et al. (2010) provide conditions when MMD is a metric, i.e. $\mu$ is injective.

---
adaptation (Zhang et al., 2013) and generalization (Blanchard et al., 2017), change-point detection (Harchaoui & Cappé, 2007), probabilistic programming (Schölkopf et al., 2015), post selection inference (Yamada et al., 2018), 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), generative adversarial networks (Dziugaite et al., 2015; Li et al., 2015; Binkowski et al., 2018), understanding the dynamics of complex dynamical systems (Klus et al., 2018; 2019), or topological data analysis (Kusano et al., 2016), among many others; Muandet et al. (2017) provide a recent 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 \( \mu_2 \) is the empirical average. Plugging this estimate into 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 \( N_1^{-1/2} + N_2^{-1/2} \) (Tolstikhin et al., 2016) in case of radial universal kernels defined on \( \mathcal{X} = \mathbb{R}^d \).

A critical property of an estimator is its robustness to 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, time series 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\(^*\) 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., 2018), 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), or as density function estimators (Vandermeulen & Scott, 2013).

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) following the approach of Catoni (2012) 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 (Vapnik, 2000) or by more sophisticated approach such as the minmax procedure (Audibert & Catoni, 2011).

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, Minsker (2015); Minsker & Strawn (2017) consider the problem of estimating the mean of a Banach-space valued random variable using “geometrical” MONs. The estimators constructed by Minsker (2015); Minsker & Strawn (2017) 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, Lugosi & Mendelson (2019a); Lecué & Lerasle (2018) proposed to combine the classical MON estimators on \( \mathbb{R} \) in a “test” procedure that can be seen as a Le Cam test estimator (Le Cam, 1973). The achievement in (Lugosi & Mendelson, 2019a; Lecué & Lerasle, 2018) is that they were able to obtain optimal deviation bounds for the resulting estimator using the powerful so-called small-ball method of Kolchinskii & Mendelson (2015); Mendelson (2015). This approach was then extended to mean estimation \( \mathbb{R}^d \) by Lugosi & Mendelson (2019b) providing the first rate-optimal sub-Gaussian deviation bounds under minimal \( L^2 \)-assumptions. The constants of Lugosi & Mendelson (2019a); Lecué & Lerasle (2018); Lugosi & Mendelson (2019b) have been improved by Catoni & Giulini (2017) for the estimation of the mean in \( \mathbb{R}^d \) under \( L^4 \)-moment negative functions integrating to one.\(^*\)

\(^*\)Smoothing kernels extensively studied in the non-parametric statistical literature (Győrfi et al., 2002) are assumed to be non-negative functions integrating to one.
MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means

assumption and in least-squares regression under $L^4/L^2$-condition that is stronger than the small-ball assumption used by Lugosi & Mendelson (2019a); Lecué & Lerasle (2018). Unfortunately, these estimators are computationally intractable; their risk bounds however serve as an important baseline for computable estimators such as the minmax MON estimators in regression (Lecué & Lerasle, 2019).

Motivated by the computational intractability of the tournament procedure underlying the first-rate-optimal sub-Gaussian deviation bound holding under minimal assumptions in $\mathbb{R}^d$ (Lugosi & Mendelson, 2019b), Hopkins (2018) proposed a convex relaxation with polynomial, $O(N^{24})$ complexity where $N$ denotes the sample size. Cherapanamjeri et al. (2019) have recently designed an alternative convex relaxation requiring $O(N^3 + dN^2)$ computation which is still rather restrictive for large sample size and infeasible in infinite dimension.

Our goal is to extend the theoretical insight of Lugosi & Mendelson (2019b) from $\mathbb{R}^d$ to kernel-enriched domains. Particularly, we prove optimal sub-Gaussian deviation bounds for MON-based mean estimators in RKHSs which hold under minimal second-order moment assumptions. In order to achieve this goal, we use a different (min-max (Audibert & Catoni, 2011; Lecué & Lerasle, 2019)) construction which combined with properties specific to RKHSs (the mean-reproducing property of mean embedding and the integral probability metric representation of MMD) give rise to our practical MONK procedures. Thanks to the usage of medians the MONK estimators 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. $[M] := \{1, \ldots, M\}$, $u_S := (u_m)_{m \in S}, S \subseteq [M]$. For a set $S$, $|S|$ denotes its cardinality. $\mathcal{E}$ stands for expectation. 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 $\sigma$-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 $K$; $(\cdot, \cdot)_K := (\cdot, \cdot)_{\mathcal{H}_K}$, $\|\cdot\|_K := \|\cdot\|_{\mathcal{H}_K}$. $\mathcal{H}_K$ is separable by the separability of $\mathcal{X}$ and the continuity of $K$ (Steinwart & Christmann, 2008, Lemma 4.33). These assumptions on $\mathcal{X}$ and $K$ are assumed to hold throughout the paper.

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

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

(2)

where the integral is meant in Bochner sense; $\mu_\mathbb{P}$ exists iff $\int_{\mathcal{X}} \|K(\cdot, x)\|_K d\mathbb{P}(x) = \int_{\mathcal{X}} \sqrt{K(x, x)} d\mathbb{P}(x) < \infty$. 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:

$$\mu_\mathbb{P} = \arg\min_{f \in \mathcal{H}_K} \int_{\mathcal{X}} \|f - K(\cdot, x)\|_K^2 d\mathbb{P}(x).$$

(3)

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

$$\mu_\mathbb{P} = \arg\min_{f \in \mathcal{H}_K} \sup_{g \in \mathcal{H}_K} J(f, g),$$

(4)

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

$$\text{MON}_Q [h] = \text{med}_{q \in [Q]}\{\mathbb{P}_{S_q} h\} = \text{med}_{q \in [Q]}\{\langle h, \mu_{S_q} \rangle_K\},$$

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: $\text{MON}_1 [h] = N^{-1} \sum_{n=1}^N h(x_n)$.

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

$$\hat{\mu}_{\mathbb{P}, Q} = \hat{\mu}_{\mathbb{P}, Q}(x_{1:N}) \in \arg\min_{f \in \mathcal{H}_K} \sup_{g \in \mathcal{H}_K} J(f, g),$$

(5)

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

$$J(f, g) =$$

$$= \text{MON}_Q \left[ x \mapsto \|f - K(\cdot, x)\|_K^2 - \|g - K(\cdot, x)\|_K^2 \right].$$

\footnote{$\mathcal{H}_K$ is separable by the separability of $\mathcal{X}$ and the continuity of $K$ (Steinwart & Christmann, 2008, Lemma 4.33). These assumptions on $\mathcal{X}$ and $K$ are assumed to hold throughout the paper.}
When \( Q = 1 \), since \( \text{MON}_1 [h] \) is the empirical mean, we obtain the classical empirical mean based estimator: \( \tilde{\mu}_P,1 = \frac{1}{N} \sum_{n=1}^N K (\cdot, x_n) \).

One can use the mean embedding (2) to get a semi-metric on probability measures: the maximum mean discrepancy (MMD) of \( P \) and \( Q \) is

\[
\text{MMD}(P, Q) = \| \mu_P - \mu_Q \|_K = \sup_{f \in B_K} \langle f, \mu_P - \mu_Q \rangle_K,
\]

where \( B_K = \{ f \in \mathcal{H}_K : \| f \|_K \leq 1 \} \) is the closed unit ball around the origin in \( \mathcal{H}_K \). The second equality shows that MMD is a specific integral probability metric (Müller, 1997; Zolotarev, 1983). Assume that we have access to \( x_{1:N} \sim P, y_{1:N} \sim Q \) samples, where we assumed the size of the two samples to be the same for simplicity. Denote by \( P_{S,x} := \frac{1}{N} \sum_{i \in S} \delta_{x_i} \) the empirical measure associated to the subset \( S \subseteq \{ 1, \ldots, N \} \). For \( X \) a Hilbert space and \( A \in \mathcal{L}(\mathcal{H}) \) a bounded linear operator, \( \| A \| := \text{sup}_{\| f \|_K = 1} \| Af \|_K / \| f \|_K \) the operator norm of \( A \). Let \( \mathcal{L}(\mathcal{H}) = \{ A : \mathcal{H} \to \mathcal{H} \text{ linear operator : } \| A \| < \infty \} \) be the space of bounded linear operators. For any \( A \in \mathcal{L}(\mathcal{H}) \), let \( A^* \in \mathcal{L}(\mathcal{H}) \) 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 \( A \in \mathcal{L}(\mathcal{H}) \) 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 \( \mathcal{H}_K \). If \( A \in \mathcal{L}(\mathcal{H}) \) is called trace-class if \( \| A \|_2 := \text{Tr}(A^* A)^{1/2} \) is finite. If \( A \) is non-negative and self-adjoint, then \( A \) is trace-class if \( \text{Tr}(A) < \infty \); this will hold for the covariance operator (\( \Sigma_p \)), see Eq. (9).

\[ A \in \mathcal{L}(\mathcal{H}) \text{ is called Hilbert-Schmidt if } \| A \|_2 < \infty \text{ the class of trace-class and (Hilbert) space of Hilbert-Schmidt operators on } \mathcal{H}_K. \]

The tensor product of \( a, b \in \mathcal{H}_K \) is \( (a \otimes b)(c) := a(b,c)_K \) for all \( c \in \mathcal{H}_K \), \( a \otimes b \in \mathcal{L}(\mathcal{H}_K) \) and \( \| a \otimes b \|_2 = \| a \|_K \| b \|_K \). \( \mathcal{L}_2(\mathcal{H}_K) = \{ A \in \mathcal{L}(\mathcal{H}_K) : \| A \|_2 < \infty \} \) is 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 \( (a \otimes b)(c) := a(b,c)_K \) for all \( c \in \mathcal{H}_K \), \( a \otimes b \in \mathcal{L}(\mathcal{H}_K) \) and \( \| a \otimes b \|_2 = \| a \|_K \| b \|_K \). \( \Sigma_p \) is non-negative, self-adjoint, moreover it has covariance-reproducing property \( f, \Sigma_p f \rangle_K = \mathbb{E}_{x \sim p} [f(x) - \mathbb{E}_p f(x)]^2 \). It is known that \( \| A \|_2 \leq \| A \|_1 \).

3. Main Results

Below we present our main results on the MONK estimators, followed by a discussion. We allow that \( N_c \) elements of \( \{(x_{n_1}, y_{n_1})\}_{j=1}^{N_c} \) of the sample \( x_{1:N} \) are arbitrarily corrupted (in MMD estimation \( \{(x_{n_1}, y_{n_1})\}_{j=1}^{N_c} \) 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.

\[ \| \mu_{q,p} - \mu_p \|_K \leq \frac{12 (1 + \sqrt{2})}{\delta} \left[ \frac{\sqrt{6} \| \Sigma_p \| \ln(1/\eta)}{\delta N}, 2 \sqrt{\frac{\text{Tr}(\Sigma_p)}{N}} \right]. \]
Theorem 2 (Consistency & outlier-robustness of \( \widehat{\text{MMD}}_Q(\mathbb{P}, \mathbb{Q}) \)). Assume that \( \Sigma_p \) and \( \Sigma_q \in \mathcal{L}_1(\mathbb{H}_K) \). Then, for any \( \eta \) with Talagrand’s contraction principle of Rademacher with a Lipschitz argument allows us to control the probability of \( \mathbb{E} \) satisfies \( Q \in \left( \frac{N_c}{1/2 - \delta}, N/2 \right) \), with probability at least \( 1 - \eta \),

\[
\left| \widehat{\text{MMD}}_Q(\mathbb{P}, \mathbb{Q}) - \text{MMD}(\mathbb{P}, \mathbb{Q}) \right| \leq 12 \max_{\delta} \left( \frac{\left( \| \Sigma_p \| + \| \Sigma_q \| \right) \ln(1/\eta)}{\delta N}, \sqrt{\frac{\text{Tr}(\Sigma_p) + \text{Tr}(\Sigma_q)^2}{N}} \right).
\]

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 \( \mathbb{H}_K \) by proving that \( \| \mu_{\mathbb{P}, Q} - \mu_{\mathbb{P}, \mathbb{Q}} \|_{K} \leq (1 + \sqrt{2})r_{Q,N} \), where \( r_{Q,N} = \sup_{f \in B_K} \text{MON}(f \mapsto \langle f, K(\cdot, x) - \mu_{\mathbb{P}} \rangle_K) = \sup_{f \in B_K} \text{med}(r(f, g)) \) with \( r(f, g) = \langle f, \mu_{\mathbb{S}_{Q}^*} - \mu_{\mathbb{P}} \rangle_K \). The Chebyshev inequality with a Lipschitz argument allows us to control the probability of the event \( \{ \psi, N \leq \epsilon \} \) using the variable \( Z = \sup_{f \in B_K} \sum_{q \in \mathcal{U}} \phi(2r(f, q)/\epsilon) - \mathbb{E} \phi(2r(f, q)/\epsilon), \) where \( U \) stands for the indices of the uncorrupted blocks and \( \psi(t) = (t - 1)I_{1 \leq t \leq 2} + I_{t > 2} \). The bounded difference property of the Z suprema 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 \( \phi \)), followed by another 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 Tolstikhin et al. (2016; 2017)’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 \( \delta \): Recall that larger \( \delta \) 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 \( \Sigma \): It is worth contrasting the rates obtained in Theorem 1 and that of the tournament procedures (Lugosi & Mendelson, 2019b) derived for the finite-dimensional case. The latter paper elegantly resolved a long-lasting open question concerning the optimal dependency in terms of \( \Sigma \). Theorem 1 proves the same dependency in the infinite-dimensional case, while giving rise to computationally tractable algorithms (Section 4).

- Separation rate: Theorem 2 also shows 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 \( N^{-1/2} \).

- Breakdown point: Our finite-sample bounds imply that the proposed MONK estimators using \( Q \) blocks is resistant to \( Q/2 \) outliers. Since \( Q \) is allowed to grow with \( N \) (it can be chosen to be almost \( N/2 \)), this specifically means that the breakdown point of our estimators can be 25%.

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

\[
\widehat{\text{MMD}}_Q(\mathbb{P}, \mathbb{Q}) = \sup_{f \in B_K} \text{med}(q \in \mathcal{Q}) \left\{ \frac{1}{|S_q|} \sum_{j \in S_q} f(x_j) - \frac{1}{|S_q|} \sum_{j \in S_q} f(y_j) \right\}.
\]

By the representer theorem (Schölkopf et al., 2001), the optimal \( f \) can be expressed as

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

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

\[
\max_{e \in \mathbb{R}^{2N} : c^e \leq 1} \text{med}(q \in \mathcal{Q}) \left\{ |S_q|^{-1} |1_q; -1_q|^* K e \right\},
\]

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

\[
\max_{e \in \mathbb{R}^{2N} : c^e \leq 1} |S_{q_m}|^{-1} |1_{q_m}; -1_{q_m}|^* K e.
\]

\(^4\)The Python code reproducing our numerical experiments is available at https://bitbucket.org/TimotheeMathieu/monk-mmd; it relies on the ITE toolbox (Szabó, 2014).
Algorithm 1 MONK BCD estimator for MMD

Input: Aggregated Gram matrix: \( \mathbf{K} \) with Cholesky factor \( \mathbf{L} (\mathbf{K} = \mathbf{LL}^*) \).
for all \( t = 1, \ldots, T \) do
  Generate a random permutation of \([N]: \sigma\).
  Shuffle the samples according to \( \sigma \): for all \( q \in [Q] \)
  \[
  S_q = \left\{ \sigma \left( (q - 1) \frac{N}{Q} + 1 \right), \ldots, \sigma \left( \frac{N}{Q} \right) \right\}.
  \]
  Find the block attaining the median \((q_m)\):
  \[
  \frac{1}{|S_q|} \sum_{j,n \in S_q} [a_n K(x_j, x_n) + b_n K(y_j, y_n)] = \med_{q \in [Q]} \frac{1}{|S_q|} [1_{q_m} - 1_{q_m}]^* \mathbf{K} \mathbf{c}.
  \]
  Compute the coefficient vector: \( \mathbf{c} = \frac{1}{|S_q|} [1_{q_m} - 1_{q_m}]^* \mathbf{K} \mathbf{c} \).
end for
Output: \( \med_{q \in [Q]} \left( \frac{1}{|S_q|} [1_{q_m} - 1_{q_m}]^* \mathbf{K} \mathbf{c} \right) \)

This optimization problem can be solved analytically: \( \mathbf{c} = \frac{1}{|S_q|} [1_{q_m} - 1_{q_m}]^* \mathbf{K} \mathbf{c} \), where \( \mathbf{L} \) is the Cholesky factor of \( \mathbf{K} (\mathbf{K} = \mathbf{LL}^*) \). The observations are shuffled after each iteration. The pseudo-code of the final MONK BCD estimator is summarized in Algorithm 1.

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) into (10) is replaced with \( \sum_{n \in S_q} \):

\[
\max_{c = [a, b] \in \mathbb{R}^{2 \times (K \times Q)}} \med_{q \in [Q]} \left\{ \frac{1}{|S_q|} \sum_{j,n \in S_q} [a_n K(x_j, x_n) + b_n K(y_j, y_n)] - \frac{\sum_{j,n \in S_q} [a_n K(y_j, x_n) + b_n K(y_j, y_n)]}{|S_q|} \right\}.
\]

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 under-
stand and demonstrate various aspects of the estimators for \((K, P, Q)\) triplets (Muandet et al., 2017, Table 3.3) when analytical expression is available for MMD. This is the case for polynomial and RBF kernels \((K)\), with Gaussian distributions \((P, Q)\). 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 bandwidth \(\sigma = 1\) for demonstration purposes and investigated the estimation error compared to the true MMD value: \(\hat{\text{MMD}}_Q(P, Q) - \text{MMD}(P, Q)\). 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
These experiments illustrate the power of the studied MONK estimators. In particular, the dataset consisted 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). We took 1532 of these samples by selecting 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 (X), chose K as the String Subsequence Kernel (Lodhi et al., 2002) to compute MMD, and performed two-sample testing based on MMD using the MONK BCD, MONK BCD-Fast and U-Stat estimators. For completeness the pseudocode of the hypothesis test is detailed in Algorithm 3 (Section D). Q, the number of blocks in the MONK techniques, was equal to 5. The significance level was α = 0.05. To assess the variability of the results 400 Monte Carlo simulations were performed, each time uniformly sampling N points without replacement resulting in (Xn)n∈[N] and (Yn)n∈[N]. To provide more detailed insights the aggregated values of MMD(EI,IE)−q1−α, MMD(EI,IE)−q1−α and MMD(EI,IE)−q1−α are summarized in Fig. 2, where q1−α is the estimated (1−α)-quantile via B = 150 bootstrap permutations. In the ideal case, MMD − q1−α is positive (negative) in the inter-class (intra-class) experiments. As Fig. 2 shows all 3 techniques are able to solve the task, both in the inter-class (when the null hypothesis does not hold; Fig. 2a) and the intra-class experiment (null holds; Fig. 2b and Fig. 2c), and they converge to a good and stable performance as a function of the sample number. It is important to note that the MONK BCD-Fast method is especially well-adapted to problems where the kernel computation (such as the String Subsequence Kernel) or the sample size is a bottleneck, as its computation is often significantly faster compared to the U-Stat technique. For example, taking all the samples (N = 766) in the DNA benchmark with Q = 15, computing MONK BCD-Fast (U-Stat) takes 32 s (1m28 s). These results illustrate the applicability of our estimators in gene analysis.

Acknowledgements

Guillaume Lecué is supported by a grant of the French National Research Agency (ANR), “Investissements d’Avenir” (LabEx Ecodec/ANR-11-LABX-0047).

1In case of unknown N, one could choose Q adaptively by the Lepski method (see for example Devroye et al., 2016) at the price of increasing the computational effort. Though the resulting Q would increase the computational time, it would be adaptive thanks to its data-driven nature, and would benefit from the same guarantee as the fixed Q appearing in Theorem 1.2.

2In the biological community, IE borders are referred to as “acceptors” while EI borders are referred to as “donors.”
MONK – Outlier-Robust Mean Embedding Estimation by Median-of-Means

Figure 1: Performance of the MMD estimators: median and quartiles of \( \ln(\left| \hat{\text{MMD}}_{Q}(P, Q) - \text{MMD}(P, Q) \right|) \). Columns from left to right: Experiment-1 – Experiment-3. Top: quadratic kernel, bottom: RBF kernel.

(a) Gaussian distribution, \( N_c = 0 \) (no outlier), quadratic kernel. (b) Gaussian distribution, \( N_c = 5 \) outliers, quadratic kernel. (c) Pareto distribution, quadratic kernel. (d) Gaussian distribution, \( N_c = 0 \) (no outlier), RBF kernel. (e) Gaussian distribution, \( N_c = 5 \) outliers, RBF kernel. (f) Pareto distribution, RBF kernel.

Figure 2: Inter-class and intra-class MMD estimates as a function of the sample number compared to the bootstrap-estimated \( (1 - \alpha) \)-quantile: \( \hat{\text{MMD}} - \hat{q}_{1-\alpha} \); mean ± std. The null hypothesis is rejected iff \( \hat{\text{MMD}} - \hat{q}_{1-\alpha} > 0 \). Notice the different scale of \( \hat{\text{MMD}} - \hat{q}_{1-\alpha} \) in the inter-class and the intra-class experiments.

(a) Inter-class: EI-IE (b) Intra-class: EI-EI (c) Intra-class: IE-IE
References


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


Supplement

The supplement contains the detailed proofs of our results (Section A), a few technical lemmas used during these arguments (Section B), the McDiarmid inequality for self-containment (Section C), and the pseudocode of the two-sample test performed in Experiment-2 (Section D).

A. Proofs of Theorem 1 and Theorem 2

This section contains the detailed proofs of Theorem 1 (Section A.1) and Theorem 2 (Section A.2).

A.1. Proof of Theorem 1

The structure of the proof is as follows:

1. We show that \( \| \hat{\mu}_{\mathcal{P},Q} - \mu_{\mathcal{P}} \|_K \leq (1 + \sqrt{2}) r_{Q,N} \), where
   \[ r_{Q,N} = \sup_{f \in B_K} \text{MON}_Q \left[ \left\{ \langle f, K(\cdot,x) - \mu_{\mathcal{P}} \rangle \right\}_K \right] \], 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 an inner product space, for any \( f \in \mathcal{H}_K \)
\[
\| f - K(\cdot,x) \|_K^2 - \| \mu_{\mathcal{P}} - K(\cdot,x) \|_K^2 \\
= \| f - \mu_{\mathcal{P}} \|_K^2 - 2 \langle f - \mu_{\mathcal{P}}, K(\cdot,x) - \mu_{\mathcal{P}} \rangle_K. \tag{14}
\]
Hence, by denoting \( e = \hat{\mu}_{\mathcal{P},Q} - \mu_{\mathcal{P}}, \tilde{g} = g - \mu_{\mathcal{P}} \) we get
\[
\| e \|_K^2 - 2 r_{Q,N} \| e \|_K \\
\leq (a) \| e \|_K^2 - 2 \text{MON}_Q \left[ \left\{ \langle e, K(\cdot,x) - \mu_{\mathcal{P}} \rangle \right\}_K \right] \| e \|_K \\
\leq (b) \text{MON}_Q \left[ \| e \|_K^2 - 2 \left\langle e, K(\cdot,x) - \mu_{\mathcal{P}} \right\rangle_K \right] \| e \|_K \\
\leq (c) \text{MON}_Q \left[ \| \hat{\mu}_{\mathcal{P},Q} - K(\cdot,x) \|_K^2 - \| \mu_{\mathcal{P}} - K(\cdot,x) \|_K^2 \right] \\
\leq (d) \sup_{g \in \mathcal{H}_K} \text{MON}_Q \left[ \| \hat{\mu}_{\mathcal{P},Q} - K(\cdot,x) \|_K^2 - \| g - K(\cdot,x) \|_K^2 \right] \\
\leq (e) \sup_{g \in \mathcal{H}_K} \text{MON}_Q \left[ \| \mu_{\mathcal{P}} - K(\cdot,x) \|_K^2 - \| g - K(\cdot,x) \|_K^2 \right] \\
\leq (f) \sup_{g \in \mathcal{H}_K} \text{MON}_Q \left[ \| g, K(\cdot,x) - \mu_{\mathcal{P}} \|_K \right] - \| g \|_K^2 \\
\leq (g) \sup_{g \in \mathcal{H}_K} \left\{ 2 \left\| \| g \|_K r_{Q,N} - \| g \|_K^2 \right\| \right\} \leq r_{Q,N}^2, \tag{15}
\]
where we used in (a) the definition of \( r_{Q,N} \), (b) the linearity of \( \text{MON}_Q \), (c) Eq. (14), (d) \( \sup_{\theta} \), (e) the definition of \( \hat{\mu}_{\mathcal{P},Q} \), (f) Eq. (14) and the linearity of \( \text{MON}_Q \), (g) the definition of \( r_{Q,N} \). In step (h), by denoting \( a = \| g \|_K^2 \), \( r = r_{Q,N} \), 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 \( a^2 - 2ra - r^2 \) where \( a := \| \theta \|_K \geq 0 \). Hence \( r^2 + 2ra - a^2 \geq 0 \), \( r_{1,2} = \pm \sqrt{4a^2 + 4a^2} = \pm 2a \), \( \theta \), thus by the nonnegativity 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
\[
\| \hat{\mu}_{\mathcal{P},Q} - \mu_{\mathcal{P}} \|_K \leq (1 + \sqrt{2}) r_{Q,N}. \tag{16}
\]
It remains to upper bound \( r_{Q,N} \).

Step-2: Our goal is to provide a probabilistic bound on
\[
\begin{align*}
& r_{Q,N} = \sup_{f \in B_K} \text{MON}_Q \{ x \mapsto \langle f, K(\cdot,x) - \mu_{\mathcal{P}} \rangle_K \} \\
& = \sup_{f \in B_K} \text{med}_{q \in [Q]} \{ \langle f, \mu_{\mathcal{S}_q} - \mu_{\mathcal{P}} \rangle_K \}.
\end{align*}
\]
The \( N_c \) corrupted samples can affect (at most) \( N_c \) of the \( (S_q)_{q \in [Q]} \) blocks. Let \( U := [Q] \setminus C \) stand for the indices of the uncorrupted sets, \( C := \{ q \in [Q] : \exists n \in S_q, j \in [N_c] \} \) contains the indices of the corrupted sets. If
\[
\forall f \in B_K : \{ |q \in U : r(f,q) \geq \epsilon | \} + N_c \leq \frac{Q}{2}, \tag{17}
\]
then for \( \forall f \in B_K, \text{med}_{q \in [Q]} \{ r(f,q) \} \leq \epsilon \), i.e., \( \sup_{f \in B_K} \text{med}_{q \in [Q]} \{ r(f,q) \} \leq \epsilon \). Thus, our task boils down to controlling the event in (17) by appropriately choosing \( \epsilon \).

- Controlling \( r(f,q) \): For any \( f \in B_K \) the random variables \( (f, k(\cdot,x) - \mu_{\mathcal{P}})_{k \in K} = (f(x_i) - \mathbb{P} f) \) are independent, have zero mean, and
\[
\begin{align*}
\mathbb{E}_{\mathcal{S}_i \sim \mathbb{P}} \{ f(k(\cdot,x_i) - \mu_{\mathcal{P}}) \}_K & = \langle f, \Sigma_f \rangle_K \\
\leq & \| f \|_K \| \Sigma_f \|_K \leq \| f \|_K^2 \| \Sigma_f \| = \| \Sigma_f \| \leq \| \Sigma_f \|_K = 1.
\end{align*}
\]
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 \( P(\{ z > a \}) \leq \mathbb{P}(\{ |z| > a \}) \leq \mathbb{E} \left( z^2 \right) / a^2 \), which implies \( P(\{ z > \sqrt{\mathbb{E} \left( z^2 \right)} / \alpha \}) \leq \alpha \) by a \( \alpha = \sqrt{\mathbb{E} \left( z^2 \right) / a^2} \) substitution. With \( z := r(f,q) \) \( q \in U \), using \( \mathbb{E} \left( z^2 \right) = \mathbb{E} \left( \langle f, \mu_{\mathcal{S}_q} - \mu_{\mathcal{P}} \rangle_K \right)^2 = \frac{Q}{N_c} \mathbb{E}_{\mathcal{S}_i \sim \mathbb{P}} \{ f(k(\cdot,x_i) - \mu_{\mathcal{P}}) \}_K \) and Eq. (18) one gets that for all \( f \in B_K, \alpha \in (0,1) \) and \( q \in U \): \( P \left( r(f,q) > 2 \sqrt{\frac{Q}{N_c}} \right) \leq \alpha \). This means \( \mathbb{P} \left( r(f,q) > \frac{\epsilon}{2} \right) \leq \alpha \) with \( \epsilon \geq 2 \sqrt{\frac{Q}{N_c}} \).
• **Reduction to \( \phi \):** 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
\[
\sum_{q \in U} \mathbb{P} \left( r(f, q) \geq \frac{\epsilon}{2} \right) = |U| \alpha + \sum_{q \in U} \mathbb{P} \left( r(f, q) \geq \frac{\epsilon}{2} \right) =: A.
\]
Let us introduce \( \phi : t \in \mathbb{R} \rightarrow (t - 1) \mathbb{I}_{2 \leq t^2} + \mathbb{I}_{t \geq 2} \). \( \phi \) is 1-Lipschitz and satisfies \( \mathbb{I}_{2 \leq t^2} \leq \phi(t) \leq \mathbb{I}_{t \geq 2} \) for any \( t \in \mathbb{R} \). Hence, we can upper bound \( A \) as
\[
A \leq |U| \alpha + \sum_{q \in U} \mathbb{P} \left( r(f, q) \geq \frac{\epsilon}{2} \right) - \mathbb{E} \phi \left( r(f, q) \right) \]
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 \( \mathbb{I}_{2 \leq t^2} \leq \phi(t) \) and the \( \phi(t) \leq \mathbb{I}_{t \geq 2} \) bound, respectively. Taking supremum over \( B_K \) we arrive at
\[
\sup_{f \in B_K} \sum_{q \in U} \mathbb{I}_r(f, q) \geq \epsilon \leq |U| \alpha + \sup_{f \in B_K} \sum_{q \in U} \mathbb{P} \left( r(f, q) \geq \frac{\epsilon}{2} \right) - \mathbb{E} \phi \left( r(f, q) \right) =: Z.
\]

• **Concentration of \( Z \) around its mean:** Notice that \( Z \) is a function of \( x_V \), the samples in the uncorrupted blocks; \( V = \bigcup_{q \in U} S_q \). 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
\[
\mathbb{P} \left( Z < \mathbb{E}_{x_V}[Z] + Q\beta \right) \geq 1 - e^{-\frac{Qa^2}{2N}}.
\]

• **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 \( g_f : \mathbb{R}^M \rightarrow \mathbb{R} \) and \( \mathbb{P}_M := \frac{1}{M} \sum_{m=1}^{M} \delta_{u_m} \) defined as
\[
g_f(u_{1:M}) = \phi \left( \frac{f, \mu_{PM} - \mu_P}{\epsilon} \right)
\]
is uniformly bounded separable Carathéodory (Lemma 5), hence the symmetrization technique (Steinwart & Christmann, 2008, Prop. 7.10), (Ledoux & Talagrand, 1991) gives
\[
\mathbb{E}_{x_V}[Z] \leq 2\mathbb{E}_{x_V}[Z] \sum_{q \in U} \mathbb{E}_q \phi \left( \frac{2r(f, q)}{\epsilon} \right),
\]
where \( e = (e_q)_{q \in U} \in \mathbb{R}^{|U|} \) with i.i.d. Rademacher entries \( \mathbb{P}(e_q = \pm 1) = \frac{1}{2} \) \((\forall q)\).

• **Discarding \( \phi \):** Since \( \phi(0) = 0 \) and \( \phi \) is 1-Lipschitz, by Talagrand’s contraction principle of Rademacher processes (Ledoux & Talagrand, 1991), (Koltchinskii, 2011, Theorem 2.3) one gets
\[
\mathbb{E}_{x_V} \mathbb{E}_e \sup_{f \in B_K} \sum_{q \in U} e_q \phi \left( \frac{2r(f, q)}{\epsilon} \right) \leq 2\mathbb{E}_{x_V} \mathbb{E}_e \sum_{q \in U} e_q \frac{2r(f, q)}{\epsilon}.
\]

• **Switching from \( |U| \) to \( N \) terms:** Applying an other symmetrization [(a)], the CBS inequality, \( f \in B_K \), and the Jensen inequality
\[
\mathbb{E}_{x_V} \mathbb{E}_e \sup_{f \in B_K} \sum_{q=1}^{Q} e_q \frac{r(f, q)}{\epsilon} \leq \frac{2Q}{eN} \mathbb{E}_{x_V} \mathbb{E}_e' \left[ \sup_{f \in B_K} \sum_{n \in V} e'_n \left( f(x_n) - \mu_P \right)_K \right] = \frac{2Q}{eN} \mathbb{E}_{x_V} \mathbb{E}_e' \left[ \sum_{n \in V} e'_n \left( f(x_n) - \mu_P \right)_K \right] \leq \frac{2Q}{eN} \mathbb{E}_{x_V} \mathbb{E}_e' \left[ \sum_{n \in V} e'_n \left( f(x_n) - \mu_P \right)_K \right]^2 \leq 2Q \sqrt{V} \text{Tr}(\Sigma_P) / eN.
\]

In (a), we proceed as follows:
\[
\mathbb{E}_{x_V} \mathbb{E}_e \sup_{f \in B_K} \sum_{q \in U} e_q \frac{r(f, q)}{\epsilon} = \mathbb{E}_{x_V} \mathbb{E}_e \sup_{f \in B_K} \sum_{q \in U} e_q \frac{f, \mu_S - \mu_P}_K \leq \frac{2Q}{eN} \mathbb{E}_{x_V} \mathbb{E}_e' \sup_{f \in B_K} \sum_{n \in V} e'_n \left( f(x_n) - \mu_P \right)_K \leq \frac{2Q}{eN} \mathbb{E}_{x_V} \mathbb{E}_e' \sup_{f \in B_K} \sum_{n \in V} e'_n \left( f(x_n) - \mu_P \right)_K,
\]
where in (c) we applied symmetrization, \( e' = (e'_n)_{n \in V} \in \mathbb{R}^{|V|} \) 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 (f, K(\cdot, x_n) - \mu_F)_{K_{n\in V}} =
\sum_{n\in V} e'_n (K(\cdot, x_n) - \mu_F)_{K_{n\in V}}.
In step (b), we had
\[ E_{\epsilon V} E_{\epsilon'} \left\| \sum_{n\in V} e'_n (K(\cdot, x_n) - \mu_F) \right\|^2_{K_{n\in V}} \]
\[ = \sum_{n\in V} e'_n (K(\cdot, x_n) - \mu_F, K(\cdot, x_n) - \mu_F)_{K_{n\in V}} \]
\[ = \left| V \right| \sum_{n\in V} \left| e'_n \right|^2 (K(\cdot, x_n) - \mu_F, K(\cdot, x_n) - \mu_F)_{K_{n\in V}} \]
\[ = \left| V \right| \sum_{n\in V} \left| e'_n \right|^2 (K(\cdot, x_n) - \mu_F, K(\cdot, x_n) - \mu_F)_{K_{n\in V}} \]
\[ = \left| V \right| \sum_{n\in V} \left| e'_n \right|^2 (K(\cdot, x_n) - \mu_F, K(\cdot, x_n) - \mu_F)_{K_{n\in V}} \]
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
$\epsilon \geq 2 \sqrt{\frac{3\delta^2}{2\alpha N}}$, then
$\sup_{f \in B_K} \frac{\sum_{q=1}^Q \| I_q(f) \|_{\geq \epsilon}}{\alpha + \beta + \frac{8Q\sqrt{\left| V \right| \operatorname{Tr}(\Sigma_p)}}{\epsilon N}}$
with probability at least $1 - \epsilon^{-\frac{Q^2}{\beta}}$. Thus, to ensure that
$\sup_{f \in B_K} \sum_{q=1}^Q \| I_q(f) \|_{\geq \epsilon} + N_c \leq Q/2$ it is sufficient to
choose $\left( \alpha, \beta, \epsilon \right)$ such that $\left( \alpha + \beta + \frac{8Q\sqrt{\left| V \right| \operatorname{Tr}(\Sigma_p)}}{\epsilon N} + N_c \leq Q/2$, and in this case $\| \mu_F, Q - \mu_F \|_{K_{n\in V}} \leq (1 + \sqrt{2})\epsilon$. Applying the $\left| U \right| \leq Q$ and $\left| V \right| \leq N$ bounds, we want to have
$\alpha + \beta + \frac{8Q\sqrt{\left| V \right| \operatorname{Tr}(\Sigma_p)}}{\epsilon N} + N_c \leq Q/2$. (19)
Choosing $\alpha = \beta = \frac{4}{3}$ in Eq. (19), the sum of the first two terms
is $Q^{\frac{2\delta}{3}}\epsilon; \epsilon \geq \max \left( 2 \sqrt{\frac{3\delta^2}{2\alpha N}}, \frac{24}{3} \sqrt{\frac{\operatorname{Tr}(\Sigma_p)}{N}} \right)$ gives
$\leq Q^{\frac{2\delta}{3}}$ for the third term. Since $N_c \leq Q\left( \frac{1}{2} - \delta \right)$, we got
$\| \mu_F, Q - \mu_F \|_{K} \leq c_1 \max \left( \sqrt{\frac{3\delta^2}{2\alpha N}}, \frac{12}{\delta} \sqrt{\frac{\operatorname{Tr}(\Sigma_p)}{N}} \right)$
with probability at least $1 - \epsilon^{-\frac{Q^2}{\beta}}$. With an $\eta = \epsilon^{-\frac{Q^2}{\beta}}$, and
hence $Q = \frac{72\ln(\frac{1}{\delta})}{\beta}$ 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 $\left\| \text{MMD}_Q(P, Q) - \text{MMD}(P, Q) \right\| \leq r_{Q,N}$, where $r_{Q,N} = \sup_{f \in B_K} \left( \text{med}_{q \in Q} \left\{ \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K \right\} \right)$.
- Then $r_{Q,N}$ is bounded.

**Step-1:**

- $\text{MMD}_Q(P, Q) - \text{MMD}(P, Q) \leq r_{Q,N}$: By the subadditivity of supremum $\left( \sup_{f \in B_K} a_f + b_f \leq \sup_{f \in B_K} a_f + \sup_{f \in B_K} b_f \right)$ one gets
  \[ \text{MMD}_Q(P, Q) = \sup_{f \in B_K} \left\{ \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K \right\} \]
  \[ + \sup_{f \in B_K} \left\{ \langle f, (\mu_P - \mu_Q) \rangle_K \right\} \]
  \[ \leq \sup_{f \in B_K} \left\{ \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K \right\} \]
  \[ + \sup_{f \in B_K} \left\{ \langle f, (\mu_P - \mu_Q) \rangle_K \right\} \]
  \[ \leq \sup_{f \in B_K} \left\{ \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K \right\} \]
  \[ = r_{Q,N} \]

- $\text{MMD}_Q(P, Q) \leq \text{MMD}(P, Q)$:
  Let $a_f := \langle f, (\mu_P - \mu_Q) \rangle_K$ and $b_f := \sup_{q \in \hat{Q}} \left\{ \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K \right\}$. Then
  \[ a_f - b_f := \langle f, (\mu_P - \mu_Q) \rangle_K \]
  \[ + \sup_{q \in \hat{Q}} \left\{ \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K \right\} \]
  \[ = \text{med}_{q \in \hat{Q}} \left\{ \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K \right\} \]
  by $\text{med}_{q \in \hat{Q}} \left\{ -z_q \right\} = -\text{med}_{q \in \hat{Q}} \left\{ z_q \right\}$. Applying the $\sup_{f \in B_K} a_f - b_f \geq \sup_{f \in B_K} a_f - \sup_{f \in B_K} b_f$ inequality (it follows from the subadditivity of sup):
  \[ \text{MMD}_Q(P, Q) \geq \text{MMD}(P, Q) \]
  \[ \geq \sup_{f \in B_K} \left\{ \langle f, (\mu_P - \mu_Q) \rangle_K \right\} \]
  \[ - \sup_{f \in B_K} \left\{ \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K \right\} \]
  \[ = \text{MMD}_Q(P, Q) \]

**Step-2:** Our goal is to control
\[ r_{Q,N} = \sup_{f \in B_K} \left\{ \text{med}_{q \in \hat{Q}} \left\{ r(f, q) \right\} \right\} \]
where $r(f, q) := \langle f, (\mu_{S_q,P} - \mu_{S_q,Q}) - (\mu_P - \mu_Q) \rangle_K$.

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

\[
\sup_{f \in B_K} \left| \text{med} \{ r(f, q) \} \right| \leq \epsilon
\]
\[
\Leftrightarrow \forall f \in B_K : -\epsilon \leq \text{med}_{q \in [Q]} \{ r(f, q) \} \leq \epsilon
\]
\[
\Leftrightarrow \forall f \in B_K : \{ q : r(f, q) \leq -\epsilon \} \leq Q/2
\]
and \[
\{ q : r(f, q) \geq \epsilon \} \leq Q/2
\]
\[
\Leftrightarrow \forall f \in B_K : \{ q : r(f, q) \geq \epsilon \} \leq Q/2.
\]
Thus, \( \forall f \in B_K : \{ q \in U : |r(f, q)| \geq \epsilon \} + N_c \leq Q/2. \)

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

• Controlling \( |r(f, q)| \): For any \( f \in B_K \) the random variables \( |f(x_i) - f(y_i)| \) are independent, zero-mean and

\[
\mathbb{E}_{(x,y) \sim \mathbb{P} \otimes \mathbb{Q}} \left[ (|f(x) - f(y)| - |f(y) - Qf|)^2 \right]
\]
\[
= \mathbb{E}_{x \sim \mathbb{P}} |f(x)|^2 + \mathbb{E}_{y \sim \mathbb{Q}} |f(y)|^2 - 2 \mathbb{E}_{x \sim \mathbb{P}} f(x) \mathbb{E}_{y \sim \mathbb{Q}} f(y)
\]
\[
\leq \| \Sigma_{\mathbb{P}} \| + \| \Sigma_{\mathbb{Q}} \|,
\]
where \( \mathbb{P} \otimes \mathbb{Q} \) is the product measure. The Chebyshev argument with \( z = |r(f, q)| \) implies that \( \forall \alpha \in (0, 1) \)

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

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

• 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

\[
\mathbb{E}_{(x,y) \sim \mathbb{P} \otimes \mathbb{Q}} \left[ \sum_{n \in V} e_n^2 (\tilde{x}_n - \tilde{y}_n) \right]_K
\]
\[
= \mathbb{E}_{(x,y) \sim \mathbb{P} \otimes \mathbb{Q}} \left[ \sum_{n \in V} e_n^2 (\tilde{x}_n - \tilde{y}_n) \right]_K
\]
\[
= |V| \mathbb{E}_{(x,y) \sim \mathbb{P} \otimes \mathbb{Q}} \left[ |K(\cdot, x) - \mu_{\mathbb{P}}| - |K(\cdot, y) - \mu_{\mathbb{Q}}| \right]_K
\]
\[
= |V| \left[ \text{Tr}(\Sigma_{\mathbb{P}}) + \text{Tr}(\Sigma_{\mathbb{Q}}) \right].
\]

These results imply

\[
Q\alpha + Q\beta + \frac{8Q\sqrt{\text{Tr}(\Sigma_{\mathbb{P}}) + \text{Tr}(\Sigma_{\mathbb{Q}})}}{\epsilon \sqrt{N}} + N_c \leq Q/2.
\]

with probability at least \( 1 - e^{-Q/2}. \)

\[
\eta = e^{-Q/2}, \text{i.e.}\]

\[
Q = \frac{72\ln(\frac{1}{\delta})}{\epsilon^2}\]
reparameterization finishes the proof of Theorem 2.

B. Technical Lemmas

Lemma 3 (Supremum).

\[
\sup_{f} |a_f - b_f| \leq \sup_{f} |a_f|.
\]

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

\[
\sup_{x_V \in \mathcal{X}^V} \sup_{x'_V \in \mathcal{X}} |Z(x_V) - Z(x'_V)| \leq 4, \forall i \in V.
\]

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

\[
q_r := q_r(f, x_V),
\]
\[
q \in U = \phi \left[ \frac{2(f, \mu_S - \mu_K)}{\epsilon} \right] - \mathbb{E}_{\phi} \left[ \frac{2(f, \mu_S - \mu_K)}{\epsilon} \right],
\]
\[
y_i := y_i(f, x_V).
\]

In this case

\[
|Z(x_V) - Z(x'_V)| = \sup_{f \in B_K} \sum_{q \in U} y_q - \sup_{f \in B_K} \sum_{q \in U \setminus \{r\}} y_q + y'_r
\]
\[
\leq \sum_{f \in B_K} |y_r - y'_r| \leq \sum_{f \in B_K} \left| \frac{y_i + y'_i}{2} \right| \leq 4,
\]

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

\[
\square
\]

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

\[
g_f(u_{1:M}) = \phi \left[ \frac{2(f, \mu_{\mathcal{X}^M} - \mu_{\mathbb{P}})}{\epsilon} \right].
\]
Then $\mathcal{F}$ is a uniformly bounded separable Carathéodory family: (i) $\sup_{f \in B_K} \|f\|_\infty < \infty$ 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 $\mathcal{H}_K \subset C(\mathcal{X}) = \{h : \mathcal{X} \to \mathbb{R}$ continuous$, \text{so } u_{1:M} \mapsto (f(u_1), \ldots, f(u_M)) \text{ is continuous. } \phi \text{ 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 $f \mapsto (f, h)_K(h = \mu_{\mathcal{P}_1} - \mu_{\mathcal{P}_2})$ and that of $\phi$.
(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

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

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

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

D. Pseudocode of Experiment-2

The pseudocode of the two-sample test conducted in Experiment-2 is summarized in Algorithm 3.

**Algorithm 3 Two-sample test (Experiment-2)**

**Input:** Two samples: $(X_n)_{n \in [N]}, (Y_n)_{n \in [N]}$. Number of bootstrap permutations: $B \in \mathbb{Z}^+$. Level of the test: $\alpha \in (0, 1)$. Kernel function with hyperparameter $\theta \in \Theta$: $K_\theta$.

Split the dataset randomly into 3 equal parts:

$$[N] = \bigcup_{i=1}^3 I_i, \quad |I_1| = |I_2| = |I_3|.$$  

Tune the hyperparameters using the 1st part of the dataset:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} J_\theta := \overline{\text{MMD}}_\theta((X_n)_{n \in I_1}, (Y_n)_{n \in I_1}).$$

Estimate the $(1 - \alpha)$-quantile of $\overline{\text{MMD}}_\hat{\theta}$ under the null, using $B$ bootstrap permutations from $(X_n)_{n \in I_2} \cup (Y_n)_{n \in I_2}$: $\hat{q}_{1-\alpha}$.

Compute the test statistic on the third part of the dataset:

$$T_{\hat{\theta}} = \overline{\text{MMD}}_\hat{\theta}((X_n)_{n \in I_3}, (Y_n)_{n \in I_3}).$$

**Output:** $T_{\hat{\theta}} - \hat{q}_{1-\alpha}$. 
