Keep it Tighter – A Story on Analytical Mean Embeddings

Linda Chamakh
Model Risk Governance and Review
JP Morgan Chase
London, UK

Zoltán Szabó
Department of Statistics
London School of Economics
London, UK

Abstract

Kernel techniques are among the most popular and flexible approaches in data science allowing to represent probability measures without loss of information under mild conditions. The resulting mapping called mean embedding gives rise to a divergence measure referred to as maximum mean discrepancy (MMD) with existing quadratic-time estimators (w.r.t. the sample size) and known convergence properties for bounded kernels. In this paper we focus on the problem of MMD estimation when the mean embedding of one of the underlying distributions is available analytically. Particularly, we consider distributions on the real line (motivated by financial applications) and prove tighter concentration for the proposed estimator under this semi-explicit setting; we also extend the result to the case of unbounded (exponential) kernel with minimax-optimal lower bounds. We demonstrate the efficiency of our approach beyond synthetic example in three real-world examples relying on one-dimensional random variables: index replication and calibration on loss-given-default ratios and on S&P 500 data.

1 INTRODUCTION

Kernel methods (Steinwart and Christmann, 2008; Paulsen and Raghupathi, 2016) form one of the most powerful tools in machine learning and statistics with a wide range of successful applications. The impressive modelling power and flexibility of kernel techniques in capturing complex nonlinear relations originates from the richness of the underlying function class called reproducing kernel Hilbert space (Aronszajn, 1950, RKHS) associated to a kernel.

Kernel functions can be used to capture the similarity of objects belonging to various domains including sequences (Király and Oberhauser, 2019), sets (Haussler, 1999), and graphs (Borgwardt et al., 2020). Having a notion of inner product realized by kernels, one can represent probability distributions on any kernel-endowed domain via mean embeddings (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007), which specifically allows to quantify the divergence between distributions by considering the RKHS distance between their corresponding mean embeddings. The resulting (semi-)metric called maximum mean discrepancy (Smola et al., 2007; Gretton et al., 2012, MMD) forms one of the most popular divergence measures in machine learning; the equivalent (Sejdinovic et al., 2013) notion in the statistic community is referred to as energy distance (Baringhaus and Franz, 2004; Székely and Rizzo, 2004, 2005) or N-distance (Zinger et al., 1992; Klebanov, 2005).

The wide popularity of MMD stems from (i) the computational tractablity of its different estimators, (ii) the existence of closed-form expressions for MMD in case of certain kernel-distribution pairs, (iii) its theoretical guarantees facilitated by the underlying Hilbert structure of RKHSs including concentration properties for bounded kernels (Gretton et al., 2012) and (iv) MMD being a metric for characteristic kernels (Fukumizu et al., 2008; Sriperumbudur et al., 2010; Szabó and Sriperumbudur, 2018). These favorable properties of MMD have given rise to various successful applications, including for instance two-sample testing (Gretton et al., 2012; Schrab et al., 2022; Hagrass et al., 2024), independence (Gretton et al., 2008; Deb et al., 2020; Albert et al., 2022), goodness-of-fit testing (Balasubramanian et al., 2021; Baum et al., 2023), and statistical inference (Briol et al., 2019a; Alquier and Gerber, 2024), among many others.

In statistical modelling, the problem of parametric estimation (Casella and Berger, 2024)—which aims to find the optimal parametric distribution from a spec-

This project was initiated during the main author's PhD and was not pursued as part of their work at JP Morgan Chase. It is an independent, extra-curricular activity and does not involve or engage the company in any capacity.

ified family, given a set of observations—is one of the most fundamental tasks. The problem can be tackled by minimizing a divergence (also referred to as a calibration metric) between the target parametric distribution and the empirical distribution associated to the data, with MMD as a natural choice. In particular, p-value and acceptance region derived by Gretton et al. (2012) can be used to assess the calibration quality and the adequacy of the chosen distribution family. Moreover, for certain kernels and parametric distributions, the mean embedding can be computed analytically (Briol et al., 2019b, Table 1). Our primary aim in this paper from theoretical perspective is to understand the benefits of such analytical knowledge (when available): we propose a new semi-explicit (one-sample) MMD estimator and prove its tightened convergence guarantees compared to its classical two-sample counterpart.

From practical angle, in finance, parametric distributions are widely applied for modelling, simulation and interpretation purposes. Common distributions arising in finance include (i) the beta distribution with a bounded support in $(0, 1)$ which is particularly well-suited to model financial ratios such as loan recovery rates (Chen and Wang, 2013) and (ii) the Gaussian distribution which is often used to model the distribution of the log-returns of stock values starting from the seminal work of Black and Scholes (1973). Distributions with non-zero skewness (the normalised third moment of the distribution) are also relevant since financial returns can divert from the Gaussian distribution by exhibiting fat tails and negative skewness (Cont, 2001). Relaxation of the Gaussian distribution in these directions, such as the skew-normal distribution, turned out (Christopher Adcock and Loperfido, 2015) to play a key role in the area.

Two key financial applications with the aforementioned distributions and our motivation from practical point of view are as follows.

1. The index replication problem consists of finding the weighted average of individual stocks matching a return distribution (Bamberg and Wagner, 2000; Roncalli and Weisang, 2009). As proposed by Chalabi and Wuertz (2012); Lassance (2019), the problem can be solved by minimizing the divergence between the distribution of weighted stocks and that of of index returns. When performing such replication, Gaussian distribution and its relaxations constitute a natural choice for the distribution of returns (Black and Scholes, 1973; Cont, 2001).
2. The modelling of the loss-given-default (LGD; which represents the percentage of the loan the client or company is not able to repay given he has defaulted) can be assumed to follow a beta distribution, as advised by the financial agency Moody's in their widely-used recovery model methodology (Gupton and Stein, 2002).

In both of these applications and throughout the paper we focus on distributions on the real line. In addition we note that further parametric estimation problems in finance arise with processes driven by a stochastic differential equation (Bishwal, 2007), quantile estimation, and tail dependence modelling (Jadhav and Ramanathan, 2009; Fortin and Kuzmics, 2002).

Our **contributions** can be summarized are as follows.

1. We propose the semi-explicit MMD estimator (relying on analytical mean embedding when available), prove its tighter concentration properties (Theorem 2) compared to its two-sample counterpart (established for bounded kernels) and extend the analysis to the unbounded exponential kernel (Theorem 3) with matching minimax lower bound (Theorem 4).
2. Accompanying our tighter concentration analysis, we derive the analytical mean embedding for new kernel-distribution pairs motivated by financial applications, covering the (Gaussian exponentiated, Gaussian) and (Matérn, beta) pairs.
3. We demonstrate the efficiency of our MMD estimator in three applications: a synthetic example, index replication, and calibration on LGD ratios and on S&P 500 data.

The paper is structured as follows. Notations are introduced in Section 2. Section 3 is dedicated to existing and the proposed semi-explicit MMD estimator. Our theoretical results are presented in Section 4. Numerical illustrations form the focus of Section 5. Proofs are deferred to the supplement.

2 NOTATIONS

This section is dedicated to definitions and to the introduction of our quantities of interest: mean embedding, maximum mean discrepancy, and our choice of studied kernel and distributions. We introduce the **notations**: $\mathbb{N}$, $\mathbb{N}^*$, $[N]$, $\mathbb{R}^{>0}$, $\mathbb{R}^{\geq 0}$, $\mathbf{v}^\top$, $\text{Diag}(\mathbf{v})$, $a \wedge b$, $a \vee b$, $\mathbb{I}_A$, $L_{(s)}$, $\mathcal{O}(\cdot)$, $o(\cdot)$, $\mathcal{O}_{a.s}(\cdot)$, $o_{a.s}(\cdot)$, $\mathcal{W}^d$, $\zeta_\mathbb{Q}$, Φ, B, $\mathcal{H}_k$, φ_k, $\mathcal{B}_k$, $\mathcal{M}_1^+(\mathcal{X})$, μ_k, MMD_k.

Natural numbers are denoted by $\mathbb{N} = \{0, 1, \ldots\}$; $\mathbb{N}^* = \{1, 2, \ldots\}$ stands for the set of positive integers. For $N \in \mathbb{N}^*$, $[N] = \{1, \ldots, N\}$. Positive reals are denoted by $\mathbb{R}^{>0}$; $\mathbb{R}^{\geq 0}$ stands for non-negative reals. The transpose of a vector $\mathbf{v} \in \mathbb{R}^d$ is denoted by $\mathbf{v}^\top$; the diagonal matrix formed of a vector $\mathbf{v} \in \mathbb{R}^d$ is given by $\text{Diag}(\mathbf{v}) \in \mathbb{R}^{d\times d}$. The minimum of two numbers $a, b \in \mathbb{R}$ is denoted by $a \wedge b$; their maximum is $a \vee b$. For a set A, $\mathbb{I}_A$ is the indicator of A: $\mathbb{I}_A(x) = 1$ if $x \in A$; $\mathbb{I}_A(x) = 0$ otherwise. Given $(L_s)_{s\in[S]} \subset \mathbb{R}$, the associated order statistics are $L_{(1)} \leq \ldots \leq L_{(S)}$. The notation $b_n = \mathcal{O}(a_n)$ (resp. $b_n = o(a_n)$) means that $(\frac{b_n}{a_n})_{n\in\mathbb{N}}$ is bounded (resp. $\lim_{n\to\infty} \frac{b_n}{a_n} = 0$). For random variables $X_n = \mathcal{O}_{a.s}(a_n)$ (resp. $X_n = o_{a.s}(a_n)$) means that $\left(\frac{X_n}{a_n}\right)_{n\in\mathbb{N}}$ is bounded (resp. converges to zero) almost surely. The $(d-1)$-dimensional simplex is $\mathcal{W}^d = \left\{\mathbf{w} \in \left(\mathbb{R}^{\geq 0}\right)^d : \sum_{j=1}^d w_j = 1\right\}$. Let $m_{\mathbb{Q}}$ and $\sigma_{\mathbb{Q}}$ denote the expectation and the standard deviation of a real-valued random variable with distribution $\mathbb{Q}$; its skewness is defined as the standardized third moment $\zeta_{\mathbb{Q}} = \mathbb{E}_{x\sim\mathbb{Q}}\left[\left((x-m_{\mathbb{Q}})/\sigma_{\mathbb{Q}}\right)^3\right]$. The cumulative density function (cdf) of the standard normal distribution is Φ; $\Phi(x) = \int_{-\infty}^x e^{-\frac{t^2}{2}}\mathrm{d}t$ $(x \in \mathbb{R})$. The beta function for $\alpha, \beta \in \mathbb{R}^{>0}$ is defined as $B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t$.

In the sequel, let $\mathcal{X}$ denote a (non-empty) subset of the real line ($\mathcal{X} \subseteq \mathbb{R}$). A function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is called kernel if there exists a feature map φ from $\mathcal{X}$ to a Hilbert space $\mathcal{H}$ such that $k(x,y) = \langle\varphi(x),\varphi(y)\rangle_{\mathcal{H}}$ for all $x, y \in \mathcal{X}$. While the feature map and the Hilbert space might not be unique, there always exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ associated to k. $\mathcal{H}_k$ is the Hilbert space of $\mathcal{X} \to \mathbb{R}$ functions characterized by two properties: $k(x,\cdot) \in \mathcal{H}_k$ $(\forall x \in \mathcal{X})$ and $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}_k}$ $(\forall x \in \mathcal{X},\ f \in \mathcal{H}_k)$.[1] The first property describes the basic elements of $\mathcal{H}_k$, the second one is called the reproducing property; combining the two properties makes the canonical feature map and feature space explicit: $k(x,y) = \langle\varphi_k(x),\varphi_k(y)\rangle_{\mathcal{H}_k}$, where $\varphi_k(x) = k(\cdot,x) \in \mathcal{H}_k$. The closed unit ball of $\mathcal{H}_k$ is denoted by $\mathcal{B}_k = \left\{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\right\}$.

Let $\mathcal{M}_1^+(\mathcal{X})$ denote the set of Borel probability measures on $\mathcal{X}$. For a given kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, let

$$\mu_k(\mathbb{P}) = \int_{\mathcal{X}} k(\cdot,x)\mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k$$

denote the mean embedding (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007) of the probability distribution $\mathbb{P} \in \mathcal{M}_1^+(\mathcal{X})$; the integral is meant in Bochner sense. The mean embedding is well-defined when $(\mathbb{P}, k)$ satisfies

$$\mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)} < \infty. \qquad (\mathcal{P}_k)$$

The maximum mean discrepancy (MMD) of two distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$ is a semi-metric defined by

$$\begin{aligned}
&\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} \\
&= \sup_{f\in\mathcal{B}_k}\left[\mathbb{E}_{x\sim\mathbb{P}}f(x) - \mathbb{E}_{y\sim\mathbb{Q}}f(y)\right] \\
&= \sqrt{\|\mu_K(\mathbb{P})\|^2_{\mathcal{H}_K} + \|\mu_K(\mathbb{Q})\|^2_{\mathcal{H}_K} - 2\left\langle\mu_K(\mathbb{P}),\mu_K(\mathbb{Q})\right\rangle_{\mathcal{H}_K}} \\
&= \sqrt{\mathbb{E}_{\substack{x\sim\mathbb{P},\\ x'\sim\mathbb{P}}}k(x,x') + \mathbb{E}_{\substack{y\sim\mathbb{Q},\\ y'\sim\mathbb{Q}}}k(y,y') - 2\mathbb{E}_{\substack{x\sim\mathbb{P},\\ y\sim\mathbb{Q}}}k(x,y)},
\end{aligned}$$

where the second form ($\sup_{f\in\mathcal{B}_k}$) encodes that the discrepancy of two probability distributions is measured by their maximal mean discrepancy over $\mathcal{B}_k$. It also shows that MMD belongs to the class of integral probability metrics (Zolotarev, 1983; Müller, 1997). MMD is well-defined when the pairs $(\mathbb{P}, k)$ and $(\mathbb{Q}, k)$ satisfy $(\mathcal{P}_k)$; this automatically holds for bounded kernels ($\sup_{x\in\mathcal{X}} k(x,x) < \infty$). MMD is a metric if and only if the kernel is characteristic (Fukumizu et al., 2008; Sriperumbudur et al., 2010); examples of characteristic kernels include the Gaussian, Laplacian, Matérn, inverse multiquadrics or the B-spline kernel.

3 MMD ESTIMATORS

In this section, we recall existing two-sample MMD estimators in Section 3.1, and we present our proposed semi-explicit ones in Section 3.2. Our motivation for the new estimators is two-fold: (i) to reduce the computational time, and (ii) to achieve tighter concentration.

3.1 Classical MMD Estimator

Given i.i.d. (independent identically distributed) samples $\{x_i\}_{i\in[N]} \sim \mathbb{P}$ and $\{y_i\}_{i\in[M]} \sim \mathbb{Q}$ from the probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathcal{X})$, one can estimate the squared MMD by using the unbiased U-statistics or the plug-in V-statistics as

$$\widehat{\mathrm{MMD}}^2_{k,U}(\mathbb{P}_N,\mathbb{Q}_M) = \frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\neq j}} k(x_i,x_j) \quad (1)$$

$$+\frac{1}{M(M-1)}\sum_{\substack{i,j\in[M]\\ i\neq j}} k(y_i,y_j) - \frac{2}{NM}\sum_{\substack{i\in[N]\\ j\in[M]}} k(x_i,y_j),$$

$$\widehat{\mathrm{MMD}}^2_{k,V}(\mathbb{P}_N,\mathbb{Q}_M) = \frac{1}{N^2}\sum_{i,j\in[N]} k(x_i,x_j) \quad (2)$$

$$+\frac{1}{M^2}\sum_{i,j\in[M]} k(y_i,y_j) - \frac{2}{NM}\sum_{i\in[N]}\sum_{j\in[M]} k(x_i,y_j),$$

where $\mathbb{P}_N = \frac{1}{N}\sum_{n\in[N]}\delta_{x_n}$ and $\mathbb{Q}_M = \frac{1}{M}\sum_{m\in[M]}\delta_{y_m}$ denote the empirical measures. The estimator $\widehat{\mathrm{MMD}}^2_{k,U}(\mathbb{P}_N,\mathbb{Q}_M)$ is unbiased, $\widehat{\mathrm{MMD}}^2_{k,V}(\mathbb{P}_N,\mathbb{Q}_M)$ is non-negative, hence they have complementary advantages; both estimators have computational complexity $\mathcal{O}\left((N+M)^2\right)$.

3.2 Proposed Semi-Explicit MMD Estimator

If the mean embedding $\mu_K(\mathbb{Q}) = \mathbb{E}_{y\sim\mathbb{Q}}k(\cdot,y)$ can be computed analytically, one can alternatively estimate the squared MMD using the plugin idea of (1), or that

[1] The shorthand $k(\cdot,x)$ stands for the function $y \in \mathcal{X} \mapsto k(y,x) \in \mathbb{R}$ while keeping $x \in \mathcal{X}$ fixed.

of (2) as

$$\widehat{\text{MMD}}^2_{k,e,U}(\mathbb{P}_N,\mathbb{Q}) = \frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\neq j}} k(x_i,x_j) + \mathbb{E}_{y\sim\mathbb{Q}}\mu_k(\mathbb{Q})(y) - \frac{2}{N}\sum_{i\in[N]}\mu_k(\mathbb{Q})(x_i), \quad (3)$$

$$\widehat{\text{MMD}}^2_{k,e,V}(\mathbb{P}_N,\mathbb{Q}) = \frac{1}{N^2}\sum_{i,j\in[N]} k(x_i,x_j) + \mathbb{E}_{y\sim\mathbb{Q}}\mu_k(\mathbb{Q})(y) - \frac{2}{N}\sum_{i\in[N]}\mu_k(\mathbb{Q})(x_i). \quad (4)$$

We will refer to (3) and (4) as the semi-explicit MMD estimators; both have computational complexity $\mathcal{O}\left(N^2\right)$.

4 RESULTS

In this section, we show the theoretical advantage of using explicit mean embedding when available. In Section 4.1 we prove tightened concentration results for our semi-explicit MMD estimators, and extend the analysis to unbounded kernels, with matching minimax lower bounds. We summarize in Section 4.2 the kernel-distribution pairs for which we derived analytical mean embeddings, extending the current literature.

4.1 Concentration of Semi-Explicit MMD

In this section, we show that explicit mean embedding, in case of both bounded and unbounded kernels, leads to better concentration properties of the MMD estimator. We start by recalling the concentration of the classical U-statistic based MMD estimator for bounded kernels (Theorem 1), followed by presenting our tighter result for the semi-explicit MMD (Theorem 2), which we also extend to the unbounded exponential kernel (Theorem 3) with matching minimax lower bound (Theorem 4).

Theorem 1 (MMD concentration - bounded kernel)**.** *Assume that $0 \le k(x,x') \le B$ for all $x,x' \in \mathcal{X}$, and let $\epsilon > 0$. Then*

$$\mathbb{P}\left(\widehat{\text{MMD}}^2_{k,U}(\mathbb{P},\mathbb{Q}) - \text{MMD}^2_k(\mathbb{P},\mathbb{Q}) > \varepsilon\right) \le e^{-\frac{\left\lfloor \frac{N}{2}\right\rfloor \varepsilon^2}{8B^2}}.$$

The same bound holds for the deviation of $-\varepsilon$ below.

Using the analytical knowledge of $\mu_K(\mathbb{Q})$ leads to tighter concentration as it is shown by our next result. We recall that $\mathcal{X} \subseteq \mathbb{R}$ ($\mathcal{X} \neq \emptyset$) throughout the manuscript.

Theorem 2 (Semi-explicit MMD concentration - bounded kernel)**.** *Assume that $A \le k(x,x') \le B$ for all $x,x' \in \mathcal{X}$, and let $\epsilon > 0$. Then*

$$\mathbb{P}\left(\widehat{\text{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \text{MMD}^2_k(\mathbb{P},\mathbb{Q}) > \varepsilon\right) \le e^{-\frac{\left\lfloor \frac{N}{2}\right\rfloor \varepsilon^2}{2(B-A)^2}} + e^{-\frac{N\varepsilon^2}{8(B-A)^2}}.$$

The same bound holds for the deviation of $-\varepsilon$ below.

Remarks:

- The proof of Theorem 2 relies on rewriting the difference $\widehat{\text{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \text{MMD}^2_k(\mathbb{P},\mathbb{Q})$ as a sum of two U-statistics of different orders, followed by applying twice the Hoeffding inequality for U-statistics and union bounding.
- Specializing Theorem 2 to $A = 0$ and comparing its concentration result with that in Theorem 1 for $\widehat{\text{MMD}}^2_{k,U}(\mathbb{P},\mathbb{Q})$, we gain in terms of constant in front of $N\varepsilon^2$ in the exponent: we have $\frac{1}{4B^2}$ and $\frac{1}{8B^2}$ instead of $\frac{1}{16B^2}$. This means that the estimator using the analytical knowledge of $\mu_K(\mathbb{Q})$ brings a factor of 2 improvement in the *exponent*.

In Theorem 2 the deviation of the estimator $\widehat{\text{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})$ was captured for bounded kernels. Our next theorem extends the result to the unbounded exponential kernel, a subcase of the Gaussian-exponentiated kernel $(x,y) \mapsto e^{-a(x-y)^2+bxy}$ with $a = 0$.

Theorem 3 (Semi-explicit MMD concentration - exponential kernel)**.** *Let us consider the exponential kernel $k(x,y) = e^{bxy}$ ($b > 0$, $x,y \in \mathbb{R}$) with probability measures $\mathbb{P},\mathbb{Q} \in \mathcal{M}^+_1(\mathbb{R})$ satisfying*

$$\mathbb{E}_{x\sim\mathbb{P}}e^{\lambda x^2} < \infty, \quad \mathbb{E}_{x\sim\mathbb{Q}}e^{\lambda x^2} < \infty \quad \forall\lambda \in \mathbb{R}. \quad (5)$$

Let the number of samples N taken from $\mathbb{P}$ be even. Then for any $p \ge 2$, there exists a universal constant $C = C_{p,\mathbb{P},K} > 0$ such that for any $\varepsilon > 0$

$$\mathbb{P}\left(\widehat{\text{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \text{MMD}^2_k(\mathbb{P},\mathbb{Q}) > \varepsilon\right) \le \frac{C}{\varepsilon^p N^{p/2}}.$$

The same bound holds for the deviation of $-\varepsilon$ below.

Remarks:

- The proof of Theorem 3 relies on combining concentration results for U-statistics and martingales. One could use similar ideas to cover the two-sample MMD estimator $\widehat{\text{MMD}}^2_{k,U}(\mathbb{P},\mathbb{Q})$ for the exponential kernel.
- **Convergence rate of $\widehat{\text{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})$:** Theorem 2 means a convergence rate $\mathcal{O}_{a.s}\left(\frac{1}{\sqrt{N}}\right)$ of the estimator $\widehat{\text{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})$ for bounded kernels. Theorem 3 implies the same convergence (when taking $\kappa \to 0$) for the unbounded exponential kernel. Indeed, for any $\kappa > 0$, one can find p such that $\kappa p > 2$. Taking

$\varepsilon_N = \left(\frac{1}{\sqrt{N}}\right)^{1-\kappa}$ in the Borel-Cantelli lemma, using Theorem 3 and that in this case $\frac{1}{\varepsilon_N^p N^{p/2}} = \frac{N^{\frac{1}{2}(1-\kappa)p}}{N^{p/2}} = N^{-\frac{\kappa p}{2}}$, one arrives at

$$\sum_{N\in\mathbb{N}^*} \mathbb{P}\left(\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q}) > \varepsilon_N\right) \leq \sum_{N\in\mathbb{N}^*} \frac{C_p}{N^{\frac{\kappa p}{2}}} < \infty.$$

- **Convergence of $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})$:** Similar rate can be proved for the V-statistics, by rewriting $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$ in terms of $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$ (see the supplement).

It is known (Tolstikhin et al., 2016, Theorem 2) that the rate $\frac{1}{\sqrt{N}}$ for bounded continuous radial kernels in the two-sample setting for the class of probability measures is optimal with infinitely differentiable density. We prove that a similar result holds for the considered one-sample setting and unbounded exponential kernel.

Theorem 4 (Minimax rate for semi-explicit MMD, exponential kernel)**.** *Let us consider the exponential kernel $k(x,y) = e^{bxy}$ ($b > 0$, $x, y \in \mathbb{R}$). Let $(\mathbb{P}, k)$ and $(\mathbb{Q}, k)$ satisfy $(\mathcal{P}_k)$, and let $m_{\mathbb{P}}$ and $m_{\mathbb{Q}}$ stand for the mean of $\mathbb{P}$ and $\mathbb{Q}$, respectively. Then*

$$\inf_{\widehat{\mathrm{MMD}}_N} \sup_{\mathbb{P},\mathbb{Q}\in\mathcal{P}} \mathbb{P}\left(\left|\widehat{\mathrm{MMD}}_N - \mathrm{MMD}_k(\mathbb{P},\mathbb{Q})\right| \geq \frac{c}{\sqrt{N}}\right) \geq \max\left(\frac{e^{-\frac{a^2 b}{2}}}{4}, \frac{1-\sqrt{\frac{a^2 b}{2}}}{2}\right)$$

for some finite constant $c > 0$, $a = \sqrt{N}(m_{\mathbb{P}} - m_{\mathbb{Q}})$, and $\widehat{\mathrm{MMD}}_N$ running over all the estimators using the samples $\{x_n\}_{n\in[N]}$.

Remarks:

- Theorem 4 shows that $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q})$ with k being the exponential kernel cannot be estimated at a rate faster than $\frac{1}{\sqrt{N}}$ by any $\widehat{\mathrm{MMD}}_N$ estimator for all $\mathbb{P},\mathbb{Q} \in \mathcal{P}_k$. The fact that the rate $\frac{1}{\sqrt{N}}$ is achievable was shown in Theorem 3.
- The proof relies on the Le Cam's method (Cam, 1973; Tsybakov, 2009). The main technical difference and challenge which were resolved are that using the unbounded exponential kernel one requires a dedicated MMD computation, and with this need the parameter dependence of MMD becomes somewhat intricate.
- The condition $\mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)} < \infty$ appearing in the definition of $\mathcal{P}_k$ can only be milder than (5), since the former is a specific case of (5) with $\lambda = \frac{b}{2}$. In fact, (5) is more restrictive as it can be seen for instance for Gaussian distributions $\mathbb{P} = \mathcal{N}\left(m, \sigma^2\right)$. Indeed, in this case a standard calculation shows that $\mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)} = e^{\frac{m^2}{1-b\sigma^2}}\frac{1}{\sqrt{1-\sigma^2 b}}$ which is finite (or equivalently $\mathbb{P} \in \mathcal{P}_k$) iff $\sigma^2 < 1/b$. However, $\mathbb{E}_{x\sim\mathbb{P}} e^{\lambda x^2} \propto \frac{1}{\sqrt{1-2\lambda\sigma^2}}$ which is finite iff $\lambda < \frac{1}{2\sigma^2}$; in other words the Gaussian distributions $\mathcal{N}\left(m, \sigma^2\right)$ do not obey (5).

Table 1: Kernel definitions. Parameters: $a, b \in \mathbb{R}^{\geq 0}$; $\sigma_0, \sigma, c, \lambda \in \mathbb{R}^{>0}$; $p \in \mathbb{N}$.

Kernel	$k(x,y)$
Gaussian-exponentiated	$e^{-a(x-y)^2+bxy}$
Matérn	$\sigma_0^2 e^{-\frac{\sqrt{2p+1}\|x-y\|}{\sigma}} \frac{p!}{(2p)!} \times \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!}\left(\frac{2\sqrt{2p+1}\|x-y\|}{\sigma}\right)^{p-i}$
Gaussian	$e^{-a(x-y)^2}$
Laplacian	$e^{-\lambda\|x-y\|}$
exponential	e^{bxy}

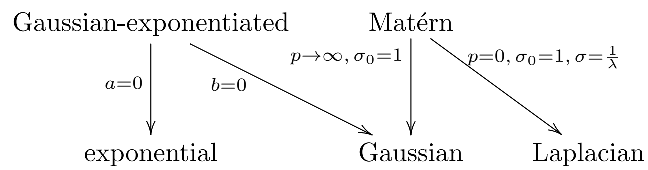

Figure 1: Relation of the kernels in Table 1.

4.2 Analytical Formulas for Mean Embedding

In Section 4.1 we showed that one can leverage with the semi-explicit MMD estimator (3) the analytical knowledge of mean embedding, and get tighter concentration. In this section, we provide a summary of our novel results on such closed-form expressions, accompanied with a discussion on existing results.

Our novel analytical mean embedding results (summarized in Table 3, available in Lemma 1 and Lemma 2—available in the supplement—and related to existing works below) are on the (Gaussian exponentiated, Gaussian) and (Matérn, beta) kernel-distribution pairs, and are motivated by financial applications. The studied Gaussian-exponentiated and the Matérn kernels generalize the widely-used Gaussian, Laplacian and exponential ones (Fig. 1), the beta distribution extends the uniform one (Fig. 2). Kernels (k) and distributions ($\mathbb{Q}$) are summarized in Table 1 and Table 2, with their relations in Fig. 1 and Fig. 2.

Regarding **related literature**, the analytical expression for the mean embedding of the (Gaussian, Gaussian) kernel-distribution pair (Song et al., 2008) can

Table 2: Target distributions. q stands for the pdf of $\mathbb{Q}$. Parameters: $s, m \in \mathbb{R}$; $\alpha, \beta, v, \sigma \in \mathbb{R}^{>0}$.

Distribution	$q(x)$
skew Gaussian	$\frac{2}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} \Phi\left(\frac{s(x-m)}{\sqrt{v}}\right)$
Gaussian	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}$
beta	$\frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{I}_{[0,1]}(x)$
uniform	$\mathbb{I}_{[0,1]}(x)$

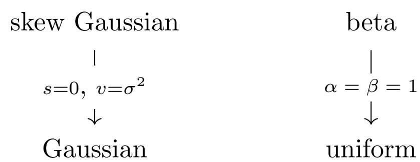

Figure 2: Relation of the distributions in Table 2.

be recovered from Lemma 1 by choosing $b = 0$ in the Gaussian-exponentiated kernel, or by taking $s = 0$ and $v = \sigma^2$ in the (Gaussian, skew Gaussian) kernel-distribution result Kennedy (1998, Section 9.2). Our result on the (Matérn, beta) kernel-distribution pair gives back with $\alpha = \beta = 1$ that by Briol et al. (2019b, Section 5.4) considering the (Matérn, uniform) pair.

5 EXPERIMENTS

In this section, we demonstrate the efficiency of the proposed semi-explicit MMD estimator. Our optimization algorithm is presented in Section 5.1, followed by the used divergence metrics in Section 5.2. Our experimental results are presented in Section 5.3–5.5.

Specifically, we designed the following experiments:

- **Experiment 1**: We compare the speed of convergence of different divergence metrics to zero when $\mathbb{P} = \mathbb{Q}$ for Gaussian distributions (Section 5.3) setting the stage for index replication.
- **Experiment 2**: In Section 5.4, we focus on the index replication problem, and aim to find the index weights matching a target distribution; we tackle the task by the minimization of various divergence measures.
- **Experiment 3**: In Section 5.5 we focus on MMD estimators to perform the calibration of parametric distributions on financial data, including LGD ratios and S&P 500.[2]

[2]S&P 500 is a widely used equity index calculated as the weighted-average value of the 500 most highly capitalised US companies.

Table 3: Summary of obtained analytical mean embedding $\mu_k(\mathbb{Q})$ results for the considered kernel (k) and distribution ($\mathbb{Q}$) pairs.

k	$\mathbb{Q}$	$\mu_k(\mathbb{Q})$
Gaussian-exponentiated	Gaussian	Lemma 1
Matérn	beta	Lemma 2

5.1 Optimization Algorithm

To **optimize** the divergence objectives we tailor the cross-entropy method (CEM; Kroese 2004) to the task, to address the possible non-convexity of our objective functions. Particularly, the CEM technique is a zero-order optimization approach constructing a sequence of pdfs $f\left(\cdot\,;\boldsymbol{\theta}^{(t)}\right)$—we considered Gaussian distributions—which gradually concentrates around the optimum as $t \to \infty$. The idea of the CEM method is to generate samples, followed by adaptively updating $f\left(\cdot\,;\boldsymbol{\theta}^{(t)}\right)$ based on maximum likelihood estimate (MLE) relying on the top ρ-percent of the samples (elite in sense of the consider objective function), and smoothing; for further details of the algorithm the reader is referred to the supplement.

To deal with the constraints arising in our problems (non-negative and sum to one index weights, non-negative variance) one can apply a softmax transformation on the generated samples from $f\left(\cdot\,;\boldsymbol{\theta}^{(t)}\right)$.

5.2 Baseline Divergence Estimators in Section 5.3 and 5.4

The performance of our semi-explicit MMD estimator (4) was compared against the two-sample MMD estimator (2), and contrasted with estimators of the Wasserstein distance and the Kullback-Leibler (KL) divergence, the latter two being classical baselines (Lassance, 2019, Chapter 5).

MMD ($\widehat{\text{MMD}}^2_{k,e,V}$, $\widehat{\text{MMD}}^2_{k,V}$). We consider the exponential kernel k_{exp} and the Gaussian kernel k_{G}, for which analytical mean-embeddings can be computed for the Gaussian distribution (Table 3). The V-statistic variant of the estimator was taken to guarantee non-negative estimates.

KL Divergence ($\widehat{D}_{\text{KL}}$). In the Gaussian setting, one can rely on the KL divergence for Gaussian distributions derived in Duchi (2007, page 13): for $\mathbb{P} = \mathcal{N}(\mu, \sigma)$, $\mathbb{Q} = \mathcal{N}(m, s)$, which can be evaluated in $\mathcal{O}(N)$ by computing the empirical mean and variance of samples from $\mathbb{P}$ when

$\mathbb{Q}$ is known,

$$D_{\mathrm{KL}}(\mathbb{P}, \mathbb{Q}) = \log\left(\frac{s}{\sigma}\right) + \frac{\sigma^2 + (m-\mu)^2}{2s^2} - \frac{1}{2}.$$

Wasserstein Distance ($\widehat{\mathrm{W}}_p$). Let $p \geq 1$. The Wasserstein distance (Peyré and Cuturi, 2019) of the probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_1^+(\mathbb{R})$ is defined as

$$\mathrm{W}_p(\mathbb{P}, \mathbb{Q}) = \left(\int_0^1 \left|F_{\mathbb{P}}^{-1}(t) - F_{\mathbb{Q}}^{-1}(t)\right|^p \mathrm{d}t\right)^{1/p},$$

where $F_{\mathbb{P}}^{-1}$ and $F_{\mathbb{Q}}^{-1}$ are the inverse cdfs of $\mathbb{P}$ and $\mathbb{Q}$, and $L^p([0,1])$ refers to the real-valued p-power Lebesgue-integrable functions on $[0,1]$. An empirical estimator for W_p is, for $\{x_i\}_{i\in[N]} \sim \mathbb{P}$:

$$\widehat{\mathrm{W}}_p(\mathbb{P}_N, \mathbb{Q}) = \left[\frac{1}{N}\sum_{j\in[N]}\left|x_{(j)} - F_{\mathbb{Q}}^{-1}\left(\frac{j}{N}\right)\right|^p\right]^{\frac{1}{p}}.$$

This estimator can be evaluated in $\mathcal{O}(N\log(N))$ time.

5.3 Speed of Convergence

In this section, we explore the speed of convergence of the divergence measures detailed in Section 5.2.

When two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ are equal, a desirable property of their divergence estimator is to converge quickly towards zero; this is what we investigate next. Particularly, we assess the convergence of $D(\mathbb{Q}_N, \mathbb{Q})$, $\mathbb{Q} \sim \mathcal{N}(m, s)$[3], based on samples $\{x_i\}_{i\in[N]} \sim \mathbb{Q}$ for varying N. We took $b = 10^{-3}$ for $k_{\exp}$, $c = 2$ for k_{G} and for $\widehat{\mathrm{W}}_p$ we chose $p = 1$ (but got similar results for other values of p). For the two-sample MMD estimator, we took $M = N$. For each fixed sample size, we performed 100 Monte Carlo simulations to assess the variability of the estimation. The obtained mean and std results are summarized in Fig. 3. As it can be seen in the figure, the semi-explicit MMD estimator converges faster than the two-sample MMD one to 0, with lower std. The exponential kernel provides lower values of the divergence. All divergence metrics show a convergence rate of $\mathcal{O}\left(1/\sqrt{N}\right)$ except for the Wasserstein metric whose slope is around -0.3 in log-log scale for the considered samples.

5.4 Index Replication

In the index replication problem, the aim is to find the weights ($\mathbf{w}$) for a basket of stocks based on the knowledge of the distributions of the stocks returns ($\mathbf{r}$) and that of the index ($\mathbb{P}_T$). The problem can be formulated by minimizing the discrepancy (measured in the sense of a divergence D) between the associated distributions $\mathbf{w}^\top\mathbf{r} \sim \mathbb{P}_{\mathbf{w}}$ and $\mathbb{P}_T$:

$$\mathbf{w}^* = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}^d} D(\mathbb{P}_{\mathbf{w}}, \mathbb{P}_T), \tag{6}$$

where $\mathcal{W}^d$ encodes non-negative weights summing to one.

In this experiment, we designed a challenging low signal-to-noise setting to replicate what is observed in financial markets. Particularly, we worked with $d = 3$ and assumed $\mathbb{P}_T = \mathcal{N}(m, s)$ with $m = \left(\mathbf{w}^0\right)^\top \boldsymbol{\mu}$, $s^2 = \left(\mathbf{w}^0\right)^\top [\mathrm{Diag}(\boldsymbol{\sigma})]^2 \left(\mathbf{w}^0\right)$, $\mathbf{w}^0 = (0.7, 0.2, 0.1)$, $\boldsymbol{\mu} = (0.05, 0.03, 0.01)$ and $\boldsymbol{\sigma} = (0.1, 0.08, 0.05)$ and consequently $m = 0.042$ and $s = 0.0719$, with a std approximately twice the mean.

We used $N = 1500$ (corresponding to 5 years of data) to generate $\{\mathbf{r}\}_{i\in[N]} \sim \mathcal{N}(\boldsymbol{\mu}, \mathrm{Diag}(\boldsymbol{\sigma}))$, and applied the CEM algorithm to solve

$$\mathbf{w}^* = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}^d} D(\mathbb{P}_{\mathbf{w},N}, \mathbb{P}_T),$$

which is the empirical counterpart of (6). We considered all the divergence measures and estimators detailed in Section 5.2; for MMD only the semi-explicit estimator was taken due to its faster convergence experienced in Section 5.3.

Our results, summarized in Fig. 4, show the convergence towards $\mathbf{w}^0$ for the different estimators. In this example the performance of the Gaussian kernel based MMD and the KL divergence supersede that of the exponential kernel based MMD; the latter shows a larger variability which could be explained by the unbounded nature of the exponential kernel.

5.5 Parametric Estimation on Financial Data

Our two sub-experiments for parametric estimation on financial data using MMD were as follows.

1. Calibration of beta distribution on a dataset of historical LGD rates provided by a European bank.[4] It includes $N = 2545$ observations on LGDs.
2. Calibration on the S&P 500[5] of a Gaussian distribution, and of a skew Gaussian distribution.[6]

In line with these calibration tasks we considered the parametric family of distributions $\{\mathbb{Q}_\theta : \theta \in \Theta\}$ as beta, Gaussian or skew Gaussian, and investigated the benefits of the semi-explicit MMD estimator compared to the two-sample one.

[3] The values of m and s were chosen as in the index replication setting (Section 5.4): $m = 0.042$ and $s = 0.0719$.

[4] The data is available at Credit Risk Analytics.

[5] Data can be found on Yahoo Finance.

[6] Although Gaussian is a special case of skew Gaussian, we will compare the goodness-of-fit in the two cases.

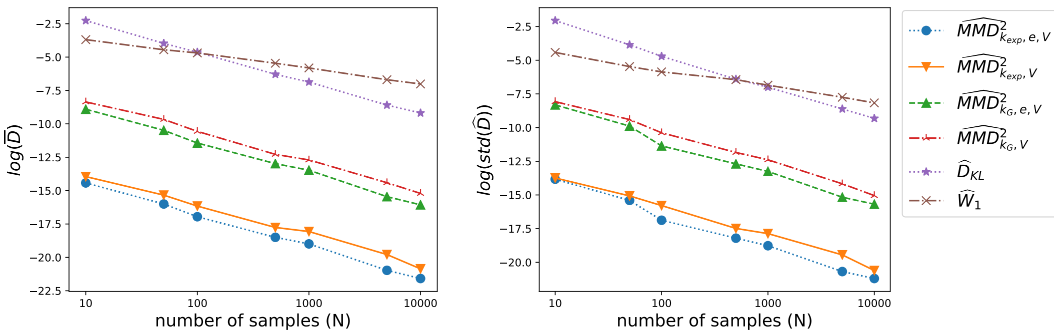

Figure 3: Mean $\pm$ std of various divergences when $\mathbb{P} = \mathbb{Q}$, on log-log scale as a function of number of sample N.

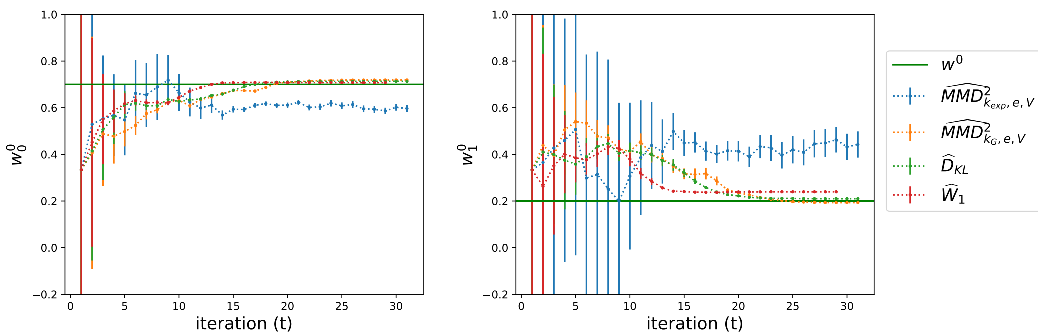

Figure 4: Estimated weights (first two coordinates of $\mathbf{w}^0$; $w^0_3 = 1 - (w^0_1 + w^0_2)$) in index replication as a function of number of iterations.

Particularly, given samples $\{x_i\}_{i\in[N]}$ used for calibration and noticing that the terms in (1) and (3) containing $k(x_i, x_j)$ are independent from θ, the associated problems boil down to the optimization tasks:

$$\begin{aligned}\theta^* &:= \arg\min_{\theta\in\Theta} \widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P}_N, \mathbb{Q}_\theta)\\ &= \arg\min_{\theta\in\Theta} \mathbb{E}_{y\sim\mathbb{Q}_\theta}(\mu_{k,\theta}(y)) - \frac{2}{N}\sum_{i\in[N]} \mu_k(\mathbb{Q}_\theta)(x_i),\\ \theta^* &:= \arg\min_{\theta\in\Theta} \widehat{\mathrm{MMD}}^2_{k,U}(\mathbb{P}_N, \mathbb{Q}_{\theta,M})\\ &= \arg\min_{\theta\in\Theta} \frac{1}{M(M-1)} \sum_{\substack{i,j\in[M]\\ i\neq j}} k(y_i, y_j)\\ &\qquad - \frac{2}{NM}\sum_{\substack{i\in[N]\\ j\in[M]}} k(x_i, y_j).\end{aligned}$$

where the dependence in the latter objective (two-sample MMD estimator) in θ is via the sample $\{y_i\}_{i\in[M]} \overset{\text{i.i.d.}}{\sim} \mathbb{Q}_\theta$.

We chose the following (kernel, distribution) pairs for which analytical mean-embedding are available:

- for the calibration of LGD ratios we selected $(k_{\mathrm{L}}, \mathbb{Q}_{\mathrm{b}}) :=$ (Laplacian, beta),
- for S&P 500 we used the pairs: $(k_{\exp}, \mathbb{Q}_{\mathrm{G}}) :=$ (exponential, Gaussian), $(k_{\mathrm{G}}, \mathbb{Q}_{\mathrm{sG}}) :=$ (Gaussian, skew Gaussian).

In Table 4 we report the mean value and the standard-deviation of the objective function in the semi-explicit and two-sample case. We also performed a test of distribution adequacy using the approximated p-value.

Table 4: Parametric estimation results (mean $\pm$ std) computed on the final elite samples of the CEM algorithm. The p-value is for the semi-explicit estimator.

$(k,\mathbb{Q})$	$\widehat{\mathrm{MMD}}^2_{k,e,U}$	$\widehat{\mathrm{MMD}}^2_{k,U}$	$\hat{p}$
$(k_{\mathrm{L}}, \mathbb{Q}_{\mathrm{b}})$	$-10^{-3} \pm 5\cdot10^{-4}$	$0.5 \pm 8\cdot10^{-2}$	0.63
$(k_{\exp}, \mathbb{Q}_{\mathrm{G}})$	$3\cdot10^{-6} \pm 8\cdot10^{-10}$	$4\cdot10^{-5} \pm 2\cdot10^{-6}$	0.66
$(k_{\mathrm{G}}, \mathbb{Q}_{\mathrm{sG}})$	$5\cdot10^{-3} \pm 2\cdot10^{-6}$	$-2\cdot10^{-3} \pm 7\cdot10^{-6}$	0.95

As it can be seen in Table 4, the semi-explicit MMD consistently provides estimates with a significantly smaller std in the last iteration of the CEM algorithm. It is worth noting that the positive skewness on the last 5 years of S&P 500 returns is correctly captured, leading to a higher p-value with a skew Gaussian rather than with a Gaussian target.

Supplementary Material

In the supplement, we provide the proofs our concentration (Section A) and analytical mean embedding (Section B) results. External statements are collected in Section C. Further details about our experiments are given in Section D.

A PROOFS

In this section, we present the proofs of our concentration results for the semi-explicit MMD estimators. In Section A.1, we give the detailed proofs of our tightened concentration results for the unbiased semi-explicit MMD estimator for bounded kernels (Theorem 2) and for the unbounded exponential kernel (Theorem 3), both leading to a $1/\sqrt{N}$ convergence rate. We follow by extending the convergence rate of the unbiased semi-explicit MMD to its V-statistics counterpart in Section A.2. We then (Section A.3) provide the proof of the optimality of the rate $1/\sqrt{N}$ in the unbounded case of exponential kernel (Theorem 4).

A.1 Proofs of Concentration Results

Proof. (Theorem 2; concentration of $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})$, bounded kernel) By the definition of $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})$ and $\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$, they have the term $\mathbb{E}_{y\sim\mathbb{Q}}\mu_k(\mathbb{Q})(y)$ in common, hence their difference writes as

$$\begin{aligned}
&\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})\\
&\overset{(a)}{=} \underbrace{\frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\neq j}} k(x_i,x_j)}_{=:U_2} - \underbrace{\mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')}_{\mathbb{E}U_2} - 2\left[\underbrace{\frac{1}{N}\sum_{i\in[N]}\mu_k(\mathbb{Q})(x_i)}_{=:U_1} - \underbrace{\mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x)}_{\mathbb{E}U_1}\right]\\
&= U_2 - \mathbb{E}U_2 - 2(U_1-\mathbb{E}U_1)
\end{aligned} \tag{7}$$

by using in (a) that U_1 and U_2 are U-statistics. The kernel of U_1 is $h_1(x)=\mu_k(\mathbb{Q})(x)$, the kernel of U_2 is $h_2(x,x')=k(x,x')$. Since by assumption the kernel k is lower bounded by A and upper bounded by B, the same property holds for h_1 and h_2. Hence applying the Hoeffding bound for U-statistics (Theorem C1), for any $t>0$

$$\begin{aligned}
\mathbb{P}(U_1-\mathbb{E}U_1 < -t) &= \mathbb{P}((-U_1)-\mathbb{E}(-U_1) > t) \le e^{-\frac{2Nt^2}{(B-A)^2}},\\
\mathbb{P}(U_2-\mathbb{E}U_2 > t) &\le e^{-\frac{2\left\lfloor\frac{N}{2}\right\rfloor t^2}{(B-A)^2}}.
\end{aligned} \tag{8}$$

Returning to our target quantity $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$, for any $\varepsilon>0$

$$\begin{aligned}
\left\{\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q}) > \varepsilon\right\} &\overset{(7)}{=} \{U_2-\mathbb{E}U_2-2(U_1-\mathbb{E}U_1)>\varepsilon\}\\
&\overset{(a)}{\subseteq} \left\{U_2-\mathbb{E}U_2 > \frac{\varepsilon}{2}\right\} \cup \left\{U_1-\mathbb{E}U_1 < -\frac{\varepsilon}{4}\right\},
\end{aligned} \tag{9}$$

where the inclusion $A\subseteq B\cup C$ in (a) is equivalent to $\bar{B}\cap\bar{C}\subseteq\bar{A}$; the latter holds as $\left\{U_2-\mathbb{E}U_2\le\frac{\varepsilon}{2}\right\}\cap\left\{-2(U_1-\mathbb{E}U_1)\le\frac{\varepsilon}{2}\right\}\subseteq\{U_2-\mathbb{E}U_2-2(U_1-\mathbb{E}U_1)\le\varepsilon\}$. Using (9) and the bound (8) with $t=\frac{\epsilon}{2}$ for U_2 and $t=\frac{\epsilon}{4}$ for U_1 one arrives at

$$\mathbb{P}\left(\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q}) > \varepsilon\right) \leq \mathbb{P}\left(U_2 - \mathbb{E}U_2 > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(U_1 - \mathbb{E}U_1 < -\frac{\varepsilon}{4}\right)$$
$$\leq e^{-\frac{2\left\lfloor\frac{N}{2}\right\rfloor\varepsilon^2}{2^2(B-A)^2}} + e^{-\frac{2N\varepsilon^2}{4^2(B-A)^2}} = e^{-\frac{\left\lfloor\frac{N}{2}\right\rfloor\varepsilon^2}{2(B-A)^2}} + e^{-\frac{N\varepsilon^2}{8(B-A)^2}}.$$

To establish the bound in $< -\varepsilon$, we can use a similar union bounding argument as in (9):

$$\begin{aligned}
&\left\{\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q}) < -\varepsilon\right\} = \{U_2 - \mathbb{E}U_2 - 2(U_1 - \mathbb{E}U_1) < -\varepsilon\}\\
&\overset{(b)}{\subseteq} \left\{U_2 - \mathbb{E}U_2 < -\frac{\varepsilon}{2}\right\} \cup \left\{-2(U_1 - \mathbb{E}U_1) < -\frac{\varepsilon}{2}\right\}\\
&= \left\{-U_2 - \mathbb{E}{-U_2} > \frac{\varepsilon}{2}\right\} \cup \left\{-(U_1 - \mathbb{E}U_1) < -\frac{\varepsilon}{4}\right\},
\end{aligned} \tag{10}$$

where the inclusion $A \subseteq B \cup C$ in (b) is equivalent to $\bar{B} \cap \bar{C} \subseteq \bar{A}$; the latter holds as $\left\{U_2 - \mathbb{E}U_2 \geq -\frac{\varepsilon}{2}\right\} \cap \left\{-2(U_1 - \mathbb{E}U_1) \geq -\frac{\varepsilon}{2}\right\} \subseteq \{U_2 - \mathbb{E}U_2 - 2(U_1 - \mathbb{E}U_1) \geq -\varepsilon\}$. Using (10) and the bound (8) replacing U_1 by $-U_1$ and U_2 by $-U_2$, with $t = \frac{\epsilon}{2}$ for U_2 and $t = \frac{\epsilon}{4}$ for U_1 we arrived at

$$\mathbb{P}\left(\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q}) > \varepsilon\right) \leq \mathbb{P}\left((-U_2) - \mathbb{E}(-U_2) > \frac{\varepsilon}{2}\right) + \mathbb{P}\left((-U_1) - \mathbb{E}(-U_1) < -\frac{\varepsilon}{4}\right)$$
$$\leq e^{-\frac{2\left\lfloor\frac{N}{2}\right\rfloor\varepsilon^2}{2^2(B-A)^2}} + e^{-\frac{2N\varepsilon^2}{4^2(B-A)^2}} = e^{-\frac{\left\lfloor\frac{N}{2}\right\rfloor\varepsilon^2}{2(B-A)^2}} + e^{-\frac{N\varepsilon^2}{8(B-A)^2}}.$$

□

Proof. (Theorem 3; concentration of $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})$, exponential kernel)

$\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$ is well-defined since $\mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)} = \mathbb{E}_{x\sim\mathbb{P}} e^{\frac{bx^2}{2}}$ and $\mathbb{E}_{x\sim\mathbb{Q}}\sqrt{k(x,x)} = \mathbb{E}_{x\sim\mathbb{Q}} e^{\frac{bx^2}{2}}$ are finite by assumption (5).

Similarly to the proof of Theorem 2, we write the difference $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$ in terms of two U-statistics

$$\begin{aligned}
&\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})\\
&= \underbrace{\frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\neq j}} k(x_i,x_j)}_{=:T_2} - \underbrace{\mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')}_{\mathbb{E}T_2} - 2\left[\underbrace{\frac{1}{N}\sum_{i\in[N]}\mu_k(\mathbb{Q})(x_i)}_{=:T_1} - \underbrace{\mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x)}_{\mathbb{E}T_1}\right]\\
&= T_2 - \mathbb{E}T_2 - 2(T_1 - \mathbb{E}T_1)
\end{aligned}$$

with T_1 having the kernel $h_1(x) = \mu_k(\mathbb{Q})(x)$ and T_2 using the kernel $h_2(x,x') = k(x,x')$. We establish a concentration result on $T_1 - \mathbb{E}T_1$ and $T_2 - \mathbb{E}T_2$ separately, and combine them with the union bound:

$$\begin{aligned}
&\left\{\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q}) > \varepsilon\right\} = \{T_2 - \mathbb{E}T_2 - 2(T_1 - \mathbb{E}T_1) > \varepsilon\}\\
&\subseteq \left\{T_2 - \mathbb{E}T_2 > \frac{\varepsilon}{2}\right\} \cup \left\{-2(T_1 - \mathbb{E}T_1) > \frac{\varepsilon}{2}\right\} = \left\{T_2 - \mathbb{E}T_2 > \frac{\varepsilon}{2}\right\} \cup \left\{T_1 - \mathbb{E}T_1 < -\frac{\varepsilon}{4}\right\}.
\end{aligned} \tag{11}$$

Below let $p \geq 2$ denote a fixed constant. By assumption $\tilde{N} := \frac{N}{2} \in \mathbb{N}^*$.

- **Bound on** T_2: Let us introduce the notation V for the sum of independent processes

$$V(x_1,x_2,\ldots,x_N) = k(x_1,x_2) + k(x_3,x_4) + \cdots + k(x_{N-1},x_N).$$

With this notation our target quantity T_2 can be rewritten (Pitcan, 2017) as

$$T_2 = \frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\neq j}} k(x_i,x_j) = \frac{2}{N}\left(\frac{1}{N!}\sum_{\sigma\in S_N} V(x_{\sigma_1},\ldots,x_{\sigma_N})\right), \tag{12}$$

where S_N denotes the set of permutations of $[N]$. Then for any $t > 0$

$$\begin{aligned}
\mathbb{P}\left(T_2 - \mathbb{E}T_2 > t\right) &= \mathbb{P}\left(\frac{1}{N(N-1)}\sum_{i\neq j} k(x_i,x_j) - \mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x') > t\right)\\
&\overset{(a)}{\leq} \frac{\mathbb{E}\left|\frac{1}{N(N-1)}\sum_{i\neq j} k(x_i,x_j) - \mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')\right|^p}{t^p} \overset{(b)}{=} \frac{\mathbb{E}\left|\frac{2}{N}\frac{1}{N!}\sum_{\sigma\in S_N} V(x_{\sigma_1},\dots,x_{\sigma_N}) - \mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')\right|^p}{t^p}\\
&\overset{(c)}{=} \left(\frac{2}{Nt}\right)^p \mathbb{E}\left|\frac{1}{N!}\sum_{\sigma\in S_N}\left[V(x_{\sigma_1},\dots,x_{\sigma_N}) - \frac{N}{2}\mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')\right]\right|^p\\
&\overset{(d)}{\leq} \left(\frac{2}{Nt}\right)^p \frac{1}{N!}\sum_{\sigma\in S_N}\mathbb{E}\Big|\underbrace{V(x_{\sigma_1},\dots,x_{\sigma_N}) - \frac{N}{2}\mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')}_{M^\sigma_{\tilde{N}}}\Big|^p.
\end{aligned} \tag{13}$$

(a) comes from the generalized Markov's inequality (Lemma C4) by choosing $\phi(x) := |x|^p$ and $I = \mathbb{R}$. In (b) we applied (12). Pulling out $\left(\frac{2}{N}\right)^p$ gives (c). (d) follows from the Jensen inequality by applying it to the argument of the expectation with the convex function $x \mapsto |x|^p$.

Let us introduce the notation $M^\sigma_{\tilde{N}} = V(x_{\sigma_1},\dots,x_{\sigma_N}) - \frac{N}{2}\mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')$ in (13). One can expand $M_{\tilde{N}}$ as a sum of centered independent processes:

$$\begin{aligned}
M^\sigma_{\tilde{N}} &= \underbrace{k(x_{\sigma_1},x_{\sigma_2}) + k(x_{\sigma_3},x_{\sigma_4}) + \cdots + k(x_{\sigma_{N-1}},x_{\sigma_N})}_{\tilde{N}=N/2 \text{ terms}} - \tilde{N}\mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')\\
&= \sum_{k\in[\tilde{N}]}\underbrace{\left[k(x_{\sigma_{2k-1}},x_{\sigma_{2k}}) - \mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')\right]}_{=:Y_k}.
\end{aligned}$$

Similarly, let us denote $M^\sigma_n = \sum_{k\in[n]} Y_k,\ n \in \left[\tilde{N}\right]$. By definition $\mathbb{E}Y_k = 0$ for all $k \in [n]$, which implies that M^σ_n is a martingale w.r.t. the filtration $\mathcal{F}_n = \sigma\left((Y_k)_{k\in[n]}\right)$:

$$\mathbb{E}\left(M^\sigma_n|\mathcal{F}_{n-1}\right) = \mathbb{E}(\underbrace{M^\sigma_{n-1}}_{\mathcal{F}_{n-1}-\text{measurable}} + \underbrace{Y_n}_{\text{independent from } \mathcal{F}_{n-1}} |\mathcal{F}_{n-1}) = M^\sigma_{n-1} + \underbrace{\mathbb{E}(Y_n)}_{0} = M^\sigma_{n-1}.$$

Hence, we can apply the Burkholder's inequality (Theorem C2) on the martingale $\{(M^\sigma_n,\mathcal{F}_n)\}_{n\in\tilde{N}}$; it ensures the existence of a constant $C_p > 0$ such that

$$\mathbb{E}\left|M^\sigma_{\tilde{N}}\right|^p \leq C_p\mathbb{E}\left(\sum_{k\in\tilde{N}} Y_k^2\right)^{p/2} = C_p\left(\tilde{N}\right)^{p/2}\mathbb{E}\left(\frac{1}{\tilde{N}}\sum_{k\in\tilde{N}} Y_k^2\right)^{p/2} \overset{(a)}{\leq} C_p\left(\tilde{N}\right)^{p/2}\underbrace{\mathbb{E}\left(\frac{1}{\tilde{N}}\sum_{k\in[\tilde{N}]}|Y_k|^p\right)}_{=:m_p=\mathbb{E}_{x,x'\sim\mathbb{P}}\left|k(x,x')-\mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')\right|^p}, \tag{14}$$

where in (a) we applied the Jensen inequality to the argument of the expectation with the convex function $x \mapsto x^{p/2}$. Moreover, m_p is finite since by assumption (5) with $\lambda = bp$, $\mathbb{E}_{x\sim\mathbb{P}}|k(x,x)|^p = \mathbb{E}_{x\sim\mathbb{P}}e^{pbx^2} < \infty$ and $\mathbb{E}_{x,x'\sim\mathbb{P}}\left|k(x,x')\right|^p \leq \sqrt{\mathbb{E}_{x\sim\mathbb{P}}|k(x,x)|^p\mathbb{E}_{x'\sim\mathbb{P}}|k(x',x')|^p}$. By (14) we arrive at

$$\mathbb{E}\left|\sum_{k\in\tilde{N}}\left[k(x_{\sigma_k},x_{\sigma_{k+1}}) - \mathbb{E}_{x,x'\sim\mathbb{P}}k(x,x')\right]\right|^p = \mathbb{E}\left|M^\sigma_{\tilde{N}}\right|^p \leq C_p m_p\left(\tilde{N}\right)^{p/2}$$

which combined with (13) gives that for any $t > 0$

$$\mathbb{P}\left(T_2 - \mathbb{E}T_2 > t\right) \leq \left(\frac{2}{Nt}\right)^p\frac{1}{N!}\sum_{\sigma\in S_N} C_p m_p\left(\frac{N}{2}\right)^{p/2} = C_p m_p\left(\frac{2}{Nt^2}\right)^{p/2}. \tag{15}$$

Note: The same bound holds for $-t$

$$\mathbb{P}\left(T_2 - \mathbb{E}T_2 < -t\right) \leq C_p m_p\left(\frac{2}{Nt^2}\right)^{p/2}, \tag{16}$$

by changing k to $-k$ in all the previous steps.

- **Bound on** T_1: Applying the generalized Markov's inequality (Lemma C4) with $\phi(x) := |x|^p$, one can bound the probability $\mathbb{P}(T_1 - \mathbb{E}T_1 > t)$ in terms of $\mathbb{E}|T_1 - \mathbb{E}T_1|^p$ for any $t > 0$ as

$$\mathbb{P}(T_1 - \mathbb{E}T_1 > t) = \mathbb{P}\left(\frac{1}{N}\sum_{n\in[N]} \mu_k(\mathbb{Q})(x_n) - \mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x) > t\right) \le \frac{\mathbb{E}\left|\frac{1}{N}\sum_{n\in[N]}\mu_k(\mathbb{Q})(x_n) - \mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x)\right|^p}{t^p}$$
$$= \frac{1}{(Nt)^p}\mathbb{E}\Big|\underbrace{\sum_{n\in[N]}\left[\mu_k(\mathbb{Q})(x_n) - \mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x)\right]}_{=:S_N}\Big|^p. \tag{17}$$

S_N is a sum of centered independent random variables and $T_1 - \mathbb{E}T_1 = \frac{1}{N}S_N$. By introducing the notation $Z_k = \mu_k(\mathbb{Q})(x_k) - \mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x)$, $S_n = \sum_{k\in[n]} Z_k$ is a martingale w.r.t. the filtration $\mathcal{F}_n := \sigma\left((Z_k)_{k\in[n]}\right)$. Hence, one can apply the Burkholder's inequality (Theorem C2) on $\{(S_n, \mathcal{F}_n)\}_{n\in[N]}$: it ensures the existence of a constant $C_p > 0$ such that

$$\mathbb{E}\,|S_N|^p \le C_p\mathbb{E}\left(\sum_{n\in[N]} Z_n^2\right)^{p/2} = C_pN^{\frac{p}{2}}\mathbb{E}\left(\frac{1}{N}\sum_{n\in[N]} Z_n^2\right)^{p/2} \overset{(a)}{\le} C_pN^{\frac{p}{2}}\mathbb{E}\left(\frac{1}{N}\sum_{n\in[N]}|Z_n|^p\right) = C_pN^{\frac{p}{2}}\mathbb{E}|Z_N|^p$$
$$= C_pN^{p/2}\underbrace{\mathbb{E}_{x'\sim\mathbb{P}}|\mu_k(\mathbb{Q})(x') - \mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x)|^p}_{m'_p}, \tag{18}$$

where in (a) we applied the Jensen inequality with the convex function $\phi(x) := x^{p/2}$. Let us show the finiteness of m'_p.
Proof (finiteness of m'_p):

- Let us first notice that assumption (5) (i.e., $\mathbb{E}_{x\sim\mathbb{P}}e^{\lambda x^2} < \infty$ and $\mathbb{E}_{x\sim\mathbb{Q}}e^{\lambda x^2} < \infty$ for all $\lambda \in \mathbb{R}$) implies that

$$\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}}e^{\lambda xy} < \infty \quad \forall\lambda \in \mathbb{R}^{>0}. \tag{19}$$

Indeed, taking $\lambda' \in \mathbb{R}^{>0}$ and using the inequality $xy \le \frac{x^2+y^2}{2}$ for any $x, y \in \mathbb{R}$, one gets

$$\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}}e^{\lambda' xy} \le \mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}}e^{\frac{\lambda'}{2}\left(x^2+y^2\right)}. \tag{20}$$

By the independence of $x \sim \mathbb{P}$ and $y \sim \mathbb{Q}$, the r.h.s. of (20) equals to $\mathbb{E}_{x\sim\mathbb{P}}e^{\frac{\lambda'}{2}x^2}\mathbb{E}_{y\sim\mathbb{Q}}e^{\frac{\lambda'}{2}y^2}$ which is finite by using (5) with $\lambda = \frac{\lambda'}{2}$.

- Let us show that $m'_p \le 2^p\mathbb{E}_{x\sim\mathbb{P},y\sim\mathbb{Q}}e^{pbxy}$. Indeed,

$$m'_p = \mathbb{E}_{x'\sim\mathbb{P}}|\mu_k(\mathbb{Q})(x') - \mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x)|^p \overset{(a)}{\le} 2^{p-1}\mathbb{E}_{x'\sim\mathbb{P}}\left(|\mu_k(\mathbb{Q})(x')|^p + |\mathbb{E}_{x\sim\mathbb{P}}\mu_k(\mathbb{Q})(x)|^p\right)$$
$$\overset{(b)}{\le} 2^p\mathbb{E}_{x'\sim\mathbb{P}}|\mu_k(\mathbb{Q})(x')|^p \overset{(c)}{=} 2^p\mathbb{E}_{x'\sim\mathbb{P}}|\mathbb{E}_{x\sim\mathbb{Q}}k(x,x')|^p \overset{(d)}{\le} 2^p\mathbb{E}_{x'\sim\mathbb{P}}\mathbb{E}_{x\sim\mathbb{Q}}|k(x,x')|^p \overset{(e)}{=} 2^p\mathbb{E}_{x'\sim\mathbb{P},x\sim\mathbb{Q}}e^{pbxx'}, \tag{21}$$

where (a) follows from the convexity inequality $|a + b|^p \le 2^{p-1}(|a|^p + |b|^p)$ for any $a, b \in \mathbb{R}$, $p \ge 1$ and the linearity of the integral, in (b) we applied the Jensen inequality with the convex function $\phi(x) := x^p$, in (c) the definition of $\mu_k(\mathbb{Q})(x)$ was used, (d) follows from the Jensen inequality, (e) is implied by the fact that $k(x, x') = e^{bxx'}$. Applying (19) with $\lambda = bp > 0$ (as $b > 0$ and $p \ge 2$) implies that $\mathbb{E}_{x'\sim\mathbb{P},x\sim\mathbb{Q}}e^{pbxx'} < \infty$ which guarantees the finiteness of m'_p by (21).

Substituting the bound (18) to (17), one gets that for any $t > 0$

$$\mathbb{P}(T_1 - \mathbb{E}T_1 > t) \le \frac{1}{(Nt)^p}\mathbb{E}\,|S_N|^p \le C_pm'_p\frac{1}{(Nt^2)^{p/2}}. \tag{22}$$

Note: the same bound holds for the deviation below with $-t$

$$\mathbb{P}(T_1 - \mathbb{E}T_1 < -t) \le C_pm'_p\frac{1}{(Nt^2)^{p/2}} \tag{23}$$

by changing T_1 to $-T_1$ in the reasoning above.

Using in (11) the bound (23) for T_1 with $t=\frac{\epsilon}{4}$ and the bound (15) for T_2 with $t=\frac{\epsilon}{2}$ one gets

$$\mathbb{P}\left(\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})>\varepsilon\right)\le\mathbb{P}\left(T_2-\mathbb{E}T_2>\frac{\varepsilon}{2}\right)+\mathbb{P}\left(T_1-\mathbb{E}T_1<-\frac{\varepsilon}{4}\right)$$
$$\le C_p m_p\underbrace{\left(\frac{2^2\times 2}{N\varepsilon^2}\right)^{\frac{p}{2}}}_{\frac{(2\sqrt{2})^p}{\varepsilon^p N^{p/2}}}+C_p m_p'\frac{4^p}{\varepsilon^p N^{p/2}}\le\frac{1}{\varepsilon^p N^{p/2}}\underbrace{\left[C_p m_p(2\sqrt{2})^p+C_p m_p' 4^p\right]}_{=:C=:C_{p,\mathbb{P},k}}.$$

The lower deviation bound with $-\varepsilon$ follows by using the bound (16) for T_2 with $t=\frac{\varepsilon}{2}$ and the bound (22) for T_1 with $t=\frac{\varepsilon}{4}$

$$\mathbb{P}\left(\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})<-\varepsilon\right)\le\mathbb{P}\left(T_2-\mathbb{E}T_2<-\frac{\varepsilon}{2}\right)+\mathbb{P}\left(T_1-\mathbb{E}T_1>\frac{\varepsilon}{4}\right)\le\frac{C_{p,\mathbb{P},k}}{\varepsilon^p N^{p/2}}.$$

□

A.2 Proof of Convergence of the V-statistic based Semi-Explicit MMD Estimator

Proof. (Convergence of $\widehat{\mathrm{MMD}}_{k,e,V}(\mathbb{P},\mathbb{Q})$)

To understand the convergence behavior of $\widehat{\mathrm{MMD}}_{k,e,V}(\mathbb{P},\mathbb{Q})$, let us start by considering the convergence of $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})$.

Convergence of $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})$: Let us rewrite $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$ in terms of $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$. By the definition of $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})$ and $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P}_N,\mathbb{Q})$ [see (4)-(3)], the two estimators only differ in their first terms which we denote as

$$T_1^V:=\frac{1}{N^2}\sum_{i,j\in[N]}k(x_i,x_j),\qquad T_1^U:=\frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\ne j}}k(x_i,x_j).$$

These two terms are closely related; let us write T_1^V in terms of T_1^U

$$T_1^V=\frac{1}{N^2}\left(\sum_{\substack{i,j\in[N]\\ i\ne j}}k(x_i,x_j)+\sum_{i\in[N]}k(x_i,x_i)\right)=\underbrace{\frac{N(N-1)}{N^2}}_{1-\frac{1}{N}}\underbrace{\left(\frac{1}{N(N-1)}\sum_{\substack{i,j\in[N]\\ i\ne j}}k(x_i,x_j)\right)}_{T_1^U}+\frac{1}{N^2}\sum_{i\in[N]}k(x_i,x_i)$$

which means that $T_1^V=\left(1-\frac{1}{N}\right)T_1^U+\frac{1}{N^2}\sum_{i\in[N]}k(x_i,x_i)$. Denoting the second and third common terms of $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})$ and $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})$ by $T_2:=\mathbb{E}_{y\sim\mathbb{Q}}\mu_k(\mathbb{Q})(y)$ and $T_3:=-2\frac{\sum_{i\in[N]}\mu_k(\mathbb{Q})(x_i)}{N}$, we hence have

$$\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})=T_1^V+T_2+T_3-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})$$
$$=\left(1-\frac{1}{N}\right)T_1^U+T_2+T_3-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})+\frac{1}{N^2}\sum_{i\in[N]}k(x_i,x_i)$$
$$=\left(1-\frac{1}{N}\right)\underbrace{\left[T_1^U+T_2+T_3-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})\right]}_{\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})}+\frac{T_2}{N}+\frac{T_3}{N}-\frac{\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})}{N}+\frac{1}{N^2}\sum_{i\in[N]}k(x_i,x_i).$$

This implies that

$$\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})=\left(1-\frac{1}{N}\right)\left[\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})\right]+o_{a.s}\left(\frac{1}{\sqrt{N}}\right)\qquad(24)$$

as T_2 and T_3 are constants, and $\frac{1}{N}\sum_{i\in[N]}k(x_i,x_i)$ converge to a constant by the law of large numbers.

Since $\widehat{\mathrm{MMD}}^2_{k,e,U}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})=\mathcal{O}_{a.s}\left(\frac{1}{\sqrt{N}}\right)$, (24) means that $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q})-\mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})=\mathcal{O}_{a.s.}\left(\frac{1}{\sqrt{N}}\right)$ also holds.

Convergence of $\widehat{\mathrm{MMD}}_{k,e,V}(\mathbb{P},\mathbb{Q})$ Throughout the proof, we assume that $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q})>0$. In this case,

$$\begin{aligned}\mathcal{O}_{a.s}\left(\frac{1}{\sqrt{N}}\right) &\overset{(*)}{=} \left|\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q})\right| \\ &= \left|\widehat{\mathrm{MMD}}_{k,e,V}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P},\mathbb{Q})\right| \Big[\underbrace{\widehat{\mathrm{MMD}}_{k,e,V}(\mathbb{P},\mathbb{Q})}_{\geq 0} + \underbrace{\mathrm{MMD}_k(\mathbb{P},\mathbb{Q})}_{>0}\Big] \\ &\geq \left|\widehat{\mathrm{MMD}}_{k,e,V}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P},\mathbb{Q})\right| \mathrm{MMD}_k(\mathbb{P},\mathbb{Q})\end{aligned}$$

one gets that $\left|\widehat{\mathrm{MMD}}_{k,e,V}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}_k(\mathbb{P},\mathbb{Q})\right| = \mathcal{O}_{a.s}\left(\frac{1}{\sqrt{N}}\right)$ by using in $(*)$ the previously established convergence $\widehat{\mathrm{MMD}}^2_{k,e,V}(\mathbb{P},\mathbb{Q}) - \mathrm{MMD}^2_k(\mathbb{P},\mathbb{Q}) = \mathcal{O}_{a.s.}\left(\frac{1}{\sqrt{N}}\right)$.

□

A.3 Proof of the Minimax Rate for the Unbounded Exponential Kernel

Proof. (Theorem 4) Let $D=(x_n)_{n\in[N]} \overset{\text{i.i.d.}}{\sim} \mathbb{P}\in\mathcal{P}$ and let $\widehat{\mathrm{MMD}}_N$ denote any estimator of $\mathrm{MMD}_k(\mathbb{P},\mathbb{Q})$ based on D. We are interested in the worst-case error (among all $\mathbb{P},\mathbb{Q}\in\mathcal{P}$) of the best estimator $\widehat{\mathrm{MMD}}_N$, in other words our target quantity is

$$\inf_{\widehat{\mathrm{MMD}}_N} \sup_{\mathbb{P},\mathbb{Q}\in\mathcal{P}} \mathbb{P}^N\left(|\widehat{\mathrm{MMD}}_N - \mathrm{MMD}_k(\mathbb{P},\mathbb{Q})| \geq s\right), \quad s>0,$$

where $\mathbb{P}^N$ denotes the N-times product measure of $\mathbb{P}$. Particularly, our goal is to show that $s=\frac{c}{\sqrt{N}}$ is a possible (hence optimal) rate, with some finite constant $c>0$. Let us define a parameteric class of distributions $\mathcal{P}_\Theta$, domain $\mathcal{X}$, and functional F

$$\mathcal{P}_\Theta := \left\{[\mathcal{N}(m,\sigma^2)]^N : (m,\sigma)\in\Theta\right\}, \quad \Theta := \left\{(m,\sigma)\in\mathbb{R}\times\left(0,\frac{1}{\sqrt{b}}\right)\right\}, \quad \mathcal{X}=\mathbb{R}^N, \quad F(\theta)=\mathrm{MMD}_k(\mathbb{P}_\theta,\mathbb{Q})$$

which we will use to invoke Theorem C3. Here $\mathbb{P}_\theta := \mathcal{N}(m_\mathbb{P},\sigma^2)$ where $\theta:=(m_\mathbb{P},\sigma)\in\Theta$ and a fixed $\mathbb{Q}=\mathcal{N}(m_\mathbb{Q},\sigma^2)$ is taken with $\theta_\mathbb{Q}=(m_\mathbb{Q},\sigma)\in\Theta$.[7] First, let us notice that $\mathcal{P}_\Theta\subset\mathcal{P}$ since $\mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)}<\infty$ means that $\sigma^2<\frac{1}{b}$[8]. Using this inclusion one gets the following lower bound (which translated to a lower bound on the target quantity by taking the infimum over $\widehat{\mathrm{MMD}}_N$)

$$\begin{aligned}\sup_{\mathbb{P},\mathbb{Q}\in\mathcal{P}} \mathbb{P}^N\left(|\widehat{\mathrm{MMD}}_N - \mathrm{MMD}_k(\mathbb{P},\mathbb{Q})| \geq s\right) &\geq \sup_{\theta,\theta_\mathbb{Q}\in\Theta} \mathbb{P}_\theta^N\left(|\widehat{\mathrm{MMD}}_N - \mathrm{MMD}_k(\mathbb{P}_\theta,\mathbb{Q}_{\theta_\mathbb{Q}})| \geq s\right) \\ &\geq \sup_{\theta\in\Theta} \mathbb{P}_\theta^N\left(|\widehat{\mathrm{MMD}}_N - \mathrm{MMD}_k(\mathbb{P}_\theta,\mathbb{Q}_{\theta_\mathbb{Q}})| \geq s\right), \quad \forall\theta_\mathbb{Q}\in\Theta, \qquad (25)\end{aligned}$$

which means that for any fixed $\theta_\mathbb{Q}\in\Theta$ we are in the realm of Theorem C3. To apply the theorem, one needs (i) an upper bound on $D_{\mathrm{KL}}\left(\mathbb{Q}_{\theta_\mathbb{Q}}^{\otimes N},\mathbb{P}_\theta^{\otimes N}\right)$, and (ii) a lower bound on $|F(\theta)-F(\theta_\mathbb{Q})|$. This is what we compute in the following.

- **Upper bound on** $D_{\mathrm{KL}}\left(\mathbb{Q}_{\theta_\mathbb{Q}}^{\otimes N},\mathbb{P}_\theta^{\otimes N}\right)$: Let p and q denote the pdf of $\mathbb{P}_\theta$ and $\mathbb{Q}_{\theta_\mathbb{Q}}$. Then the Kullback-Leibler

[7] Notice that the variance parameter of $\mathbb{P}_\theta$ and $\mathbb{Q}_{\theta_\mathbb{Q}}$ are chosen to be identical.

[8] A standard calculation shows that if $\mathbb{P}=\mathcal{N}(m,\sigma)$, $\mathbb{E}_{x\sim\mathbb{P}}\sqrt{k(x,x)} = e^{\frac{m^2}{1-b\sigma^2}}\frac{1}{\sqrt{1-\sigma^2 b}}$ which is finite iff $\sigma^2<1/b$.

divergence can be computed as

$$D_{\mathrm{KL}}\left(\mathbb{Q}_{\theta_{\mathbb{Q}}}^{\otimes N},\mathbb{P}_{\theta}^{\otimes N}\right)=\int_{\mathbb{R}^N}\log\left(\frac{\prod_{n\in[N]}q(x_n)}{\prod_{n\in[N]}p(x_n)}\right)\prod_{j\in[N]}q(x_j)\,\mathrm{d}x_1\ldots\mathrm{d}x_N$$

$$=\sum_{n\in[N]}\underbrace{\int_{\mathbb{R}^N}\log\left(\frac{q(x_n)}{p(x_n)}\right)\prod_{j\in[N]}q(x_j)\,\mathrm{d}x_1\ldots\mathrm{d}x_N}_{\underbrace{\int_{\mathbb{R}}\log\left(\frac{q(x_n)}{p(x_n)}\right)\mathrm{d}q(x_n)}_{D_{\mathrm{KL}}\left(\mathbb{Q}_{\theta_{\mathbb{Q}}},\mathbb{P}_{\theta}\right)}\prod_{j\in[N],j\neq n}\underbrace{\int_{\mathbb{R}}q(x_j)\mathrm{d}x_j}_{=1}}=\sum_{n\in[N]}D_{\mathrm{KL}}\left(\mathbb{Q}_{\theta_{\mathbb{Q}}},\mathbb{P}_{\theta}\right)\overset{(a)}{=}N\frac{(m_{\mathbb{P}}-m_{\mathbb{Q}})^2}{2\sigma^2}=\underbrace{\frac{a^2}{2\sigma^2}}_{=:\alpha},\tag{26}$$

where in (a) we used Lemma C5, and in (b) we assumed that

$$m_{\mathbb{P}}=m_{\mathbb{Q}}+\frac{a}{\sqrt{N}}\tag{27}$$

for some $a>0$.

- **Lower bound on** $|F(\theta)-F(\theta_{\mathbb{Q}})|$: Since $F(\theta_{\mathbb{Q}})=\mathrm{MMD}_k(\mathbb{Q},\mathbb{Q})=0$, it is sufficient to compute $F(\theta)=\mathrm{MMD}_k\left(\mathbb{P}_\theta,\mathbb{Q}_{\theta_{\mathbb{Q}}}\right)$. By Lemma 1,one has

$$[F(\theta)]^2=\frac{1}{\sqrt{1-r^2}}\left[e^{\frac{bm_{\mathbb{P}}^2}{1-r}}+e^{\frac{bm_{\mathbb{Q}}^2}{1-r}}-2e^{\frac{bc\left(m_{\mathbb{P}}^2+m_{\mathbb{Q}}^2\right)+2bm_{\mathbb{P}}m_{\mathbb{Q}}}{2(1-r^2)}}\right]\tag{28}$$

where $r=b\sigma^2$. We are going to show that

$$\left[\mathrm{MMD}_k\left(\mathbb{P}_\theta,\mathbb{Q}_{\theta_{\mathbb{Q}}}\right)\right]^2\geq\frac{(2K)^2}{N}\tag{29}$$

for some constant $K>0$. Let $x=\frac{a}{\sqrt{N}}$, in other words $m_{\mathbb{P}}=m_{\mathbb{Q}}+x$ (in accordance with (27)); we are going to rewrite the squared MMD in (29) as a function of $x=m_{\mathbb{P}}-m_{\mathbb{Q}}$. To do so we will apply a Taylor expansion of the squared MMD around $x=0$. By introducing the notation

$$f_1(x):=e^{\frac{b(m_{\mathbb{Q}}+x)^2}{1-r}}=e^{\frac{bm_{\mathbb{P}}^2}{1-r}},\qquad f_2(x):=e^{\frac{2b(1+c)xm_{\mathbb{Q}}+bcx^2}{2(1-r^2)}},$$

our target quantity writes as

$$\left[\mathrm{MMD}_k\left(\mathbb{P}_\theta,\mathbb{Q}_{\theta_{\mathbb{Q}}}\right)\right]^2=\frac{1}{\sqrt{1-r^2}}\left[f_1\left(\frac{a}{\sqrt{N}}\right)+f_1(0)-2f_1(0)f_2\left(\frac{a}{\sqrt{N}}\right)\right].\tag{30}$$

Indeed, by substituting $m_{\mathbb{P}}=m_{\mathbb{Q}}+x$ in the third term of (28), one gets

$$e^{\frac{bc\left(m_{\mathbb{P}}^2+m_{\mathbb{Q}}^2\right)+2bm_{\mathbb{P}}m_{\mathbb{Q}}}{2(1-r^2)}}=e^{\frac{bc\left[(m_{\mathbb{Q}}+x)^2+m_{\mathbb{Q}}^2\right]+2b(m_{\mathbb{Q}}+x)m_{\mathbb{Q}}}{2(1-r^2)}}=\underbrace{e^{\frac{2bcm_{\mathbb{Q}}^2+2bm_{\mathbb{Q}}^2}{2(1-r^2)}}}_{e^{\frac{2b(c+1)m_{\mathbb{Q}}^2}{2(1-r)(1+c)}}}e^{\frac{bc\left(2xm_{\mathbb{Q}}+x^2\right)+2bxm_{\mathbb{Q}}}{2(1-r^2)}}=\underbrace{e^{\frac{bm_{\mathbb{Q}}^2}{1-r}}}_{f_1(0)}\underbrace{e^{\frac{2b(1+c)xm_{\mathbb{Q}}+bcx^2}{2(1-r^2)}}}_{f_2(x)}.$$

Let us first form the second-order Taylor expansion of f_1 and f_2 around $x=0$; for this approximation the derivatives are

$$f_1'(x)=f_1(x)\left[\frac{2b}{1-r}(m_{\mathbb{Q}}+x)\right],\qquad f_1''(x)=f_1(x)\left(\frac{2b}{1-r}+\left[\frac{2b}{1-r}(m_{\mathbb{Q}}+x)\right]^2\right),$$

$$f_2'(x)=f_2(x)\left(\frac{bm_{\mathbb{Q}}}{1-r}+\frac{bcx}{1-r^2}\right),\qquad f_2''(x)=f_2(x)\left[\frac{bc}{1-r^2}+\left(\frac{bm_{\mathbb{Q}}}{1-r}+\frac{bcx}{1-r^2}\right)^2\right],$$

which means that for $x=0$ one has

$$f_1'(0)=f_1(0)\frac{2bm_{\mathbb{Q}}}{1-r},\qquad f_1''(0)=f_1(0)\left[\frac{2b}{1-r}+\left(\frac{2b}{1-r}m_{\mathbb{Q}}\right)^2\right],$$

$$f_2'(0)=\underbrace{f_2(0)}_{=1}\frac{bm_{\mathbb{Q}}}{1-r},\qquad f_2''(0)=\underbrace{f_2(0)}_{=1}\left[\frac{bc}{1-r^2}+\left(\frac{bm_{\mathbb{Q}}}{1-r}\right)^2\right].$$

Consequently, the 2nd-order Taylor expansion of f_1 and f_2 takes the form

$$f_1\left(\frac{a}{\sqrt{N}}\right) = f_1(0) + f_1(0)\frac{2bm_{\mathbb{Q}}}{1-r}\frac{a}{\sqrt{N}} + f_1(0)\left[\frac{2b}{1-r} + \left(\frac{2b}{1-r}m_{\mathbb{Q}}\right)^2\right]\frac{a^2}{2N} + \mathcal{O}\left(\frac{a^2}{2N}\right),$$
$$f_2\left(\frac{a}{\sqrt{N}}\right) = 1 + \frac{bm_{\mathbb{Q}}}{1-r}\frac{a}{\sqrt{N}} + \left[\frac{bc}{1-r^2} + \left(\frac{bm_{\mathbb{Q}}}{1-r}\right)^2\right]\frac{a^2}{2N} + \mathcal{O}\left(\frac{a^2}{2N}\right).$$

Using these expansions in (30), the $f_1(0)$ and $f_1(0)\frac{2bm_{\mathbb{Q}}}{1-r}\frac{a}{\sqrt{N}}$ terms simplify and one gets

$$\left[\mathrm{MMD}_k\left(\mathbb{P}_\theta, \mathbb{Q}_{\theta_{\mathbb{Q}}}\right)\right]^2 = \frac{1}{\sqrt{1-r^2}}f_1(0)\Bigg(\underbrace{\left[\frac{2b}{1-r} + \left(\frac{2bm_{\mathbb{Q}}}{1-r}\right)^2\right] - 2\left[\frac{rb}{1-r^2} + \left(\frac{bm_{\mathbb{Q}}}{1-r}\right)^2\right]}_{=\underbrace{\frac{2b}{1-r}}_{=\frac{2b(1+c)}{1-r^2}} - \frac{2rb}{1-r^2} + 2\left(\frac{bm_{\mathbb{Q}}}{1-r}\right)^2}\Bigg)\frac{a^2}{2N} + \mathcal{O}\left(\frac{a^2}{2N}\right)$$
$$= \frac{1}{\sqrt{1-r^2}}f_1(0)\left[\frac{2b}{1-r^2} + 2\left(\frac{bm_{\mathbb{Q}}}{1-r}\right)^2\right]\frac{a^2}{2N} + \mathcal{O}\left(\frac{a^2}{2N}\right).$$

This means that the term in $\frac{a^2}{2N}$ will be smaller than the remaining term $\mathcal{O}\left(\frac{a^2}{2N}\right)$ for large enough N. Hence there exists a constant $K > 0$ such that

$$\left[\mathrm{MMD}_k\left(\mathbb{P}_\theta, \mathbb{Q}_{\theta_{\mathbb{Q}}}\right)\right]^2 \geq \frac{(2K)^2}{N}.$$

Hence we have that

$$|F(\theta) - F(\theta_{\mathbb{Q}})| = |F(\theta)| \geq \frac{2K}{\sqrt{N}} := 2s. \tag{31}$$

By using the derived bounds (26) and (31), Theorem C3 can be applied with $\alpha = \frac{a^2}{2\sigma^2}$ and $s = \frac{K}{\sqrt{N}}$, and the bound (25) implies that

$$\inf_{\widehat{\mathrm{MMD}}_N} \sup_{\mathbb{P},\mathbb{Q}\in\mathcal{P}} \mathbb{P}\left(|\widehat{\mathrm{MMD}}_N - \mathrm{MMD}_k\left(\mathbb{P},\mathbb{Q}\right)| \geq \frac{K}{\sqrt{N}}\right) \geq \max\left(\frac{e^{-\frac{a^2}{2\sigma^2}}}{4}, \frac{1-\sqrt{\frac{a^2}{2\sigma^2}}}{2}\right).$$

Since the bound is valid for any σ for which $\sigma^2 < \frac{1}{b}$, by continuity one can also take the limit $\sigma^2 = \frac{1}{b}$ for which the lower bound is maximized and writes as

$$\max\left(\frac{e^{-\frac{a^2b}{2}}}{4}, \frac{1-\sqrt{\frac{a^2b}{2}}}{2}\right).$$

□

B ANALYTICAL MEAN EMBEDDINGS

In Section B.1 we state our novel analytical mean-embedding results, followed their proofs (Section B.2). Auxiliary results related to Lemma 2 are given in Section B.3.

B.1 Our Results on Analytical Mean Embeddings

Below we present our results on analytical mean-embeddings obtained for the (Gaussian-exponentiated, Gaussian) and (Matérn, beta) kernel-distribution pairs.

Lemma 1 (Mean embedding: Gaussian-exponentiated kernel - Gaussian target). *Let the target distribution be Gaussian* $q(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}$*, the kernel be Gaussian-exponentiated* $k(x,y) = e^{-a(x-y)^2+bxy}$ *where* $m \in \mathbb{R}$*,* $\sigma \in \mathbb{R}^{>0}$ $a \geq 0, b \geq 0$*. Then the mean embedding* $\mu_k(\mathbb{Q})$ *can be computed analytically as*

$$\mu_k(\mathbb{Q})(x) = \frac{e^{-\frac{a(x-m)^2}{1+2a\sigma^2}+\frac{2bmx+b(b+4a)\sigma^2x^2}{2(1+2a\sigma^2)}}}{\sqrt{1+2a\sigma^2}}.$$

Lemma 2 (Mean embedding: Matérn kernel - beta target). *Let the target distribution be beta* $q(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\mathbb{I}_{[0,1]}(x)$ *with* $\alpha \in \mathbb{R}^{>0}$*,* $\beta \in \mathbb{R}^{>0}$*, and let the kernel be Matérn with half-integer* ν *(*$\nu = p+\frac{1}{2}$*,* $p \in \mathbb{N}$*),* $\sigma_0 \in \mathbb{R}^{>0}$*,* $\sigma \in \mathbb{R}^{>0}$

$$k(x,y) = \sigma_0^2 e^{-\frac{\sqrt{2p+1}|x-y|}{\sigma}}\frac{p!}{(2p)!}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{2\sqrt{2p+1}|x-y|}{\sigma}\right)^{p-i}.$$

Then the mean embedding $\mu_k(\mathbb{Q})$ *can be analytically computed as*

$$\begin{aligned}\mu_k(\mathbb{Q})(x) = \frac{\sigma_0^2}{B(\alpha,\beta)}\frac{p!}{(2p)!}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{2\sqrt{2p+1}}{\sigma}\right)^{p-i} \times \\ \sum_{k=0}^{p-i}\binom{p-i}{k}x^k\Big[(-1)^{p-i-k}e^{-\frac{\sqrt{2p+1}x}{\sigma}}E_1^{\frac{\sqrt{2p+1}}{\sigma}}\left((0\vee x)\wedge 1, p-i-k+\alpha-1, \beta-1\right) \\ + (-1)^k e^{\frac{\sqrt{2p+1}x}{\sigma}}E_2^{\frac{\sqrt{2p+1}}{\sigma}}\left((0\vee x)\wedge 1, p-i-k+\alpha-1, \beta-1\right)\Big],\end{aligned}$$

where for $a, b > -1$*,* E_1^λ *and* E_2^λ *are defined as*

$$E_1^\lambda(z,a,b) = \int_0^z y^a(1-y)^b e^{\lambda y}\mathrm{d}y, \quad E_2^\lambda(z,a,b) = \int_z^1 y^a(1-y)^b e^{-\lambda y}\mathrm{d}y$$

and can be evaluated using Lemma 3.

B.2 Proofs of Analytical Mean Embeddings

Below we provide the proofs of our analytical mean-embedding results.

Proof. (Lemma 1; mean embedding: Gaussian target - Gaussian-exponentiated kernel) Our target quantity is

$$\mu_k(\mathbb{Q})(x) = \int_{\mathbb{R}}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-m)^2}{2\sigma^2}}e^{-a(x-y)^2+bxy}\mathrm{d}y.$$

Completing the square, one gets

$$\begin{aligned}\frac{(y-m)^2}{2\sigma^2} + a(x-y)^2 - bxy &= \frac{(y-m)^2 + 2\sigma^2 a(x-y)^2 - 2\sigma^2 bxy}{2\sigma^2} \\ &= \frac{\left(1+2a\sigma^2\right)y^2 - 2\left(m+2a\sigma^2x+b\sigma^2x\right)y + 2a\sigma^2x^2+m^2}{2\sigma^2} \\ &= \underbrace{\frac{(1+2a\sigma^2)}{2\sigma^2}}_{=:\frac{1}{2\sigma^{*2}}}\left(y^2 - 2\underbrace{\frac{m+2a\sigma^2x+b\sigma^2x}{1+2a\sigma^2}}_{=:m^*}y\right) + \frac{2a\sigma^2x^2+m^2}{2\sigma^2} \\ &= \frac{1}{2\sigma^{*2}}\left[(y-m^*)^2 - m^{*2}\right] + \frac{2a\sigma^2x^2+m^2}{2\sigma^2} \\ &= \frac{1}{2\sigma^{*2}}(y-m^*)^2 \underbrace{-\frac{(m+2a\sigma^2x+b\sigma^2x)^2}{2\sigma^2(1+2a\sigma^2)} + \frac{2a\sigma^2x^2+m^2}{2\sigma^2}}_{=:-c^*}.\end{aligned}$$

Bringing the two terms in c^* to common denominator, after simplification on arrives at

$$\begin{aligned} c^* &= \frac{\left(m+2a\sigma^2x+b\sigma^2x\right)^2-\left(2a\sigma^2x^2+m^2\right)\left(1+2a\sigma^2\right)}{2\sigma^2\left(1+2a\sigma^2\right)} = \frac{4a\sigma^2mx-2a\sigma^2(m^2+x^2)+b\sigma^2x\left(b\sigma^2x+4a\sigma^2x+2m\right)}{2\sigma^2(1+2a\sigma^2)} \\ &= \frac{-2a\sigma^2(m-x)^2+b\sigma^2x\left(b\sigma^2x+4a\sigma^2x+2m\right)}{2\sigma^2\left(1+2a\sigma^2\right)} = -\frac{a(x-m)^2}{1+2a\sigma^2}+\frac{2bmx+b\sigma^2(b+4a)x^2}{2(1+2a\sigma^2)}. \end{aligned}$$

This means that our target quantity can be rewritten as

$$\mu_k(\mathbb{Q})(x) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^{*2}}(y-m^*)^2+c^*}\mathrm{d}y = \frac{\sigma^*}{\sigma}e^{c^*}\underbrace{\int_{\mathbb{R}} \frac{1}{\sqrt{2\pi(\sigma^*)^2}} e^{-\frac{1}{2\sigma^{*2}}(y-m^*)^2}\mathrm{d}y}_{=1} = \frac{e^{-\frac{a(x-m)^2}{1+2a\sigma^2}+\frac{2bmx+b(b+4a)\sigma^2x^2}{2(1+2a\sigma^2)}}}{\sqrt{1+2a\sigma^2}}.$$

□

Proof. (Lemma 2; mean embedding: beta target - Matérn kernel) Our target quantity is

$$\begin{aligned} &\mu_k(\mathbb{Q})(x) \\ &= \frac{\sigma_0^2}{B(\alpha,\beta)}\int_0^1 e^{-\frac{\sqrt{2p+1}|x-y|}{\sigma}}\frac{p!}{(2p)!}\sum_{i=0}^p\frac{(p+i)!}{i!(p-i)!}\left(\frac{2\sqrt{2p+1}|x-y|}{\sigma}\right)^{p-i} y^{\alpha-1}(1-y)^{\beta-1}\mathbb{I}_{y\in[0,1]}\mathrm{d}y \\ &\overset{(*)}{=} \frac{\sigma_0^2}{B(\alpha,\beta)}\frac{p!}{(2p)!}\sum_{i=0}^p\frac{(p+i)!}{i!(p-i)!}\left(\frac{2\sqrt{2p+1}}{\sigma}\right)^{p-i}\left[\int_0^{(0\vee x)\wedge 1}(x-y)^{p-i}y^{\alpha-1}(1-y)^{\beta-1}e^{-\frac{\sqrt{2p+1}(x-y)}{\sigma}}\mathrm{d}y\right. \\ &\qquad\left. + \int_{(0\vee x)\wedge 1}^1 (y-x)^{p-i}y^{\alpha-1}(1-y)^{\beta-1}e^{-\frac{\sqrt{2p+1}(y-x)}{\sigma}}\mathrm{d}y\right]. \end{aligned}$$

where in $(*)$ we applied the decomposition trick: for $f:\mathbb{R}\to\mathbb{R}$ and $x\in\mathbb{R}$,

$$\int_0^1 f(|x-y|)\mathrm{d}y = \int_0^{(0\vee x)\wedge 1} f(x-y)\mathrm{d}y + \int_{(0\vee x)\wedge 1}^1 f(y-x)\mathrm{d}y.$$

Using the binomial formula on $(x-y)^{p-i} = \sum_{k=0}^{p-i}\binom{p-i}{k}x^k(-y)^{p-i-k}$ and on $(y-x)^{p-i}$, one can rewrite the two integrals as

$$\begin{aligned} &\int_0^{(0\vee x)\wedge 1}(x-y)^{p-i}y^{\alpha-1}(1-y)^{\beta-1}e^{-\frac{\sqrt{2p+1}(x-y)}{\sigma}}\mathrm{d}y = \\ &\quad= \sum_{k=0}^{p-i}\binom{p-i}{k}(-1)^{p-i-k}x^k\int_0^{(0\vee x)\wedge 1} y^{p-i-k}y^{\alpha-1}(1-y)^{\beta-1}e^{-\frac{\sqrt{2p+1}(x-y)}{\sigma}}\mathrm{d}y \\ &\quad= \sum_{k=0}^{p-i}\binom{p-i}{k}(-1)^{p-i-k}x^k e^{-\frac{\sqrt{2p+1}x}{\sigma}}\underbrace{\int_0^{(0\vee x)\wedge 1} y^{p-i-k+\alpha-1}(1-y)^{\beta-1}e^{\frac{\sqrt{2p+1}y}{\sigma}}\mathrm{d}y}_{E_1^{\frac{\sqrt{2p+1}}{\sigma}}((0\vee x)\wedge 1,\, p-i-k+\alpha-1,\, \beta-1)}, \end{aligned}$$

and

$$\begin{aligned} &\int_{(0\vee x)\wedge 1}^1 (y-x)^{p-i}y^{\alpha-1}(1-y)^{\beta-1}e^{-\frac{\sqrt{2p+1}(y-x)}{\sigma}}\mathrm{d}y = \\ &\quad= \sum_{k=0}^{p-i}\binom{p-i}{k}(-x)^k\int_0^{(0\vee x)\wedge 1} y^{p-i-k}y^{\alpha-1}(1-y)^{\beta-1}e^{-\frac{\sqrt{2p+1}(y-x)}{\sigma}}\mathrm{d}y \\ &\quad= \sum_{k=0}^{p-i}\binom{p-i}{k}(-x)^k e^{\frac{\sqrt{2p+1}x}{\sigma}}\underbrace{\int_{(0\vee x)\wedge 1}^1 y^{p-i-k+\alpha-1}(1-y)^{\beta-1}e^{-\frac{\sqrt{2p+1}y}{\sigma}}\mathrm{d}y}_{E_2^{\frac{\sqrt{2p+1}}{\sigma}}((0\vee x)\wedge 1,\, p-i-k+\alpha-1,\, \beta-1)}, \end{aligned}$$

Hence,

$$\mu_k(\mathbb{Q})(x) = \frac{\sigma_0^2}{B(\alpha,\beta)}\frac{p!}{(2p)!}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{2\sqrt{2p+1}}{\sigma}\right)^{p-i}\times$$
$$\sum_{k=0}^{p-i}\binom{p-i}{k}x^k\Bigg[(-1)^{p-i-k}e^{-\frac{\sqrt{2p+1}x}{\sigma}}E_1\left((0\vee x)\wedge 1, p-i-k+\alpha-1, \beta-1\right)$$
$$+(-1)^k e^{\frac{\sqrt{2p+1}x}{\sigma}}E_2\left((0\vee x)\wedge 1, p-i-k+\alpha-1,\beta-1\right)\Bigg].$$

□

B.3 Auxiliary Results for Analytical Mean Embeddings

In this section, we present an approximation formula on E_1^λ and E_2^λ which can be used to evaluate the (Matérn kernel, beta target) mean-embedding in Lemma 2.

Lemma 3 (Infinite sum formulation of E_1^λ and E_2^λ, and truncation error control)**.** *For $z\in[0,1]$, $a>-1$, $b>-1$, let $E_1^\lambda(z,a,b)=\int_0^z y^a(1-y)^b e^{\lambda y}\mathrm{d}y$ and $E_2^\lambda(z,a,b)=\int_z^1 y^a(1-y)^b e^{-\lambda y}\mathrm{d}y$. Then we have the following infinite sum formulation:*

$$E_1^\lambda(z,a,b)=\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}B_{inc}\left(a+k+1,b+1,z\right),$$
$$E_2^\lambda(z,a,b)=\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}\left[B\left(a+k+1,b+1\right)-B_{inc}\left(a+k+1,b+1,z\right)\right].$$

Let $K\in\mathbb{N}$ be fixed and let us denote the truncated E_1^λ and E_2^λ by

$$E_1^{\lambda,tr}=\sum_{k=0}^{K}\frac{\lambda^k}{k!}B_{inc}\left(a+k+1,b+1,z\right),$$
$$E_2^{\lambda,tr}=\sum_{k=0}^{K}\frac{(-\lambda)^k}{k!}\left[B\left(a+k+1,b+1\right)-B_{inc}\left(a+k+1,b+1,z\right)\right].$$

Then the following bounds hold for the truncation errors

$$E_1^\lambda-E_1^{\lambda,tr}\le\frac{\lambda^{K+1}e^{\lambda z}}{(K+1)!}B_{inc}\left(a+K+2,b+1,z\right):=\mathcal{E}_1^{\lambda,tr},$$
$$E_2^\lambda-E_2^{\lambda,tr}\le\frac{(-\lambda)^{K+1}e^{-\lambda z}}{(K+1)!}\left[B\left(a+K+2,b+1\right)-B_{inc}\left(a+K+2,b+1,z\right)\right]:=\mathcal{E}_2^{\lambda,tr}.$$

Proof. (Lemma 3) For $z\in[0,1]$, $a>-1$ and $b>-1$, let

$$E_1^\lambda(z,a,b)=\int_0^z y^a(1-y)^b e^{\lambda y}\mathrm{d}y,\quad E_2^\lambda(z,a,b)=\int_z^1 y^a(1-y)^b e^{-\lambda y}\mathrm{d}y.$$

The infinite sum formulations follow from the exponential series expansions $e^{\lambda y}=\sum_{k=0}^\infty\frac{(\lambda y)^k}{k!}$ and $e^{-\lambda y}=$

$\sum_{k=0}^{\infty} \frac{(-\lambda y)^k}{k!}$:

$$E_1^{\lambda}(z,a,b) = \int_0^z y^a(1-y)^b e^{\lambda y}\mathrm{d}y = \int_0^z y^a(1-y)^b \sum_{k=0}^{\infty}\frac{(\lambda y)^k}{k!}\mathrm{d}y = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\underbrace{\int_0^z y^{a+k}(1-y)^b\mathrm{d}y}_{B_{\text{inc}}(a+k+1,b+1,z)},$$

$$\begin{aligned}E_2^{\lambda}(z,a,b) &= \int_z^1 y^a(1-y)^{b+1}e^{-\lambda y}\mathrm{d}y = \sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}\int_z^1 y^{a+k}(1-y)^b\mathrm{d}y \\ &= \sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}\Bigg[\underbrace{\int_0^1 y^{a+k}(1-y)^b\mathrm{d}y}_{B(a+k+1,b+1)} - \underbrace{\int_0^z y^{a+k}(1-y)^b\mathrm{d}y}_{B_{\text{inc}}(a+k+1,b+1,z)}\Bigg].\end{aligned}$$

Let us now fix $K \in \mathbb{N}$, and truncate $E_1^{\lambda}(z,a,b)$ and $E_2^{\lambda}(z,a,b)$ to the first $K+1$ terms:

$$\begin{aligned}E_1^{\lambda,tr} &= \sum_{k=0}^{K}\frac{\lambda^k}{k!}B_{\text{inc}}\left(a+k+1,b+1,z\right), \\ E_2^{\lambda,tr} &= \sum_{k=0}^{K}\frac{(-\lambda)^k}{k!}\left[B\left(a+k+1,b+1\right) - B_{\text{inc}}\left(a+k+1,b+1,z\right)\right].\end{aligned}$$

By the Taylor-Lagrange theorem, in case of

- E_1^{λ}: for any $y \in [0,z]$ there is a $y_K \in (0,y)$ such that

$$e^{\lambda y} = \sum_{k=0}^{K}\frac{(\lambda y)^k}{K!} + e^{\lambda y_K}\frac{(\lambda y)^{K+1}}{(K+1)!}.$$

- E_2^{λ}: for any $y \in [z,1]$ there is a $y_K' \in (0,y)$ such that

$$e^{-\lambda y} = \sum_{k=0}^{K}\frac{(-\lambda y)^k}{K!} + e^{-\lambda y_K'}\frac{(-\lambda y)^{K+1}}{(K+1)!}.$$

Hence, the truncation errors can be bounded as

$$\begin{aligned}E_1^{\lambda} - E_1^{\lambda,tr} &= \int_0^z y^a(1-y)^b e^{\lambda y_K}\frac{(\lambda y)^{K+1}}{(K+1)!}\mathrm{d}y \overset{(a)}{\leq} \int_0^z y^a(1-y)^{-1/2}e^{\lambda z}\frac{(\lambda y)^{K+1}}{(K+1)!}\mathrm{d}y \\ &= \frac{\lambda^{K+1}e^{\lambda z}}{(K+1)!}\underbrace{\int_0^z y^{a+K+1}(1-y)^b\mathrm{d}y}_{B_{\text{inc}}(a+K+2,b+1,z)}, \\ E_{21}^{\lambda} - E_{21}^{\lambda,tr} &= \int_z^1 y^a(1-y)^b e^{-\lambda y_K'}\frac{(-\lambda y)^{K+1}}{(K+1)!}\mathrm{d}y \overset{(b)}{\leq} \int_z^1 y^a(1-y)^b e^{-\lambda z}\frac{(-\lambda y)^{K+1}}{(K+1)!}\mathrm{d}y \\ &= \frac{(-\lambda)^{K+1}}{(K+1)!}e^{-\lambda z}\int_z^1 y^{a+K+1}(1-y)^b\mathrm{d}y \\ &= \frac{(-\lambda)^{K+1}}{(K+1)!}e^{-\lambda z}\left[\underbrace{\int_0^1 y^{a+K+1}(1-y)^b\mathrm{d}y}_{B(a+K+2,b+1)} - \underbrace{\int_0^z y^{a+K+1}(1-y)^b\mathrm{d}y}_{B_{\text{inc}}(a+K+2,b+1,z)}\right] \\ &= \frac{(-\lambda)^{K+1}}{(K+1)!}e^{-\lambda z}\left[B\left(a+K+2,b+1\right) - B_{\text{inc}}\left(a+K+2,b+1,z\right)\right],\end{aligned}$$

where in (a) we used that $e^{\lambda y_K} \leq e^{\lambda z}$ and in (b) that $e^{-\lambda y_K'} \leq e^{-\lambda z}$. □

C EXTERNAL STATEMENTS

This section contains external statements used in the proofs of our concentration results.

Theorem C1 (Hoeffding inequality for U-statistic; Hoeffding (1963), Pitcan (2017))**.** *Assume that we have n i.i.d. samples* $\{X_i\}_{i\in[n]} \sim \mathbb{P}$*. Let* I_m^n *be the set m-tuples chosen without repetition from* $[n]$*. Suppose that* $h : \mathbb{R}^m \to \mathbb{R}$ *is bounded:* $a \le h(x_1,\dots,x_m) \le b$ *for all* $(x_1,\dots,x_m)$*. We denote* $m_h = \mathbb{E}h(X_1,\dots,X_m)$ *and its U-statistic based estimator* $U_n = \frac{1}{\binom{n}{m}}\sum_{(i_1,\dots,i_m)\in I_m^n} h(X_{i_1},\dots,X_{i_m})$*. Then, for any* $\varepsilon > 0$

$$\mathbb{P}\left(U_n - m_h > \varepsilon\right) \le e^{-\frac{2\left\lfloor \frac{n}{m}\right\rfloor \varepsilon^2}{(b-a)^2}},$$

and the same deviation bound holds for $-\varepsilon$ *below, i.e.* $\mathbb{P}\left(U_n - m_h < -\varepsilon\right) \le e^{-\frac{2\left\lfloor \frac{n}{m}\right\rfloor \varepsilon^2}{(b-a)^2}}$.

Lemma C4 (Generalized Markov's inequality; (2.1) in Boucheron et al. (2013))**.** *Let* ϕ *denote a nondecreasing and nonnegative function defined on* $I \subseteq \mathbb{R}$ *and let* Y *denote a random variable taking values in* I*. Then Markov's inequality implies that for every* $t \in I$ *with* $\phi(t) > 0$

$$\mathbb{P}\left(Y \ge t\right) \le \frac{\mathbb{E}\phi(Y)}{\phi(t)}.$$

Theorem C2 (Burkholder's inequality; Theorem 2.10 in Hall and Heyde (1980))**.** *Assume that* $\{(S_i,\mathcal{F}_i)\}_{i\in[n]}$ *is a martingale sequence and its filtration,* $1 < p < \infty$*. Let the associated martingale increments be denoted by* $X_1 = S_1$ *and* $X_i = S_i - S_{i-1}, 2 \le i \le n$*. Then there exist a constant* C_p *depending on p such that*

$$\mathbb{E}\left|S_n\right|^p \le C_p \mathbb{E}\left|\sum_{i=1}^n X_i^2\right|^{p/2}.$$

Theorem C3 (Theorem 2.2 in Tsybakov (2009))**.** *Let* $\mathcal{X}$ *and* Θ *denote two measurable spaces. Let* $F : \Theta \to \mathbb{R}$ *be a functional. Let* $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\}$ *be a class of probability measures on* $\mathcal{X}$ *indexed by* Θ*. We observe the data* D *distributed according* $\mathbb{P}_\theta \in \mathcal{P}_\Theta$ *with some unknown* θ*. The goal is to estimate* $F(\theta)$*. Let* $\hat{F} := \hat{F}(D)$ *be an estimator of* $F(\theta)$ *based on* D*. Assume that there exist* $\theta_0, \theta_1 \in \Theta$ *such that* $|F(\theta_0) - F(\theta_1)| \ge 2s > 0$ *and* $D_{KL}\left(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_0}\right) \le \alpha$ *with* $0 < \alpha < \infty$*. Then*

$$\inf_{\hat{F}} \sup_{\theta\in\Theta} \mathbb{P}_\theta\left(|\hat{F} - F(\theta)| \ge s\right) \ge \max\left(\frac{e^{-\alpha}}{4}, \frac{1-\sqrt{\alpha/2}}{2}\right).$$

Note: Typically $\mathcal{X}$*,* D *and* $\mathbb{P}_\theta$ *depend on the sample size* N*.*

Lemma C5 (Kullback-Leibler divergence for univariate Gaussian variables; page 13 in Duchi (2007))**.** *Let* $\mathbb{P} = \mathcal{N}(m_\mathbb{P}, \sigma_\mathbb{P})$*,* $\mathbb{Q} = \mathcal{N}(m_\mathbb{Q}, \sigma_\mathbb{Q})$*,* $m_\mathbb{P}, m_\mathbb{Q} \in \mathbb{R}$*,* $\sigma_\mathbb{P}, \sigma_\mathbb{Q} \in \mathbb{R}^{>0}$*. Then*

$$D_{KL}\left(\mathbb{P},\mathbb{Q}\right) = \log\left(\frac{\sigma_\mathbb{Q}}{\sigma_\mathbb{P}}\right) + \frac{\sigma_\mathbb{P}^2 + (m_\mathbb{Q} - m_\mathbb{P})^2}{2\sigma_\mathbb{Q}^2} - \frac{1}{2}.$$

D FURTHER EXPERIMENTAL DETAILS

In this section, we present the CEM algorithm and the transformation functions used to deal with constraints on the parameters.

D.1 CEM Algorithm

The CEM algorithm maximizing an objective function L is given in Alg. 1.

D.2 Parameter Settings in our Experiments

In this section, we detail how we used the CEM algorithm in our numerical experiments.

Algorithm 1 Maximization of L with CEM

Input: Initial value $\boldsymbol{\theta}^{(0)}$, quantile parameter $\rho > 0$, smoothing parameter $\omega \in (0, 1]$, sample size $S \in \mathbb{N}^*$, accuracy $\epsilon > 0$, iterations number $T \in \mathbb{N}^*$.
Initialize the iteration and the elite level: $t = 1$, $\gamma_0 = +\infty$.
repeat
  Generate samples: $\{\mathbf{x}_s\}_{s\in[S]} \overset{\text{i.i.d.}}{\sim} f\left(\cdot\,; \boldsymbol{\theta}^{(t-1)}\right)$.
  Evaluate performance: $L_s = L(\mathbf{x}_s)$, $s \in [S]$.
  Set level: $\gamma_t = L_{(\lceil(1-\rho)S\rceil)}$ $\{(1-\rho)$-quantile of $\{L(\mathbf{x}_s)\}_{s\in[S]}\}$.
  Estimate new parameter: $\tilde{\boldsymbol{\theta}}^{(t)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{S}\sum_{s\in[S]} \mathbb{I}_{\{L(\mathbf{x}_s)\geq\gamma_t\}} \log\left[f\left(\mathbf{x}_s; \boldsymbol{\theta}\right)\right]$ {MLE on the elite}.
  Smoothing: $\boldsymbol{\theta}^{(t)} = (1-\omega)\boldsymbol{\theta}^{(t-1)} + \omega\tilde{\boldsymbol{\theta}}^{(t)}$.
until $(t \leq T)$ **and** $(\max(|\gamma_t - \gamma_{t-1}|, \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|_\infty) \geq \epsilon)$
Output: $\hat{\mathbf{x}} = \mathbb{E}_{\mathbf{x}\sim f\left(\cdot\,;\boldsymbol{\theta}^{(T)}\right)}\mathbf{x}$.

CEM choice of hyperparameters: In all our experiments, we chose the pdf $f(\cdot\,; \boldsymbol{\theta})$ to be Gaussian with dimension adapted to the size of the problem, with a mean value initialized taking into account the constraints on the parameters, and a covariance matrix set to the identity: $\theta^{(t)} = (\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, $\boldsymbol{\mu}_0 \in \mathbb{R}^d$, $\boldsymbol{\Sigma}_0 = \mathbf{I}_d \in \mathbb{R}^{d\times d}$. We considered $N = 150$, $\rho = 0.1$ and set the maximum number of iterations to $T = 30$. A stopping criteria on the update of the elite parameter γ_t and on the samples distribution parameter $\boldsymbol{\theta}^{(t)}$ was considered, under the form $\max(|\gamma_t - \gamma_{t-1}|, \|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\|_\infty) < \epsilon$, so that the algorithm stops when these parameters do not change much. The threshold in the stopping criteria was conservatively set to $\epsilon = 10^{-8}$.

CEM adaptation to constraints: Since the considered Gaussian distribution has unbounded support, and our experiments involve loss functions on specific domains of parameters, a transformation was applied on the samples to ensure this constraint. Denoting by g the transformation function, the loss function was evaluated on $\{g(\boldsymbol{x}_s)\}_{s\in[S]}$; we chose g in CEM as follows.

- **Experiment 2**: The index replication problem involved optimising the loss function on $\mathcal{W}^3 := \left\{\mathbf{w} \in \left(\mathbb{R}^{\geq 0}\right)^3 : \sum_{j=1}^3 w_j = 1\right\}$. For $\mathbf{w} \in \mathcal{W}^3$, by $w_3 = 1 - w_1 - w_2$, the problem can be reduced to $d = 2$. This corresponds to taking $g((x_1, x_2)) = (x_1, x_2, 1 - x_1 - x_2)$. We chose $\boldsymbol{\mu}_0 = \left(\frac{1}{3}, \frac{1}{3}\right)$. The positivity constraint could also be enforced by considering the transformation

$$g(\boldsymbol{x}) = \left(\frac{e^{x_1}}{1 + \sum_{i=1}^2 e^{x_i}}, \frac{e^{x_2}}{1 + \sum_{i=1}^2 e^{x_i}}, \frac{1}{1 + \sum_{i=1}^2 e^{x_i}}\right),$$

  with $\boldsymbol{\mu}_0 = (0, 0)$ (to ensure that $g(\boldsymbol{\mu}_0) = \left(\frac{1}{3}, \frac{1}{3}\right)$). We explored both options in our experiments. We got slightly better results by just enforcing the sum-to-one constraint.
- **Experiment 3**: Here, we performed the calibration of a beta distribution, for which the parameters have to be positive (or even smaller than one), and of a Gaussian and skew Gaussian distribution, for which the variance has to be positive.
  - beta calibration: In our specific application of beta distribution on LGD ratios, the distribution was displaying a U-shape (see Fig. 5), which is typical of a beta distribution with parameters $\alpha, \beta \in (0, 1)$. To enforce this constraint on α and β, the following sigmoid transformation from $\mathbb{R}^2$ to $(0, 1)^2$ was applied:

$$g(\boldsymbol{x}) = \left(\frac{e^{x_1}}{1 + e^{x_1}}, \frac{e^{x_2}}{1 + e^{x_2}}\right).$$

  - Gaussian and skew Gaussian calibration: To deal with positive variance (parameters σ and υ), the transformation $x \mapsto e^x$ mapping $\mathbb{R}$ to $\mathbb{R}^{>0}$ was applied on respective coordinates.

References

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022). Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858–879.

Alquier, P. and Gerber, M. (2024). Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1):71–92.

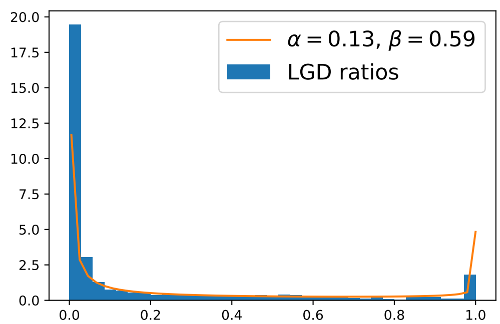

Figure 5: Histogram of LGD ratios and the calibrated beta distribution.

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

Balasubramanian, K., Li, T., and Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 22(1):1–45.

Bamberg, G. and Wagner, N. (2000). Equity index replication with standard and robust regression estimators. *OR-Spektrum*, 22:525–543.

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

Baum, J., Kanagawa, H., and Gretton, A. (2023). A kernel Stein test of goodness of fit for sequential models. In *International Conference on Machine Learning (ICML)*, pages 1936–1953.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer.

Bishwal, J. P. (2007). *Parameter estimation in stochastic differential equations*. Springer.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., and Riec, B. (2020). Graph kernels: State-of-the-art and future challenges. *Foundations and Trends in Machine Learning*, 13(5-6):531–712.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019a). Statistical inference for generative models with maximum mean discrepancy. Technical report. (`https://arxiv.org/abs/1906.05944`).

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., and Sejdinovic, D. (2019b). Probabilistic integration: a role in statistical computation? *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, 34(1):38–42.

Cam, L. L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1:38–53.

Casella, G. and Berger, R. (2024). *Statistical inference*. CRC Press.

Chalabi, Y. and Wuertz, D. (2012). Portfolio optimization based on divergence measures. Technical report. (`https://mpra.ub.uni-muenchen.de/43332/`).

Chen, R. and Wang, Z. (2013). Curve fitting of the corporate recovery rates: the comparison of beta distribution estimation and kernel density estimation. Technical report. (`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3707900/`).

Christopher Adcock, M. E. and Loperfido, N. (2015). Skewed distributions in finance and actuarial science: a review. *The European Journal of Finance*, 21(13-14):1253–1281.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

Deb, N., Ghosal, P., and Sen, B. (2020). Measuring association on topological spaces using kernels and geometric graphs. Technical report. (https://arxiv.org/pdf/2010.01768/).

Duchi, J. (2007). Derivations for linear algebra and optimization. Technical report, The University of Stanford. (https://web.stanford.edu/~jduchi/projects/general_notes.pdf).

Fortin, I. and Kuzmics, C. (2002). Tail-dependence in stock-return pairs. *Intelligent Systems in Accounting, Finance & Management*, 11(2):89–107.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 498–496.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592.

Gupton, G. and Stein, R. (2002). Losscalctm: Model for predicting loss given default (lgd).

Hagrass, O., Sriperumbudur, B., and Li, B. (2024). Spectral regularized kernel two-sample tests. *The Annals of Statistics*, 52(3):1076–1101.

Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London.

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

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Jadhav, D. and Ramanathan, T. (2009). Parametric and non-parametric estimation of value-at-risk. *The Journal of Risk Model Validation*, 3(1):51–71.

Kennedy, M. (1998). Bayesian quadrature with non-normal approximating functions. *Statistics and Computing*, 8(4):365–375.

Király, F. J. and Oberhauser, H. (2019). Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20:1–45.

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

Kroese, R. Y. R. D. P. (2004). *The Cross-Entropy Method*. Springer.

Lassance, N. (2019). *Information-theoretic approaches to portfolio selection*. PhD thesis, Louvain School of Management.

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

Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

Pitcan, Y. (2017). A Note on Concentration Inequalities for U-Statistics. Technical report, University of Berkeley. (https://arxiv.org/abs/1712.06160).

Roncalli, T. and Weisang, G. (2009). Tracking problems, hedge fund replication and alternative beta. Technical report. (https://ssrn.com/abstract=1325190).

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022). Efficient aggregated kernel tests using incomplete U-statistics. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18793–18807.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31.

Song, L., Zhang, X., Smola, A., Gretton, A., and Schölkopf, B. (2008). Tailoring density estimation via reproducing kernel moment matching. In *International Conference on Machine Learning (ICML)*, page 992–999.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer.

Szabó, Z. and Sriperumbudur, B. (2018). Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29.

Székely, G. and Rizzo, M. (2004). Testing for equal distributions in high dimension. *InterStat*, 5:1249–1272.

Székely, G. and Rizzo, M. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80.

Tolstikhin, I., Sriperumbudur, B., and Schölkopf, B. (2016). Minimax estimation of maximal mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1930–1938.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer.

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

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