Bernoulli **28**(1), 2022, 181–213
https://doi.org/10.3150/21-BEJ1338

Finite sample properties of parametric MMD estimation: Robustness to misspecification and dependence

BADR-EDDINE CHÉRIEF-ABDELLATIF[1] and PIERRE ALQUIER[2]

[1]*Department of Statistics, University of Oxford. E-mail: badr-eddine.cherief-abdellatif@stats.ox.ac.uk*
[2]*RIKEN AIP, Tokyo, Japan. E-mail: pierrealain.alquier@riken.jp*

Many works in statistics aim at designing a universal estimation procedure, that is, an estimator that would converge to the best approximation of the (unknown) data generating distribution in a model, without any assumption on this distribution. This question is of major interest, in particular because the universality property leads to the robustness of the estimator. In this paper, we tackle the problem of universal estimation using a minimum distance estimator presented in (Briol et al. (2019)) based on the Maximum Mean Discrepancy. We show that the estimator is robust to both dependence and to the presence of outliers in the dataset. Finally, we provide a theoretical study of the stochastic gradient descent algorithm used to compute the estimator, and we support our findings with numerical simulations.

Keywords: Minimum distance estimation; kernel methods; universal estimation; robust statistics; RKHS; weak dependence

1. Introduction

One of the main challenges in statistics is the design of a *universal* estimation procedure. Given data, a universal procedure is an algorithm that provides an estimator of the generating distribution which is simultaneously statistically consistent when the true distribution belongs to the model, and robust otherwise. Typically, a universal estimator is consistent for any model, with minimax-optimal or fast rates of convergence and is robust to small departures from the model assumptions [10] such as sparse instead of dense effects or non-Gaussian errors in high dimensional linear regression. Unfortunately, most statistical procedures are based upon strong assumptions on the model or on the corresponding parameter set, and very famous estimation methods such as maximum likelihood estimation (MLE), method of moments or Bayesian posterior inference may fail even on simple problems when such assumptions do not hold. For instance, even though MLE is consistent and asymptotically normal with optimal rates of convergence in parametric estimation under suitable regularity assumptions [62, 91] and in nonparametric estimation under entropy conditions, this method behaves poorly in case of misspecification when the true generating distribution of the data does not belong to the chosen model.

Let us investigate a simple example presented in [12] that illustrates the non-universal characteristic of MLE. We observe a collection of n independent and identically distributed (i.i.d.) random variables $X_1, \dots, X_n$ that are distributed according to some mixture distribution $P_n^0 = (1-2n^{-1})\mathcal{U}([0,1/10]) + 2n^{-1}\mathcal{U}([1/10,9/10])$ where $\mathcal{U}([a,b])$ is the uniform distribution between a and b. We consider the parametric model of independent uniform distributions $\mathcal{U}([0,\theta])$, $0 \le \theta < 1$, and we choose the squared Hellinger distance $h^2(\cdot,\cdot)$ as the risk measure. Here the maximum likelihood is the maximum of the observations $X_{(n)} := \max(X_1, \dots, X_n)$, and $\mathcal{U}([0,1/10])$ is a good approximation of the generating distribution P_n^0 as $h^2(P_n^0, \mathcal{U}([0,1/10])) < 5/4n$ for $n \ge 4$. Hence, one would expect that $\mathbb{E}[h^2(P_n^0, \mathcal{U}([0,X_{(n)}]))]$ goes to 0 as $n \to +\infty$, which is actually not the case. We do not even have

consistency: $\mathbb{E}[h^2(P_n^0, \mathcal{U}([0, X_{(n)}]))] > 0.38$. Hence, the MLE is not robust to this small deviation from the parametric assumption. The same happens in Bayesian statistics: the regular posterior distribution is not always robust to model misspecification. Indeed, authors of [7,49] show pathologic cases where the posterior does not concentrate to the true distribution.

Universal estimation is all the more important since it provides a generic approach to tackle the more and more popular problem of robustness to outliers under the i.i.d. assumption, although definitions and goals involved in robust statistics are quite different from the universal estimation perspective. Hüber introduced a framework that models situations where a small fraction ε of data is contaminated, and he assumes that the true generated distribution can be written $(1-\varepsilon)P_{\theta_0} + \varepsilon Q$ where Q is the contaminating distribution and ε is the proportion of corrupted observations [57]. The goal when using this approach is to estimate the true parameter θ_0 given a misspecified model $\{P_\theta / \theta \in \Theta\}$ with $\theta_0 \in \Theta$. A procedure is then said to be robust in this case if it leads to a good estimation of the true parameter θ_0. More generally, when a procedure is able to provide a good estimate of the generating distribution of i.i.d. data when a small proportion of them is corrupted, whatever the values of these outliers, then such an estimator is considered as robust.

Interestingly enough, none of the aforementioned works questioned the independence assumption on the observation. We believe that a universal estimation procedure should still produce sensible estimations under small deviations from this assumption.

1.1. Related work

Several authors attempted to design a general universal estimation method. Sture Holm [10] suggested that Minimum Distance Estimators (MDE) were the most natural procedures being robust to misspecification. Motivated by [81,92], MDE consists in minimizing some probability distance d between the empirical distribution and a distribution in the model. The MDE $\hat{\theta}_n$ is defined by:

$$d(\hat{P}_n, P_{\hat{\theta}_n}) = \inf_{\theta \in \Theta} d(\hat{P}_n, P_\theta)$$

where $\hat{P}_n = n^{-1} \sum_{i=1}^n \delta_{\{X_i\}}$ is the empirical measure and Θ the parameter set associated to the model. If the minimum does not exist, then one can consider an ε-approximate solution. In fact, this minimum distance estimator is used in many usual procedures. Indeed, the generalized method of moments [51] is actually defined as minimizing the weighted Euclidean distance between moments of $\hat{P}_n$ and P_θ while the MLE minimizes the KL divergence, at least for discrete measures. When the distance d is wisely chosen, for example, when it is bounded, then MDE can be robust and consistent.

A popular metric is the Total Variation (TV) distance [35,94]. [94] built an estimator that is uniformly consistent in TV distance and is robust to misspecification under the i.i.d. assumption, but without any assumption on the true distribution of the data. The rate of convergence depends on the Kolmogorov entropy of the model. A few decades later, Devroye and Lugosi studied in details the skeleton estimate, a variant of the estimator of [94] that is based on the TV-distance restricted to the so-called Yatracos sets, see [35]. Unfortunately, the skeleton estimate and the original Yatracos estimate are not computationally tractable.

In [5] and [6], Baraud, Birgé and Sart introduced ρ-estimation, a universal method that retains some appealing properties of the MLE such as efficiency under some regularity assumptions, while being robust to deviations, measured by the Hellinger distance. This ρ-estimation procedure is inspired from T-estimation [12], itself inspired from earlier works of Le Cam [63,64] and Birgé [11], and goes beyond the classical compactness assumption used in T-estimation. In compact models, ρ-estimators can be seen as variants of T-estimators also based on robust tests, but they can be extended to noncompact

models such as linear regression with fixed or random design with various error distributions. As T-estimators, they enjoy robustness properties, but involve other metric dimensions which lead to optimal rates of convergence with respect to the Hellinger distance even in cases where T-estimators can not be defined. Moreover, when the sample size is large enough, ρ-estimation recovers the usual MLE in density estimation when the model is parametric, well-specified and regular enough. Hence, ρ-estimation can be seen as a robust version of the MLE. Unfortunately, this strategy is also intractable. The Wasserstein distance became recently extremely popular. Some attempts to obtain universal estimation with the Wasserstein distance can be found in [8,67].

More recently, [15] showed that using the Maximum Mean Discrepancy (MMD) [47] to build a minimum distance estimator leads to both robust estimation in the i.i.d. case, without any assumption on the model $\{P_\theta, \theta \in \Theta\}$. Moreover, this estimator is tractable as soon as the model is generative, that is, when one can sample efficiently from any P_θ. MMD, a metric based on embeddings of probability measures into a reproducing kernel Hilbert space, has been applied successfully in a wide range of problems such as kernel Bayesian inference [87], approximate Bayesian computation [80], two-sample [47] and goodness-of-fit testing [59], and MMD GANs [42,68] and autoencoders [95], to name a few prominent examples. Such minimum MMD-based estimators are proven to be consistent, asymptotically normal and robust to model misspecification. The trade-off between the statistical efficiency and the robustness is made through the choice of the kernel. The authors investigated the geometry induced by the MMD on a finite-dimensional parameter space and introduced a (natural) gradient descent algorithm for efficient computation of the estimator. This algorithm is inspired from the stochastic gradient descent (SGD) used in the context of MMD GANs where the usual discriminator is replaced with a two-sample test based on MMD [42]. These results were extended in the Bayesian framework by [26].

Finally, a whole branch of probability and statistics study limit theorems (LLN, CLT) under the assumptions that the data is not exactly independent, but that in some sense, the dependence between the observations is not strong. Since the seminal work of [85], many mixing conditions, that is, restrictions on the dependence between observations, were defined. These conditions lead to limit theorems useful to analyze the asymptotic behavior of estimators computed on time series [39]. Nevertheless, checking mixing assumptions is difficult in practice and many classes of processes that are of interest in statistics such as elementary Markov chains are sometimes not mixing. More recently, [40] proposed a new weak dependence condition for time series that is built on covariance-based coefficients which are much easier to compute than mixing ones, and that is more general than mixing as it stands for most relevant classes of processes. We believe that it is important to study robust estimators in this setting, in order to check that they are also robust from small deviations to the independence assumption.

1.2. Contributions

In this paper, we further investigate universality properties of minimum distance estimation based on MMD distance [15]. Inspired by the related literature, our contributions in this paper are the following:

- We go beyond the classical i.i.d. framework. Indeed, we prove that the the estimator is robust to dependence between observations. To do so, we introduce a new dependence coefficient expressed as a covariance in some reproducing kernel Hilbert space, and which is very simple to use in practice.
- We show that our oracle inequalities imply robust estimation under the i.i.d. assumption in the Hüber contamination model and in the case of adversarial contamination.
- We propose a theoretical analysis of the SGD algorithm used to compute this estimator in [15] and [42] for some finite dimensional models. Thanks to this algorithm, we provide numerical simulations to illustrate our theoretical results.

The first result of this paper is a generalization bound in the non-i.i.d. setting. It states that under a very general dependence assumption, the generalization error with respect to the MMD distance decreases in $n^{-1/2}$ as $n \to +\infty$. This result extends the inequalities in [15] that are only available in the i.i.d. framework, and is obtained using dependence concepts for stochastic processes. We introduce in this paper a new dependence coefficient in the wake of [40] which can be expressed as a covariance in some reproducing kernel Hilbert space associated with MMD. This coefficient can be easily computed in many situations and which may be related to usual mixing coefficients such as the popular β-mixing one. We show that a weak assumption on this new dependence coefficient can relax the i.i.d. assumption of [15] and can lead to valid generalization bounds even in the dependent setting.

Regarding robustness, we prove that our generalization bounds for the MMD estimator implies that this estimator is robust to the presence of outliers. Note that this includes Hüber's type contamination, and adversarial contamination as well. In particular, we compare the rate of convergence of the MMD estimator to the minimax estimators in the example of the estimation of the mean of a Gaussian.

Regarding computational issues, we provide a Stochastic Gradient Descent (SGD) algorithm as in [15,42] involving a U-statistic approximation of the expectation in the formula of the MMD distance. We theoretically analyze this algorithm in parametric estimation using a convex parameter set. We also perform numerical simulations that illustrate the efficiency of our method, especially by testing the behavior of the algorithm in the presence of outliers.

The rest of the paper is organized as follows. Section 2 defines the MMD-based minimum distance estimator and our new dependence coefficient based on the kernel mean embedding. Section 3 provides nonasymptotic bounds in the dependent and misspecified framework, with their implications in terms of robust parametric estimation. Section 4 illustrates the efficiency of our method in several different frameworks. We finally present an SGD algorithm with theoretical convergence guarantees in Section 5 and we perform numerical simulations in Section 6. The proofs of the theorems of Section 3 are provided in Section 7. The supplementary material is dedicated to the remaining proofs [27].

2. Background and definitions

In this section, we first introduce some notations and present the statistical setting of the paper in Section 2.1. Then, we remind in Section 2.2 some theory on reproducing kernel Hilbert spaces (RKHS) and we define both the maximum mean discrepancy (MMD) and our minimum distance estimator based on the MMD. Finally, we introduce in Section 2.3 a new dependence coefficient expressed as a covariance in a RKHS.

2.1. Statistical setting

We shall consider a dependent setting throughout the paper. We observe in a measurable space $(\mathbb{X}, \mathcal{X})$ a collection of n random variables $X_1,\ldots,X_n$ generated from a stationary process. This implies that the X_i's are identically distributed, and we will let P^0 denote their marginal distribution. Note that this include as an example the case where the X_i's are i.i.d. with generating distribution P^0. We introduce a statistical model $\{P_\theta / \theta \in \Theta\}$ indexed by a parameter space Θ.

2.2. Maximum mean discrepancy

We consider a positive definite kernel function k, that is, a symmetric function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ such that for any integer $n \geq 1$, for any $x_1, \ldots, x_n \in \mathbb{X}$ and for any $c_1, \ldots, c_n \in \mathbb{R}$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0.$$

We then consider the reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ associated with the kernel k which satisfies the reproducing property $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}_k}$ for any function $f \in \mathcal{H}_k$ and any $x \in \mathbb{X}$. From now on, we assume that the kernel is bounded by some positive constant, that will be assumed to be 1 without loss of generality. That is, for any $x, y \in \mathbb{X}$, $|k(x, y)| \leq 1$.

Now we introduce the notion of *kernel mean embedding*, a Hilbert space embedding of a probability measure that can be viewed as a generalization of the original feature map used in support vector machines and other kernel methods. Given a probability measure P, we define the mean embedding $\mu_P \in \mathcal{H}_k$ as:

$$\mu_P(\cdot) := \mathbb{E}_{X \sim P}\big[k(X, \cdot)\big] \in \mathcal{H}_k.$$

All the applications and the theoretical properties of those embeddings have been well studied [76]. In particular, the mean embedding μ_P satisfies the relationship $\mathbb{E}_{X \sim P}[f(X)] = \langle f, \mu_P \rangle_{\mathcal{H}_k}$ for any function $f \in \mathcal{H}_k$, and induces a semi-metric [1] on measures called maximum mean discrepancy (MMD), defined for two measures P and Q as follows:

$$\mathbb{D}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}$$

or equivalently

$$\mathbb{D}_k^2(P, Q) = \mathbb{E}_{X, X' \sim P}\big[k\big(X, X'\big)\big] - 2\mathbb{E}_{X \sim P, Y \sim Q}\big[k(X, Y)\big] + \mathbb{E}_{Y, Y' \sim Q}\big[k\big(Y, Y'\big)\big].$$

A kernel k is said to be characteristic if $P \mapsto \mu_P$ is injective. This ensures that $\mathcal{D}_k$ is a metric, and not only a semi-metric. Section 3.3.1 of the thorough survey [76] provides a wide range of conditions ensuring that k is characteristic. They also provide many examples of characteristic kernels, see their Table 3.1. Among others, when $\mathbb{X} \subset \mathbb{R}^d$ equipped with the Euclidean norm $\|\cdot\|$, the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/\gamma^2)$ and the Laplace kernel $k(x, y) = \exp(-\|x - y\|/\gamma)$, are known to be characteristic. We actually mostly use these two kernels in our applications. From now on, we will assume that k is characteristic.

Note that there are many applications of the kernel mean embedding and MMD in statistics such as two-sample testing [47], change-point detection [4], detection [65], we also refer the reader to [69] for a thorough introduction to the applications of kernels and MMD to computationnal biology.

Here, we will focus on estimation of parameters based on MMD. This principle was used to train generative networks [42,68], it's only recently that it was studied as a general principle for estimation [15]. Following these papers, we define the MMD estimator $\hat{\theta}_n$ such that,

$$\mathbb{D}_k(P_{\hat{\theta}_n}, \hat{P}_n) = \inf_{\theta \in \Theta} \mathbb{D}_k(P_\theta, \hat{P}_n)$$

[1] This means that $P \to \|\mu_P\|_{\mathcal{H}_k}$ satisfies the requirements of a norm besides $\|\mu_P - \mu_Q\|_{\mathcal{H}_k} = 0$ only for $\mu_P = \mu_Q$.

where $\hat{P}_n = (1/n)\sum_{i=1}^n \delta_{X_i}$ is the empirical measure, that is,

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\left\{\mathbb{E}_{X,X'\sim P_\theta}\big[k\big(X, X'\big)\big] - \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{X\sim P_\theta}\big[k(X, X_i)\big]\right\}.$$

It could be that there is no minimizer, see the discussion in Theorem 1 page 9 in [15]. In this case, we can use an approximate minimizer. More precisely, for any $\varepsilon > 0$ we can always find a $\hat{\theta}_{n,\varepsilon}$ such that,

$$\mathbb{D}_k(P_{\hat{\theta}_{n,\varepsilon}}, \hat{P}_n) \leq \inf_{\theta\in\Theta} \mathbb{D}_k(P_\theta, \hat{P}_n) + \varepsilon.$$

In what follows, we will consider the case where the minimizer exists (that is, $\varepsilon = 0$) but when this is not the case, everything can be easily extended by considering $\hat{\theta}_{n,1/n}$.

2.3. Covariances in RKHS

In this subsection, we introduce and discuss a new dependence coefficient based on the kernel mean embedding. This coefficient allows to go beyond the i.i.d. case in the study of the MMD estimator of [15], and to show that it is actually robust to dependence.

Definition 2.1. We define, for any $t \in \mathbb{N}$,

$$\varrho_t = \left|\mathbb{E}\left\langle k(X_t, \cdot) - \mu_{P^0}, k(X_0, \cdot) - \mu_{P^0}\right\rangle_{\mathcal{H}_k}\right|.$$

In the i.i.d. case, note that $\varrho_t = 0$ for any $t \geq 1$. In general, the following assumption will ensure the consistency of our estimator.

Assumption 2.1. There is a $\Sigma < +\infty$ such that, for any n, $\sum_{t=1}^n \varrho_t \leq \Sigma$.

Our mean embedding dependence coefficient may be seen as a covariance expressed in the RKHS $\mathcal{H}_k$. We shall see throughout the paper that the kernel mean embedding coefficient ϱ_t can be easily computed in many situations, and that it is closely related to widely used mixing coefficients. In particular, we will show in Section 4.2 that our coefficient ϱ_t is upper-bounded by the popular β-mixing coefficient. For the reader who would not be familiar with β-mixing, we also show that any real-valued auto-regressive process $X_t = aX_{t-1} + \varepsilon_t$ satisfies Assumption 2.1 as long as $|a| < 1$, the ε_t are i.i.d. and $\mathbb{E}(|\varepsilon_0|) < \infty$. Also, we show that some special cases of such auto-regressive processes are not β-mixing, which proves that Assumption 2.1 is more general than β-mixing: an explicit example is given in Section 4.3. Hence, Assumption 2.1 may be referred to as a weak dependence condition in the wake of the concept of weak dependence introduced in [40]. We will show in the next section that under Assumption 2.1, we can obtain a nonasymptotic generalization bound of the same order than in the i.i.d. case.

3. Nonasymptotic bounds in the dependent, misspecified case

In this section, we provide nonasymptotic generalization bounds in MMD distance for the minimum MMD estimator. In particular, we show in Section 3.1 that under a weak dependence assumption, it is robust to both dependence and misspecification, and is consistent at the same $n^{-1/2}$ rate than in the i.i.d. case. In particular, we give explicit bounds in the Hüber contamination model and in a more general adversarial setting in Section 3.2.

3.1. Estimation with respect to the MMD distance

First, we begin with a theorem that gives an upper bound on the generalization error, that is, the expectation of $\mathbb{D}_k(P_{\hat{\theta}_n}, P^0)$. The rate of convergence of this error is of order $n^{-1/2}$ independently of the dimension of the parameter space Θ. In fact, note that there is actually no assumption at all on the model $\{P_\theta, \theta \in \Theta\}$ in this theorem.

Theorem 3.1. *We have*:

$$\mathbb{E}\big[\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big)\big] \leq \inf_{\theta\in\Theta} \mathbb{D}_k\big(P_\theta, P^0\big) + 2\sqrt{\frac{1 + 2\sum_{t=1}^n \varrho_t}{n}}.$$

As a consequence, under Assumption 2.1:

$$\mathbb{E}\big[\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big)\big] \leq \inf_{\theta\in\Theta} \mathbb{D}_k\big(P_\theta, P^0\big) + 2\sqrt{\frac{1 + 2\Sigma}{n}}.$$

We remind that the proofs of the results in this section are deferred to Section 7. It is also possible to provide a result that holds with large probability as in [15,42]. Naturally, this requires stronger assumptions, and the conditions on the dependence become more intricate in this case. Here, we use a condition introduced in [82,83] for generic metric spaces that we adapt to the kernel embedding and to stationarity.

Assumption 3.1. Assume that there is a family $(\gamma_\ell)_\ell$ of non-negative numbers such that, for any integer n, for any $\ell \in \{1, \dots, n-1\}$ and any function $g : \mathcal{H}_k^\ell \to \mathbb{R}$ such that

$$\big|g(a_1, \dots, a_\ell) - g(b_1, \dots, b_\ell)\big| \leq \sum_{i=1}^{\ell} \|a_i - b_i\|_{\mathcal{H}_k},$$

we have: $|\mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}}, \dots, \mu_{\delta_{X_n}})|X_1, \dots, X_\ell] - \mathbb{E}[g(\mu_{\delta_{X_{\ell+1}}}, \dots, \mu_{\delta_{X_n}})]| \leq \gamma_1 + \cdots + \gamma_{n+\ell-1}$, almost surely. Assume that $\Gamma := \sum_{\ell\geq 1} \gamma_\ell < \infty$.

This assumption is more technical than Assumption 2.1. The idea is quite similar: the coefficient γ_s is a measure of the dependence between X_t and X_{t+s}, and the assumption will be satisfied if X_t and X_{t+s} are "almost independent" when s is large – but the sense given to "almost independent" is not exactly the same as in Assumption 2.1. For example, we show in Section 4.3 that auto-regressive processes $X_{t+1} = aX_t + \varepsilon_{t+1}$ with $|a| < 1$ and i.i.d. ε_t satisfy this assumption under the additional condition that the ε_t are almost surely bounded. Again, note that in the case of independence, we can take all the $\gamma_i = 0$ and hence $\Gamma = 0$ in addition to $\Sigma = 0$. We can now state our result in probability.

Theorem 3.2. *Assume that Assumptions 2.1 and 3.1 are satisfied. Then, for any* $\delta \in (0, 1)$,

$$\mathbb{P}\left[\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big) \leq \inf_{\theta\in\Theta} \mathbb{D}_k\big(P_\theta, P^0\big) + 2\frac{\sqrt{1 + 2\Sigma} + (1 + \Gamma)\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}\right] \geq 1 - \delta.$$

Assumption 3.1 is fundamental to obtain a result in probability. Indeed, the rate of convergence in Theorem 3.2 is characterized by some concentration inequality upper bounding the MMD distance between the empirical and the true distribution as done in [15]. Nevertheless, the proof of this inequality

in [15] is based on a Hoeffding-type inequality known as McDiarmid's inequality [74] that is only valid for independent variables (that is, all the $\gamma_i = 0$), which makes this inequality not applicable in our dependent setting. Hence, we use a version of McDiarmid's inequality for time series obtained by Rio [82,83] which is available under the assumption that $\sum_{\ell \geq 1} \gamma_\ell < \infty$ (Assumption 3.1).

Remark 3.1 (The i.i.d. case). Note that when the X_i's are i.i.d., Assumptions 2.1 and 3.1 are always satisfied with $\Sigma = \Gamma = 0$ and thus Theorem 3.1 gives simply

$$\mathbb{E}\big[\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big)\big] \leq \inf_{\theta \in \Theta} \mathbb{D}_k\big(P_\theta, P^0\big) + \frac{2}{\sqrt{n}}$$

while Theorem 3.2 gives

$$\mathbb{P}\left[\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big) \leq \inf_{\theta \in \Theta} \mathbb{D}_k\big(P_\theta, P^0\big) + 2\frac{1 + \sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}\right] \geq 1 - \delta.$$

Remark 3.2 (Connection between the MMD distance and the L^2 norm). In Section 4, we study the connection between the convergence of $\hat{P}_{\hat{\theta}_n}$ in terms of MMD distance, and the convergence of $\hat{\theta}_n$, is some classical models. However, it is also worth mentioning a connection between the MMD distance and the quadratic distance on densities. Indeed, assume $\mathbb{X} = \mathbb{R}^d$ and that P and Q have density p and q, respectively with respect to the Lebesgue measure. Using the Gaussian kernel $k_\gamma(x, y) = \exp(-\|x - y\|^2/\gamma^2)$, we expect that, when $\gamma \to 0$, under suitable assumptions,

$$\mathbb{E}_{X \sim P, Y \sim Q}\big[k(X, Y)\big] \sim \pi^{\frac{d}{2}} \gamma^d \int p(x) q(x)\, \mathrm{d}x$$

and so that

$$\mathbb{D}_{k_\gamma}(P, Q) \sim \pi^{\frac{d}{4}} \gamma^{\frac{d}{2}} \|p - q\|_{L^2}. \tag{1}$$

Corollary 4 page 1527 of [88] provides a formal statement of this claim. Thus, the convergence in the MMD distance has connections with the convergence of the densities (when they exist) in L^2.

Note that [5,35] argue that the L^2-norm is not suitable for universal estimation: indeed, in some models, P_θ does not have a density with respect to the Lebesgue measure. But (1) allows the interpretation of the MMD distance (with the Gaussian kernel) as an approximation of the L^2 distance, that is however well defined for *any* model (P_θ).

3.2. Robust parametric estimation

3.2.1. *Contamination models*

As explained in the introduction, when all observations but a small proportion of them are sampled independently from a generating distribution P_{θ_0} $(\theta_0 \in \Theta)$, robust parametric estimation consists in finding estimators being both rate optimal and resistant to outliers. Two among the most popular frameworks for studying robust estimation are the so-called Hüber's contamination model and the adversarial contamination model.

Hüber's contamination model is as follows. We observe a collection of random variables $X_1, \ldots, X_n$. We consider a contamination rate $\varepsilon \in (0, 1/2)$, latent i.i.d. random variables $Z_1, \ldots, Z_n \sim \mathrm{Ber}(\varepsilon)$ and some noise distribution Q, such that the distribution of X_i given $Z_i = 0$ is P_{θ_0}, and that the distribution of X_i given $Z_i = 1$ is Q. Hence, the observations X_i's are independent and sampled from the mixture $P^0 = (1 - \varepsilon)P_{\theta_0} + \varepsilon Q$.

The adversarial model is more general. Contrary to Hüber's contamination where outliers were all sampled from the contaminating distribution, we do not make any particular assumption on the outliers here. Hence, we shall adopt slightly different notations. We assume that $X_1, \dots, X_n$ are identically distributed from P_{θ_0} for some $\theta_0 \in \Theta$. However, the statistician only observes $\tilde{X}_1, \dots, \tilde{X}_n$ where $\tilde{X}_i$ can be any arbitrary value for $i \in \mathcal{O}$, where $\mathcal{O}$ is an arbitrary set subject to the constraint $|\mathcal{O}| \leq \varepsilon n$, and $\tilde{X}_i = X_i$ for $i \notin \mathcal{O}$. The estimators are built based on these observations $\tilde{X}_1, \dots, \tilde{X}_n$.

3.2.2. *Literature*

One hot research trend in robust statistics is focused on the search of both statistically optimal and computationally tractable procedures for the Gaussian mean estimation problem $\{P_\theta = \mathcal{N}(\theta, I_d)/\theta \in \mathbb{R}^d\}$ in the presence of outliers under the i.i.d. assumption, which remains a major challenge. Usual robust estimators such as the coordinatewise median and the geometric median are known to be suboptimal in this case, and there is a need to look at more complex estimators such as Tukey's median that achieves the minimax optimal rate of convergence $\max(\frac{d}{n}, \varepsilon^2)$ with respect to the squared Euclidean distance, where d is the dimension, n is the sample size and ε is the proportion of corrupted data. Unfortunately, computation of Tukey's median is not tractable and even approximate algorithms lead to an $\mathcal{O}(n^d)$ complexity [3,22]. This has led to the rise of the recent studies in robust statistics which address how to build robust and optimal statistical procedures, in the wake of the works of [90] and [57], but that are also computationally efficient.

This research area started with two seminal works presenting two procedures for the normal mean estimation problem: the *iterative filtering* [36] and the *dimension halving* [61]. These algorithms are based upon the idea of using higher moments in order to obtain a good robust moment estimation, and are minimax optimal up to a poly-logarithmic factor in polynomial time. This idea was then used in several other problems in robust statistics, for instance in sparse functionals estimation [41], clustering [60], mixtures of spherical Gaussians learning [37], and robust linear regression [38]. In Hüber's contamination model, [29] achieves the minimax rate without any extra factor in the $\varepsilon = \mathcal{O}(\min(d^{-1/2}, n^{-1/4}))$ regime with an improved overall complexity. Meanwhile, [43] offers a different perspective on robust estimation and connects the robust normal mean estimation problem with Generative Adversarial Networks (GANs) [9,46], what enables computing robust estimators using efficient tools developed for training GANs. Hence, the authors compute depth-like estimators that retain the same appealing robustness properties than Tukey's median and that can be trained using stochastic gradient descent (SGD) algorithms that were originally designed for GANs.

Another popular approach for the more general problem of mean estimation under the i.i.d. assumption in the presence of outliers is the study of finite-sample sub-Gaussian deviation bounds. Indeed, designing estimators achieving sub-Gaussian performance under minimal assumptions ensures robustness to outliers that are inevitably present when the generating distribution is heavy-tailed. In the univariate case, some estimators present a sub-Gaussian behavior for all distributions under first and second order moments. A simple but powerful strategy, the Median-of-Means (MOM), dates back to [1,58,78]. This method consists in randomly splitting the data into several equal-size blocks, then computing the empirical mean within each block, and finally taking the median of them. Most MOM-based procedures lead to estimators that are simultaneously statistically optimal [28,34,65,66,73] and computationally efficient [24,33,54]. Moreover, this approach can be easily extended to the multivariate case [56,75]. An important advantage is that the MOM estimator has good performance even for distributions with infinite variance. An elegant alternative to the MOM strategy is due to Catoni, whose estimator is based on PAC-Bayesian truncation in order to mitigate heavy tails [20]. It has the same performance guarantees than the MOM method but with sharper and near-optimal constants. In [21], Catoni and Giulini proposed a very simple and trivial-to-compute multidimensional extension of Catoni's M-estimator defined as an empirical average of the data, with the observations with large norm

shrunk towards zero, and that still satisfies a sub-Gaussian concentration using PAC-Bayes inequalities. The influence function of Catoni and Giulini has been widely used since then, see [44,45,50,52,53]. We refer the reader to the beautiful review of [72] for more details on those mean estimation procedures.

3.2.3. *Robust MMD estimation*

In this section, we show the properties of our MMD-based estimator in robust parametric estimation with outliers, both in Hüber's contamination model and in the adversarial case. Our bounds are obtained by working directly in the RKHS rather than in the parameter space. the consequence of these results in terms of the Euclidean distance in the parameter space will be explored in Section 4.

First, we consider Hüber's contamination model [57]. The objective is to estimate P_{θ_0} by observing contaminated random variables $X_1,\dots, X_n$ with actual distribution is $P^0 = (1-\alpha)P_{\theta_0} + \alpha Q$ for some Q, and some $0 \leq \alpha \leq \varepsilon$. We state the key following lemma.

Lemma 3.3. *We have, for any* $\theta \in \Theta$, $|\mathbb{D}_k(P_\theta, P^0) - \mathbb{D}_k(P_\theta, P_{\theta_0})| \leq 2\varepsilon$.

Applying Lemma 3.3 to the left-hand side, and to the right-hand side, of Theorem 3.1, we have the following result.

Corollary 3.4. *Assume that* $X_1, \dots, X_n$ *are identically distributed from* $P^0 = (1-\alpha)P_{\theta_0} + \alpha Q$ *for some* $\theta_0 \in \Theta$, *some* Q, *with* $0 \leq \alpha \leq \varepsilon$. *Then*

$$\mathbb{E}\big[\mathbb{D}_k(P_{\hat{\theta}_n}, P_{\theta_0})\big] \leq 4\varepsilon + 2\sqrt{\frac{1 + 2\sum_{t=1}^n \varrho_t}{n}}.$$

If moreover we assume that Assumptions 2.1 and 3.1 are satisfied, then for any $\delta \in (0,1)$,

$$\mathbb{P}\left[\mathbb{D}_k(P_{\hat{\theta}_n}, P_{\theta_0}) \leq 2\left(2\varepsilon + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}\right)\right] \geq 1-\delta.$$

We obtain a rate $\max(1/\sqrt{n}, \varepsilon)$ in MMD distance (note once again that the convergence rate with respect to more standard distances is studied in Section 4). When $\varepsilon \lesssim 1/\sqrt{n}$, then we recover the rate of convergence without contamination, and when $1/\sqrt{n} \lesssim \varepsilon$, then the rate is dominated by the contamination ratio ε. Hence, the maximum number of outliers which can be tolerated without breaking down the rate is $n\varepsilon \asymp \sqrt{n}$.

This result can also be extended to the adversarial contamination setting, where no assumption is made on the outliers.

Proposition 3.5. *Assume that* $X_1, \dots, X_n$ *are identically distributed from from* $P^0 = P_{\theta_0}$ *for some* $\theta_0 \in \Theta$. *However, the statistician only observes* $\tilde{X}_1, \dots, \tilde{X}_n$ *where* $\tilde{X}_i$ *can be any arbitrary value for* $i \in \mathcal{O}$, $\mathcal{O}$ *is any arbitrary set subject to the constraint* $|\mathcal{O}| \leq \varepsilon n$, *and* $\tilde{X}_i = X_i$ *for* $i \notin \mathcal{O}$ *and builds the estimator* $\tilde{\theta}_n$ *based on these observations*:

$$\mathbb{D}_k\left(P_{\tilde{\theta}_n}, \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{X}_i}\right) = \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, \frac{1}{n}\sum_{i=1}^n \delta_{\tilde{X}_i}\right).$$

Then:

$$\mathbb{D}_k(P_{\tilde{\theta}_n}, P_{\theta_0}) \leq 4\varepsilon + 2\mathbb{D}_k(P_{\hat{\theta}_n}, P_{\theta_0}).$$

Thus

$$\mathbb{E}\big[\mathbb{D}_k(P_{\tilde{\theta}_n}, P_{\theta_0})\big] \le 4\varepsilon + 4\sqrt{\frac{1 + 2\sum_{t=1}^{n} \varrho_t}{n}}$$

and, under Assumptions 2.1 and 3.1, for any $\delta \in (0, 1)$,

$$\mathbb{P}\left[\mathbb{D}_k(P_{\hat{\theta}_n}, P_{\theta_0}) \le 4\left(\varepsilon + \frac{\sqrt{1 + 2\Sigma} + (1 + \Gamma)\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}\right)\right] \ge 1 - \delta.$$

One can see that the rate of convergence we obtain without making any assumption on the outliers is exactly the same than in Hüber's contamination model. The only difference is that the constant in the right hand side of the inequality is tighter in Hüber's contamination model.

4. Examples

4.1. Independent observations

In the previous section, we studied the convergence of $P_{\hat{\theta}_n}$ with the MMD distance. In this subsection, we show what are the consequences of these results in terms of the convergence of $\hat{\theta}$ in some classical models. For the sake of simplicity, we focus on i.i.d. observations. That is, $\varrho_t = 0$ for any $t \ge 1$. Moreover, we will only use the Gaussian kernel $k_\gamma(x, y) = \exp(-\|x - y\|^2/\gamma^2)$.

4.1.1. *Estimation of the mean in a Gaussian model*

Here, $\mathbb{X} = \mathbb{R}^d$ and we are interested in the estimation of the mean in a Gaussian model. For the sake of simplicity, we assume that the variance is known.

Proposition 4.1. *Assume that* $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ *for* $\theta \in \Theta = \mathbb{R}^d$. *Moreover, assume that we are in an adversarial contamination model where a proportion at most* ε *of the observations is contaminated. Then, with probability* $1 - \delta$,

$$\|\theta - \tilde{\theta}_n\|^2 \le -\big(4\sigma^2 + \gamma^2\big)\log\left\{1 - 8\mathrm{e}^{\frac{2\sigma^2 d}{\gamma^2}}\left(\varepsilon + \frac{1 + \sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}\right)^2\right\}. \tag{2}$$

In particular, the choice $\gamma = \sigma\sqrt{2d}$ *leads to*

$$\|\tilde{\theta}_n - \theta_0\|^2 \le -2\sigma^2(d + 2)\log\left[1 - 8\mathrm{e}\left(\varepsilon + \frac{1 + \sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}\right)^2\right].$$

The complete proof can be found in the supplementary material. Note that when ε is small and n is large,

$$\|\tilde{\theta}_n - \theta_0\|^2 \le -2\sigma^2(d + 2)\log\left[1 - 16\mathrm{e}\left(\varepsilon^2 + \frac{(1 + \sqrt{2\log(\frac{1}{\delta})})^2}{n}\right)\right]$$

$$\sim 32\mathrm{e}\sigma^2(d + 2)\left(\varepsilon^2 + \frac{(1 + \sqrt{2\log(\frac{1}{\delta})})^2}{n}\right).$$

We can see that our MMD estimator achieves a rate of convergence $d\varepsilon^2 + d/n$ which is the same than for several median-based estimators such as the geometric median or the coordinatewise median (see Proposition 2.1 in [23]). We have a quadratic dependence in ε, contrary to many robust estimators such as Median-of-Means which dependence in ε is linear. Hence, as soon as the dimension is no larger than the square root of the sample size $d \leq \sqrt{n}$, the MMD method tolerates a larger number of outliers without affecting the usual rate of convergence (i.e., the rate with no contamination).

Unfortunately, it seems that our method performs poorly compared to such estimators in large dimension. Indeed, according to Theorems 2.1 and 2.2 in [23], the minimax optimal rate with respect d, ε and n is $\varepsilon^2 + d/n$. Furthermore, numerical experiments and the investigation conducted for the population limit case when one has access to infinitely many samples in [93] (that has been published since the first version of this paper) suggest that the MMD estimator can not match the minimax rate of convergence. Nevertheless, this non-optimality in the minimax sense does not necessarily imply inaccurate mean estimation in general, and MMD can still lead to efficient estimation in most contamination scenarios.

To understand why the MMD estimator can not match the minimax rate of convergence in high dimension, and why this is not necessarily a problem, we need to analyze the landscape of the optimization program.

Let us first investigate the population limit case, where we do not work with the MMD distance to the empirical distribution $\hat{P}_n$ but to the true distribution $(1-\varepsilon)\mathcal{N}(\theta_0, \sigma^2 I_d) + \varepsilon Q$, as if we had access to infinitely many samples, and with a point-mass delta Dirac contamination $Q = \delta_{\{\theta_c\}}$. The optimization program is, for any value of γ,

$$\min_{\theta\in\mathbb{R}^d} \mathbb{D}_{k_\gamma}\left(P_\theta, (1-\varepsilon)\mathcal{N}\left(\theta_0, \sigma^2 I_d\right) + \varepsilon\delta_{\{\theta_c\}}\right)$$
$$= \max_{\theta\in\mathbb{R}^d}\left\{(1-\varepsilon)\exp\left(-\frac{\|\theta-\theta_0\|^2}{\gamma^2+4\sigma^2}\right) + \varepsilon\left(\frac{\gamma^2+4\sigma^2}{\gamma^2+2\sigma^2}\right)^{d/2}\exp\left(-\frac{\|\theta-\theta_c\|^2}{\gamma^2+2\sigma^2}\right)\right\}.$$

Even though the objective function is nonconvex in θ, it is easy to see that the solution belongs to the line between θ_0 and θ_c. More precisely, if θ_0 and θ_c are far from each other, then the solution is simply θ_0. At the opposite, if θ_0 and θ_c are closed, then the solution will be very close to θ_0. In the situation in between where $\|\theta_0 - \theta_C\|^2 \approx d$, then it is proven in [93] that the solution is at least $\varepsilon\sqrt{d}$ far from the true parameter θ_0, which explains the term $d\varepsilon^2$ in the rate of convergence of the MMD estimator. Hence, we understand that the worst-case rate of the MMD estimator does not correspond to cases where θ_c is far from θ_0 but to cases where the distance is quite large in high dimensions only (of order $\sqrt{d}$).

The previous reasoning can be easily generalized to the MMD estimator with a finite sample. In this situation with $Q = \delta_{\{\theta_c\}}$, the optimization program can be written, denoting by $\mathcal{O}$ the set of outliers,

$$\max_{\theta\in\mathbb{R}^d}\left\{\sum_{i\notin\mathcal{O}}\exp\left(-\frac{\|\theta-X_i\|^2}{\gamma^2+2\sigma^2}\right) + |\mathcal{O}|\exp\left(-\frac{\|\theta-\theta_c\|^2}{\gamma^2+2\sigma^2}\right)\right\},$$

and the solution belongs to the convex hull of the set of points composed of the (random) inliers in the random variables $X_1, \ldots, X_n$ and of the contamination point θ_c. A remarkable point in high dimensional probability is that samples from a multivariate standard Gaussian distribution are concentrated on the sphere of radius $\sqrt{d}$ centered at θ_0, which means that the typical distance $\|X_i - \theta_0\|$ of a datapoint X_i from the mean θ_0 is roughly $\sqrt{d}$. Then, if the contamination is such that $\|\theta_0 - \theta_c\|^2 \approx d$, the outliers lie at a distance $\sqrt{d}$ from θ_0 without being detected, and thus look harmless but shift the mean by approximately $\sqrt{d}\varepsilon$, see Figure 1.

Hence, perhaps counter-intuitively at first sight, the worst contamination does not come from a value of θ_c that is very far away from θ_0 (in which case the estimation will simply be the mean of the inliers),

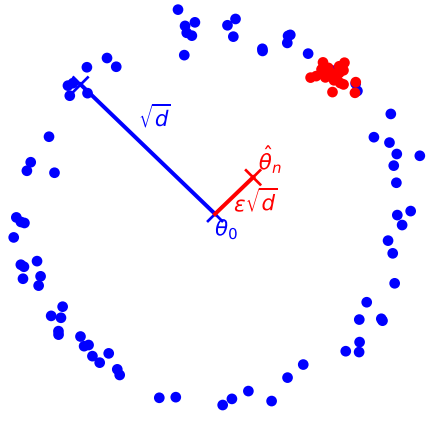

Figure 1. Illustration of the behaviour of the MMD estimator in the high-dimensional Gaussian mean estimation problem. The true parameter θ_0 and datapoints sampled from the true distribution $\mathcal{N}(\theta_0, I_d)$ are colored in blue. Outliers and the MMD estimator $\hat{\theta}_n$ are colored in red. We can see that outliers lying at a distance $\sqrt{d}$ are not detected and shift the mean by $\varepsilon\sqrt{d}$.

but that is only $\sqrt{d}$ away from θ_0, and hence there is mainly one "worst-case contamination" that explains the non-optimality in the minimax sense. Figure 1a of [93] even seems to show that the error of the MMD estimator when γ is of order $\sqrt{d}$ increases with $\|\theta_0 - \theta_c\|$ until achieving $\sqrt{d}$, and then decreases. The same applies to a Gaussian contamination with a small variance.

4.1.2. *Cauchy model*

Here, $\mathbb{X} = \mathbb{R}$ and $P_\theta = \mathcal{C}(\theta, 1)$ where $\mathcal{C}(\theta, s)$ has density $1/[\pi s(1 + (x-\theta)^2/s^2)]$.

Proposition 4.2. *Assume that $P_\theta = \mathcal{C}(\theta, 1)$ for $\theta \in \Theta = \mathbb{R}$. Moreover, assume that we are in an adversarial contamination model where a proportion at most ε of the observations is contaminated. Then, taking $\gamma = 2$ leads to, for any $\delta > 0$,*

$$(\tilde{\theta}_n - \theta_0)^2 \leq 4\left(1 - \frac{1}{1 - 96\pi(\varepsilon^2 + \frac{2+4\log(1/\delta)}{n})}\right).$$

Note that

$$(\tilde{\theta}_n - \theta_0)^2 \leq 4\left(1 - \frac{1}{1 - 128\pi(\varepsilon^2 + \frac{2+4\log(1/\delta)}{n})}\right) \sim 512\pi\left(\varepsilon^2 + \frac{2 + 4\log(1/\delta)}{n}\right).$$

4.1.3. *Estimation with a dictionary*

We consider here estimation of P^0 by a linear combination of measures in a dictionary. This framework actually appears in various models:

- first, when the dictionary contains probability distributions, this is simply a mixture of known components. In this case, the linear combination is actually a convex combination. This context is for example studied in [30];
- assuming that P^0 has a density, in nonparametric density estimation, we can use this setting, the dictionary being defined by a basis of L_2. This is for example, the point of view in [2,16,17].

We will here focus on the first setting, but an extension to the second one is quite straightforward. Let $\{\Phi_1, \dots, \Phi_D\}$ be a family of probability measures over $\mathbb{X} = \mathbb{R}^d$. For $1 \leq i \leq D$ we remind that

$$\mu_{\Phi_i}(\cdot) = \int k(x, \cdot)\Phi_i(\mathrm{d}x).$$

Define the measure $P_\theta = \mathcal{D}(\theta; \Phi_1, \dots, \Phi_D) = \sum_{i=1}^D \theta_i \Phi_i$, and define the model $\{P_\theta, \theta \in \Theta\}$ with $\Theta \subseteq \mathcal{S}_D = \{\theta \in \mathbb{R}_+^D : \sum_{i=1}^D \theta_i = 1\}$. The estimator is then

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} \left\| \sum_{\ell=1}^D \theta_\ell \mu_{\Phi_\ell}(\cdot) - \mu_{\hat{P}_n} \right\|_{\mathcal{H}_k}^2.$$

An application of Theorem 3.2 leads to the following proposition.

Proposition 4.3. *Assume that $P_\theta = \sum_{i=1}^D \theta_i \Phi_i$ where Φ_i is a probability distribution. Define the matrix $G_\gamma = (\langle \mu_{\Phi_i}, \mu_{\Phi_j} \rangle_{\mathcal{H}_{k_\gamma}})_{1 \leq i,j \leq D}$. Letting $\lambda_{\min}(\cdot)$ denote the smallest eigenvalue of a symmetric matrix, we have*:

$$\mathbb{P}\left[\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big) \leq \inf_{\theta \in \Theta} \mathbb{D}_k\big(P_\theta, P^0\big) + 2\frac{1 + \sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}} \right] \geq 1 - \delta$$

and, in the well specified case where $P^0 = P_{\theta_0}$,

$$\mathbb{P}\left[\|\hat{\theta} - \theta_0\|^2 \leq 2\frac{1 + \sqrt{2\log(\frac{1}{\delta})}}{\lambda_{\min}(G_\gamma)\sqrt{n}} \right] \geq 1 - \delta.$$

4.2. β-mixing observations

We now consider non-independent random variables: as in the general framework presented above, $(X_t)_{t \in \mathbb{Z}}$ is a strictly stationary time series, with stationary distribution P^0, and we observe $X_1, \dots, X_n$. We will exhibit some condition on the dependence of the X_i's ensuring that we can still estimate P^0 with the MMD method.

There is a very rich literature on limit theorems and exponential inequalities under conditions on various dependence coefficients. Mixing coefficients and their applications are detailed in the monographs [39,84], weak dependence coefficients in [31]. In this subsection, we show that our coefficient ϱ_t can be upper-bounded by the β-mixing coefficients. So for any β-mixing process, the estimation of P^0 using MMD remains possible. We also remind some examples of β-mixing processes. Note that we will show in the next subsection that Theorem 3.1 can be successfully applied to non β-mixing processes.

4.2.1. *β-mixing and coefficients ϱ_t*

We start by a reminder of the definition of the β-mixing coefficients, from page 4 (Chapter 1) in [31].

Definition 4.1. Given two σ-algebras $\mathcal{A}$ and $\mathcal{B}$,

$$\beta(\mathcal{A}, \mathcal{B}) = \frac{1}{2} \sup_{\substack{I,J \geq 1 \\ U_1,\dots,U_I \\ V_1,\dots,V_J}} \sum_{1 \leq i \leq I} \sum_{1 \leq j \leq J} \big| \mathbb{P}(U_i \cap V_j) - \mathbb{P}(U_i)\mathbb{P}(V_j) \big|$$

where $(U_1, \dots, U_I)$ is any partition of $\mathcal{A}$ and $V_1, \dots, V_j$ any partition of $\mathcal{B}$. Put:

$$\beta_t^{(X)} = \beta\big(\sigma(X_0, X_{-1}, \dots), \sigma(X_t, X_{t+1}, \dots)\big).$$

Section 1.5 in [39] provides summability conditions on the $\beta_t^{(X)}$ leading to a law of large numbers and to a central limit theorem. Examples are also discussed.

Example 4.1. Assume in this example that (X_t) is an homogeneous Markov chain given by its transition kernel P and $X_0 \sim \pi$ where $\pi P = \pi$. Assume that there is a $0 < c \leq 1$ and a probability measure Q on $\mathbb{R}^d$ such that, for some integer $r \geq 1$ and for any measurable A, $P^r(x, A) \geq cQ(A)$. Then it is known, see, for example, Theorem 1 page 88 in [39] that

$$\beta_t^{(X)} \leq 2(1-c)^{\frac{t}{r}-1}.$$

We now compare our ϱ coefficients with the β-mixing coefficients.

Proposition 4.4. *Assume that $k(x, y) = F(\|x - y\|)$ were $F(a) = \int_a^\infty f(b)\,\mathrm{d}b$ for some nonnegative continuous function f with $\int_0^\infty f(b)\,\mathrm{d}b = 1$. Then we have*

$$\varrho_t \leq 2\beta\big(\sigma(X_0), \sigma(X_t)\big) \leq 2\beta_t.$$

Note that $k(x, y) = \exp(-\|x - y\|/\gamma)$ and $k(x, y) = \exp(-\|x - y\|^2/\gamma^2)$ for example, trivially work, respectively with $f(b) = \exp(-b/\gamma)/\gamma$ and $f(b) = 2b\exp(-b^2/\gamma^2)/\gamma^2$.

4.2.2. *Application*: *Hidden Markov chains*

Assume here that $(Y_t)_{t\in\mathbb{N}}$ is a Markov chain on $\{1, \dots, d\}$, and that $X_t|(Y_t = i)$ is independent from all the other values $Y_{t'}$ and is drawn in $\mathbb{R}^D$ from a probability measure Φ_i. The Φ_i's are known and $X_1, \dots, X_n$ are observed but the $(Y_t)_{t\in\mathbb{N}}$ are not observed. Note that this is a dependent extension of the mixture model $\mathcal{D}(\theta; \Phi_1, \dots, \Phi_d)$ discussed above. Indeed, we consider this as a case of misspecification: the statistician uses the mixture model $\mathcal{D}(\theta; \Phi_1, \dots, \Phi_d)$ with $\Theta = \mathcal{S}_d$, being not aware that the data is actually not independent.

Letting P denote the transition matrix of Y, we assume that there exists $c > 0$ and an integer $r \geq 0$ such that $P^r(i, j) \geq c/d$ for any $(i, j) \in \{1, \dots, d\}^2$. Then we have $\beta_t^{(Y)} \leq 2(1-c)^{t/r-1}$. This also implies that there is a unique π such that $\pi P = \pi$ and we assume that $Y_0 \sim \pi$. Then the distribution P^0 of each X_t is given by $P^0(x) = \sum_{i=1}^d \pi_i \Phi_i(x)$.

Also, note that

$$\varrho_t = \beta\big(\sigma(X_0), \sigma(X_t)\big) \leq \beta\big(\sigma(X_0, Y_0), \sigma(X_t, Y_t)\big) = \beta\big(\sigma(Y_0), \sigma(Y_t)\big) \leq 2(1-c)^{t/r-1}.$$

So, a direct application of Theorem 3.1 gives:

$$\mathbb{E}\big[\big\|G_k^{-1/2}(\hat{\theta} - \pi)\big\|\big] = \mathbb{E}\big[\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big)\big] \leq 2\sqrt{\frac{1 + (1-c)^{\frac{1}{r}-1}(3+c)}{n[1-(1-c)^{\frac{1}{r}}]}}.$$

Note that we can add a second layer in the process: assume that an opponent is allowed to replace a fraction ε of the X_t, as in Proposition 3.5. This result in the observation of $\tilde{X}_t$ such that $\tilde{X}_t = X_t$ for a proportion $(1 - \varepsilon)$ of the data, and $\tilde{X}_t$ can be anything for the remaining ε. For example, the opponent

can try fo fool the learner, by drawing from the wrong Φ_i. The MMD estimator $\tilde{\theta}$ still satisfies, from Proposition 3.5,

$$\mathbb{E}\big[\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big)\big] \leq 4\varepsilon + 4\sqrt{\frac{1+(1-c)^{\frac{1}{r}-1}(3+c)}{n[1-(1-c)^{\frac{1}{r}}]}}.$$

4.3. Auto-regressive observations

In this section, we provide examples of auto-regressive processes satisfying Assumption 2.1 and Assumption 3.1, which allows to apply Theorem 3.2. Interestingly, for one of these examples, $\beta_t = 1/4$ and so $\sum_{t=1}^{\infty} \beta_t = \infty$, but still $\sum_{t=1}^{\infty} \varrho_t < \infty$: this means that Assumption 2.1 is more general than β-mixing.

4.3.1. *Auto-regressive processes*

Proposition 4.5. *Assume that X_t takes values in $\mathbb{R}^d$ and that $k(x,y) = F(\|x-y\|)$ where F is an L-Lipschitz function. Assume that*

$$X_{t+1} = AX_t + \varepsilon_{t+1}$$

where the (ε_t) are i.i.d. with $\mathbb{E}\|\varepsilon_0\| < \infty$, and A is a matrix with $\|A\| = \sup_{\|x=1\|} \|Ax\| < 1$. Then Assumption 2.1 is satisfied with

$$\varrho_t \leq \|A\|^t \frac{2L\mathbb{E}\|\varepsilon_0\|}{1-\|A\|} \quad \text{and} \quad \Sigma = \sum_{t=1}^{\infty} \varrho_t = \frac{2\|A\|L\mathbb{E}\|\varepsilon_0\|}{(1-\|A\|)^2}.$$

Moreover, assume that almost surely, $\|\varepsilon_t\| \leq c$. Then Assumption 3.1 is satisfied with

$$\gamma_i = \frac{2c\sqrt{L}\|A\|^{\frac{i}{2}}}{1-\|A\|} \quad \text{and} \quad \Gamma = \sum_{i=1}^{\infty} \gamma_i = \frac{2c\sqrt{L\|A\|}}{(1-\|A\|)(1-\sqrt{\|A\|})}.$$

4.3.2. *Examples of non-mixing processes with $\sum_t \varrho_t < \infty$*

First, we remind a classical example of non-mixing process, in the sense that $\sum_{t=1}^{\infty} \beta_t = \infty$. See, for example, Section 1.5 page 8 in [31] where it is also proven that it is neither α-mixing. The process is real-valued, it is defined by $X_{t+1} = (X_t + \eta_{t+1})/2$, where the η_t are i.i.d. $\mathcal{B}e(1/2)$ and $X_0 \sim \mathcal{U}([0,1])$. Note that the noise is there $\varepsilon_t = \eta_t/2$. As for any t, $X_t = f(X_{t+1})$ where f is the measurable function $f(x) = 2x - \lfloor 2x \rfloor$, it is possible to take $I = J = 2$, $V_1 = U_1$ and $V_2 = U_2 = U_1^c$ for some U_1 with $\mathbb{P}(U_1) = 1/2$ in Definition 4.1. This leads to $\beta(\sigma(X_0), \sigma(X_t)) \geq 1/4$.

However, according to Proposition 4.5, as $\mathbb{E}|\varepsilon_0| = 1/4$, we have

$$\varrho_t \leq \frac{L}{2^t}, \quad \Sigma = \sum_{t=1}^{\infty} \varrho_t = 2L < \infty, \quad \text{and} \quad \gamma_i = 2^{1-\frac{i}{2}}\sqrt{L}, \quad \Gamma = \sum_{i=1}^{\infty} \gamma_i = \frac{2\sqrt{L}}{\sqrt{2}-1} < \infty.$$

Another classical example of non-mixing process is a reversed version of the previous one. We draw $X_0 \sim \mathcal{U}([0,1])$ and simply define $X_{t+1} = f(X_t)$ where we still have $f(x) = 2x - \lfloor 2x \rfloor$. Note that once X_0 is given, the process $(X_t)_{t \geq 0}$ is entirely deterministic, and thus non-mixing. Properties of (generalized versions) of such processes are studied in Section 3.3 page 28 in [31]. It is not difficult to check that $Y_t = X_{-t}$ actually satisfies $Y_{t+1} = (Y_t + \varepsilon_t)/2$ where the ε_t are independent $\mathcal{B}e(1/2)$. Thus, a straightforward adaptation of the proof of Proposition 4.5 leads to $\varrho_t \leq 2/L^t$.

5. Stochastic gradient algorithm for MMD estimation

In this section, we briefly discuss gradient-based algorithms to compute the estimator $\hat{\theta}_n$ when $\Theta \subset \mathbb{R}^d$. In Section 5.1, we provide an expression of the gradient of the criterion to be minimized. We briefly provide a special case where this gradient can be computed explicitly. However, in general, this is not the case, but we can provide unbiased estimators of this gradient as soon as we are able to sample from P_θ, in this case the model is often referred to as a *generative model*. Thus, it is possible to use a stochastic gradient algorithm when $\{P_\theta, \theta \in \Theta\}$ is a generative model. We describe this algorithm in Section 5.2, and remind its theoretical properties in Section 5.3.

Note that the idea to use a stochastic gradient algorithm to compute $\hat{\theta}_n$ was first used to train a generative neural network by [42]. In [15], the authors propose to use a stochastic natural gradient algorithm instead. By providing adaptation to the geometry of the problem, the natural gradient will lead to better results but increase the computational burden when the dimension of the problem is large.

5.1. Gradient of the MMD distance

We remind that in this whole section, $\Theta \subset \mathbb{R}^d$. To compute $\hat{\theta}_n$, one must minimize, with respect to $\theta \in \Theta$,

$$\mathbb{D}_k^2(P_\theta, \hat{P}_n) = \mathbb{E}_{X,X' \sim P_\theta}\big[k\big(X, X'\big)\big] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{X \sim P_\theta}\big[k(X_i, X)\big] + \frac{1}{n^2}\sum_{1 \le i,j \le n} k(X_i, X_j)$$

or, equivalently,

$$\mathrm{Crit}(\theta) = \mathbb{E}_{X,X' \sim P_\theta}\big[k\big(X, X'\big)\big] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{X \sim P_\theta}\big[k(X_i, X)\big].$$

In order to use gradient algorithms or any first order method, a first step is to compute the gradient of this quantity with respect to θ.

Proposition 5.1. *Assume that each P_θ has a density p_θ with respect to the Lebesgue measure. Assume that for any x, $\theta \mapsto p_\theta(x)$ is differentiable with respect to θ and that there is a non-negative function $g(x, x')$ such that, for any $\theta \in \Theta$, $|k(x, x')\nabla_\theta[p_\theta(x)p_\theta(x')]| \le g(x, x')$ and $\iint g(x, x')\mu(\mathrm{d}x)\mu(\mathrm{d}x') < \infty$. Then*

$$\nabla_\theta \,\mathrm{Crit}(\theta) = 2\mathbb{E}_{X,X' \sim P_\theta}\left[\left(k\big(X, X'\big) - \frac{1}{n}\sum_{i=1}^{n} k(X_i, X)\right)\nabla_\theta\big[\log p_\theta(X)\big]\right].$$

Note that the gradient of $\mathrm{Crit}(\theta)$ is given by an expectation with respect to P_θ. So, as soon as it is feasible to sample from P_θ, on can provide unbiased estimates of $\nabla_\theta \,\mathrm{Crit}(\theta)$, and thus implement a stochastic gradient algorithm.

Remark 5.1. It might be that in special cases, we have explicit formulas for the expectations in $\mathrm{Crit}(\theta)$ and its gradient. For example, assume that we are a translation parameter, that is: $p_\theta(x) = f(x - \theta)$ for some density f, and that the kernel k is given by $k(x, x') = K(x - x')$ for some function K.

Then

$$\mathrm{Crit}(\theta) = \iint K\big(x-x'\big) f(x-\theta) f\big(x'-\theta\big) \mu(\mathrm{d}x) \mu\big(\mathrm{d}x'\big) - \frac{2}{n} \sum_{i=1}^{n} \int K(X_i - x) f(x-\theta) \mu(\mathrm{d}x)$$

$$= \iint K\big(x-x'\big) f(x) f\big(x'\big) \mu(\mathrm{d}x) \mu\big(\mathrm{d}x'\big) - \frac{2}{n} \sum_{i=1}^{n} \int K(\theta + x - X_i) f(x) \mu(\mathrm{d}x).$$

For example, in the case $P_\theta = \mathcal{U}[\theta - 1/2, \theta + 1/2]$ we have

$$\mathrm{Crit}(\theta) = \iint_{[-1/2,1/2]^2} K\big(x-x'\big)\,\mathrm{d}x\,\mathrm{d}x' - \frac{2}{n} \sum_{i=1}^{n} \int_{-1/2}^{1/2} K(\theta + x - X_i)\,\mathrm{d}x$$

$$= \iint_{[-1/2,1/2]^2} K\big(x-x'\big)\,\mathrm{d}x\,\mathrm{d}x' - \frac{2}{n} \sum_{i=1}^{n} \int_{\theta-1/2-X_i}^{\theta+1/2-X_i} K(u)\,\mathrm{d}u$$

and thus

$$\nabla_\theta \mathrm{Crit}(\theta) = -\frac{2}{n} \sum_{i=1}^{n} \big[K(\theta + 1/2 - X_i) - K(\theta - 1/2 - X_i)\big].$$

So, in this special case, the estimation of the gradient is unnecessary and we can use a proper gradient algorithm to compute $\hat{\theta}_n$.

5.2. Projected stochastic gradient algorithm for the MMD estimator

From Proposition 5.1,

$$\nabla_\theta \mathrm{Crit}(\theta) = 2\mathbb{E}_{X,X' \sim P_\theta}\left[\left(k\big(X, X'\big) - \frac{1}{n}\sum_{i=1}^{n} k(X_i, X)\right) \nabla_\theta \big[\log p_\theta(X)\big]\right].$$

So, if we can compute $\nabla[\log p_\theta(x)]$ and if it is feasible to simulate from P_θ, we can easily compute a Monte Carlo estimator of $\nabla_\theta \mathrm{Crit}(\theta)$ and thus use a stochastic gradient descent (SGD). First, simulate $(Y_1, \ldots, Y_M)$ i.i.d. from P_θ, then put

$$\widehat{\nabla_\theta \mathrm{Crit}}(\theta) = \frac{2}{M} \sum_{j=1}^{M} \left(\frac{1}{M-1} \sum_{\ell \neq j} k(Y_j, Y_\ell) - \frac{1}{n} \sum_{i=1}^{n} k(X_i, Y_j)\right) \nabla_\theta \big[\log p_\theta(Y_j)\big].$$

We provide the details of a projected stochastic gradient algorithm (PSGA) in Algorithm 1. The projection step is necessary if $\Theta \subsetneq \mathbb{R}^d$. Thus, we assume that $\Theta \subseteq \mathbb{R}^d$ is a closed and convex subset and let Π_Θ denote the orthogonal projection on Θ.

5.3. Theoretical analysis of the algorithm

In its original version, the stochastic gradient algorithm was proposed with a sequence of steps $(\eta)_t$ such that $\eta_t \to 0$ and $\sum_t \eta_t = \infty$. However, [77] proved that the method can be made more robust by taking a constant step size $\eta_t = \eta$ and by averaging the parameters. The following proposition is actually a direct application of the results of [77].

Algorithm 1 PSGA for MMD

1: **Input**: a dataset $(X_1, \dots, X_n)$, a model $(P_\theta, \theta \in \Theta \subset \mathbb{R}^d)$ a kernel k, a sequence of steps $(\eta_t)_{t\geq 1}$, an integer M, a stopping time T.
2: **Initialize** $\theta^{(0)} \in \Theta$, $t = 0$.
3: **For** $t = 1, \dots, T$
4: draw $(Y_1, \dots, Y_M)$ i.i.d from $P_{\theta^{(t-1)}}$,
5: $\theta^{(t)} = \Pi_\Theta\{\theta^{(t-1)} - \frac{2\eta_t}{M}\sum_{j=1}^{M}[\frac{1}{M-1}\sum_{\ell\neq j} k(Y_j, Y_\ell) - \frac{1}{n}\sum_{i=1}^{n} k(X_i, Y_j)]\nabla_{\theta^{(t-1)}}[\log p_{\theta^{(t-1)}}(Y_j)]\}$
6: **End for**

Proposition 5.2. *Under the conditions of Proposition* (5.1) *above, and under the assumption that* Θ *is closed, convex and bounded with* $\mathcal{D} = \sup_{(\theta,\theta')\in\Theta^2} \|\theta - \theta'\|$, *define*

$$\hat{\theta}_n^{(T)} = \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}$$

where the $\theta^{(t)}$*'s are given by Algorithm* 1. *Assume that, for any* $\theta \in \Theta$,

$$\mathbb{E}\big[\big\|\widehat{\nabla_\theta \mathrm{Crit}}(\theta)\big\|^2\big] \leq M^2.$$

Assume that $\mathrm{Crit}(\theta)$ *is a convex function of* θ. *Then the choice* $\eta = \mathcal{D}/(M\sqrt{T})$ *leads to*

$$\mathbb{E}\big[\mathrm{Crit}\big(\hat{\theta}_n^{(T)}\big) - \mathrm{Crit}(\hat{\theta}_n)\big] \leq \frac{\mathcal{D}M}{\sqrt{T}}, \tag{3}$$

where the expectation $\mathbb{E}$ *is taken with respect to drawings of the* Y_i*'s in Algorithm* 1. *Moreover,*

$$\mathbb{E}\big[\mathbb{D}_k\big(P_{\hat{\theta}_n^{(T)}}, P^0\big)\big] \leq \inf_{\theta\in\Theta}\mathbb{D}_k\big(P_\theta, P^0\big) + 3\sqrt{\frac{1 + 2\sum_{t=1}^{n}\varrho_t}{n}} + \sqrt{\frac{\mathcal{D}M}{\sqrt{T}}}$$

where the expectation is taken with respect to the sample and to the Y_i*'s, and the choice* $T = n^2$ *leads to*

$$\mathbb{E}\big[\mathbb{D}_k\big(P_{\hat{\theta}_n^{(n^2)}}, P^0\big)\big] \leq \inf_{\theta\in\Theta}\mathbb{D}_k\big(P_\theta, P^0\big) + \frac{\sqrt{\mathcal{D}M} + 3\sqrt{1 + 2\sum_{t=1}^{n}\varrho_t}}{n}.$$

The restrictive assumption in this proposition is the convexity assumption on the criterion. However, it is satisfied in some of the examples of Section 4.

Example 5.1. Let us come back to the "estimation with a dictionary" example of Section 4: P_θ is given by its density

$$p_\theta = \sum_{\ell=1}^{D}\theta_\ell\Phi_\ell.$$

Let us assume that $\Theta = \mathcal{S}_D$ and the Φ_ℓ's are probability densities. Then Θ closed, convex and bounded with $\mathcal{D} = 1$. Moreover,

$$\widehat{\nabla_\theta \operatorname{Crit}}(\theta) = \frac{2}{M}\sum_{j=1}^{M}\left[\frac{1}{M-1}\sum_{\ell\neq j} k(Y_j, Y_\ell) - \frac{1}{n}\sum_{i=1}^{n} k(X_i, Y_j)\right]\nabla_\theta\big[\log p_\theta(Y_j)\big].$$

and

$$\nabla_\theta\big[\log p_\theta(Y_j)\big] = \begin{pmatrix} \dfrac{\Phi_1(Y_j)}{\sum_{\ell=1}^{D}\theta_\ell\Phi_\ell(Y_j)} \\ \vdots \\ \dfrac{\Phi_D(Y_j)}{\sum_{\ell=1}^{D}\theta_\ell\Phi_\ell(Y_j)} \end{pmatrix}.$$

Consequently,

$$\begin{aligned}\big\|\widehat{\nabla_\theta \operatorname{Crit}}(\theta)\big\|^2 &= \sum_{\ell=1}^{D}\left(\frac{2}{M}\sum_{j=1}^{M}\left[\frac{1}{M-1}\sum_{\ell\neq j} k(Y_j, Y_\ell) - \frac{1}{n}\sum_{i=1}^{n} k(X_i, Y_j)\right]\frac{\Phi_\ell(Y_j)}{\sum_{\ell=1}^{D}\theta_\ell\Phi_\ell(Y_j)}\right)^2 \\ &\leq \sum_{\ell=1}^{D}\frac{4}{M^2}\sum_{1\leq j,k\leq M}\frac{\Phi_\ell(Y_j)\Phi_\ell(Y_k)}{(\sum_{\ell=1}^{D}\theta_\ell\Phi_\ell(Y_j))(\sum_{\ell=1}^{D}\theta_\ell\Phi_\ell(Y_k))} \\ &= \frac{4}{M^2}\sum_{1\leq j,k\leq M}\frac{\sum_{\ell=1}^{D}\Phi_\ell(Y_j)\Phi_\ell(Y_k)}{p_\theta(Y_j)p_\theta(Y_k)},\end{aligned}$$

and then

$$\begin{aligned}\mathbb{E}\big(\big\|\widehat{\nabla_\theta \operatorname{Crit}}(\theta)\big\|^2\big) &\leq \iint 4\frac{\sum_{\ell=1}^{D}\Phi_\ell(y)\Phi_\ell(y')}{p_\theta(y)p_\theta(y')}p_\theta(y)p_\theta\big(y'\big)\,\mathrm{d}y\,\mathrm{d}y' \\ &= 4\iint\sum_{\ell=1}^{D}\Phi_\ell(y)\Phi_\ell\big(y'\big)\,\mathrm{d}y\,\mathrm{d}y' = 4D.\end{aligned}$$

Hence Proposition 5.2 leads to

$$\mathbb{E}\big[\operatorname{Crit}\big(\hat{\theta}_n^{(T)}\big) - \operatorname{Crit}(\hat{\theta}_n)\big] \leq \sqrt{\frac{4D}{T}} = 2\sqrt{\frac{D}{T}}.$$

Remark 5.2. In many examples, the MMD criterion is not convex, so we cannot apply Proposition 5.2. This includes for example, the estimation of the mean of a Gaussian distribution. However, the simulation study below shows that the stochastic gradient algorithm still provides excellent results, even though we cannot prove that it reached a global minimum of the criterion.

In some sense, the situation is similar to the estimation of mixtures with the EM algorithm: we cannot prove that the algorithm will not get trapped in a local minimum, but the algorithm is still extremely valuable in practice. Note that many strategies were proposed in order to avoid local minima for EM: for example, multiple runs of the algorithm, with randomized initializations. This strategy works well, at least in reasonably small dimension, even though improvements are possible [79].

6. Simulation study

In this section, we test our stochastic gradient algorithm in the Gaussian mean estimation and in the Gaussian mixture estimation settings. In all the experiments, we chose a number of Monte-Carlo samples equal to n and a step-size of $\eta_t = 1/\sqrt{t}$, and we used the Gaussian kernel $k(x, y) = e^{-\|x-y\|_2^2/d}$ where d is the dimension. Each experiment is repeated 10 times.

6.1. Gaussian mean estimation

First, we estimate the mean of a Gaussian distribution $\mathcal{N}(0, I_d)$ where I_d is the identity matrix of dimension d and where more generally, $\mathbf{a}$ is the vector with all components equal to $a \in \mathbb{R}$. We provide numerical experiments to illustrate and validate our theory on the non-minimax optimality given in Section 4. We verify the rates we obtained from a theoretical point of view by exploring via numerical experiments various types of contamination distributions Q and different proportions of outliers. The MMD gradient descent is compared with the maximum likelihood estimator which is here the arithmetic mean, the componentwise median, and the JS-GAN studied in [43], which is known to be robust and even minimax optimal. Note that in [43], JS-GAN outperforms iterative filtering and dimension halving in all experiments, so we don't include these two methods here. The metric considered here is the square root of the mean square error (MSE) over all the 50 repetitions. We focus on the scenario where $d/n < \varepsilon$ with $d = 10$, $n = 500$ and $\varepsilon = 0.2$. We are interested here in the influence of the contamination distribution Q on the MSE. The results are reported on Table 1, where the bold character marks the lowest MSE among all methods for each contamination Q. As expected, we can see that there is mainly one "worst-case contamination" (the point-mass distribution in a $\sqrt{d}$-far contamination parameter) for which the MMD estimator performs poorly. At the opposite, in all other situations, the MMD estimator is one of the best methods and is not really affected by the distance of the contamination parameter to θ_0. In particular, MMD estimation is competitive with the minimax-optimal JS-GAN procedure.

Additionally, Figure 2 shows that the estimation error is linear with respect to the proportion of outliers ε in dimension $d = 10$ with a sample size $n = 5000$. We chose the contamination $Q = \mathcal{N}(5, I_d)$,

Table 1. Square root of the MSE for several choices of Q (with standard deviations in parenthesis from the 50 repeated experiments). Here, $\varepsilon = 0.2$, $d = 10$ and $n = 500$ such that $\sqrt{d/n} < \varepsilon$. Chosen structure of the network for the GAN: 1 hidden layer with 5 hidden units (as suggested in [43]). The bold character marks the lowest MSE among all methods for each Q. The MMD estimator performs poorly only when there are outliers on the sphere of radius $\sqrt{d}$ centered at 0, i.e. $Q = \delta_{\{\theta_c\}}$ with $\|\theta_0 - \theta_c\| = \sqrt{d}$

Method	$\mathcal{N}(0.2, I_d)$	$\mathcal{N}(0.5, I_d)$	$\mathcal{N}(1, I_d)$	$\mathcal{N}(5, I_d)$	$\mathcal{N}(10, I_d)$	$\mathcal{C}(0.5)$	$\delta_{\{1\}}$	$\delta_{\{10\}}$
Mean	**0.0379**	0.0954	0.2033	1.0166	1.9915	0.3577	0.2057	2.0048
	(0.0046)	(0.0039)	(0.0115)	(0.0145)	(0.0153)	(0.6451)	(0.0115)	(0.0156)
Median	0.0387	**0.0893**	0.1756	0.3106	0.3345	**0.0769**	0.3194	0.3258
	(0.0158)	**(0.0098)**	(0.0058)	(0.0109)	(0.0164)	**(0.0232)**	(0.0215)	(0.0098)
JS-GAN	0.1848	0.2036	0.2172	0.1879	0.2204	0.2276	**0.1969**	0.1877
	(0.0443)	(0.0346)	(0.0241)	(0.0287)	(0.0423)	(0.0376)	**(0.0342)**	(0.0324)
MMD	0.0654	0.1172	**0.1730**	**0.0634**	**0.0681**	0.0882	0.3622	**0.0601**
	(0.0132)	(0.0199)	**(0.0077)**	**(0.0081)**	**(0.0157)**	(0.0140)	(0.0212)	**(0.0157)**

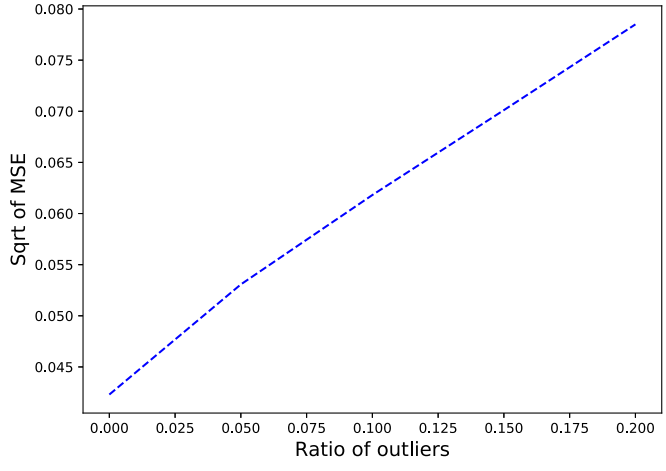

Figure 2. Mean square error as a function of the outliers ratio ε, for a dimension $d = 10$, a sample size $n = 5000$, and a Gaussian contamination $Q = \mathcal{N}(5, I_d)$. The error grows linearly as the ratio increases.

but this choice is not crucial and other choices of the contamination distribution would lead to the same results.

Similarly, Figure 3 shows the effect of the dimension on the estimation error, using $\varepsilon = 0.1$ and $n = 5000$. We can see that for a "harmless contamination" $Q = \mathcal{N}(5, I_d)$ (blue curve), there is no effect at all of the dimension on the estimation error. This result is still valid for other contaminations. At the opposite, when choosing the "worst-case contamination" $Q = \delta_{\{1\}}$ (red curve), such that the distance between the true parameter and the corrupted mean is $\sqrt{d}$, then the estimation error grows linearly in the square root of the dimension, which explains the rate of convergence of the MMD estimator.

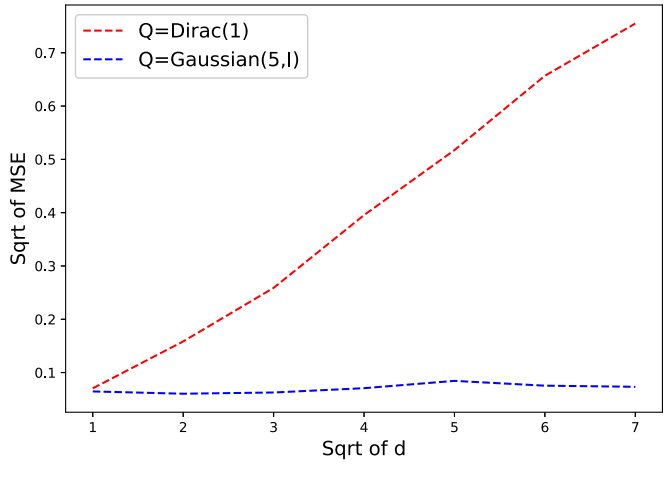

Figure 3. Mean square error as a function of the square root of the dimension $\sqrt{d}$, for an outlier ratio $\varepsilon = 0.1$, a sample size $n = 5000$, and two different contaminations: a "harmless" Gaussian $Q = \mathcal{N}(5, I_d)$ and a "worst-case" Dirac $Q = \delta_{\{1\}}$. The error grows linearly in the Dirac case but is not affected by the dimension in the Gaussian case.

6.2. Gaussian mixture estimation

In the second experiment of this paper, we sample data according to a three component Gaussian mixture $0.3\mathcal{N}(-3.72, 1) + 0.3\mathcal{N}(0.11, 1) + 0.4\mathcal{N}(4.54, 1)$. Here, we use the same approach than in Section 4.1.3. We try to estimate the mixture as a linear combination of mixture in a dictionary composed of all Gaussians of variance 1 and whose means range from -5 to 5 with a stepsize of 0.02. Note that the Gaussian $\mathcal{N}(0.11, 1)$ is not even in the dictionary. The goal is to estimate the weights of each Gaussian in the dictionary using MMD estimation. This estimation method is compared to the gold standard Expectation-Maximization (EM) [32] algorithm and to the tempered Coordinate Ascent Variational Inference (CAVI) algorithm [13,25] that estimate directly the means and the weights of the three-component mixture, using ten random initializations. The experiment is conducted first without any outlier, and then with an outlier equal to 100. Here, the metric is a mean average error (MAE) between the densities. First, we sample 10.000 datapoints independently according to the true mixture. Then, we evaluate the difference between the true density p^0 and the estimated density $p_{\hat{\theta}_n}$ evaluated at each of the 10.000 datapoints, and we finally take the average:

$$\begin{cases} z_1, \dots, z_N \overset{\text{i.i.d.}}{\sim} p^0 & \text{where } N = 10.000, \\ \text{MAE} = \dfrac{1}{N} \displaystyle\sum_{\ell=1}^{N} \left| p^0(z_\ell) - p_{\hat{\theta}_n}(z_\ell) \right|. \end{cases}$$

Again, the final metric is the average over 50 repetitions of the experiment. Figures 4, 5 and 6, and Table 2 clearly show that our estimator performs comparably to both the EM and the CAVI algorithms in the well-specified case, while it is the only one that is not sensitive to the outlier and that gives a consistent estimate.

7. Proofs of the main theorems

7.1. A preliminary lemma: Convergence of $\hat{P}_n$ to P^0 with respect to $\mathbb{D}_k$

Lemma 7.1. *We have*

$$\mathbb{E}\left[\mathbb{D}_k^2\left(\hat{P}_n, P^0\right)\right] \leq \frac{1 + 2\sum_{t=1}^{n}(1 - \frac{t}{n})\varrho_t}{n}.$$

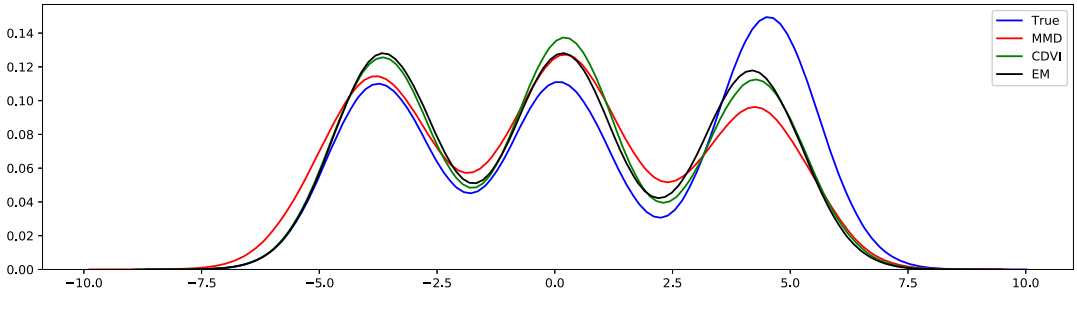

Figure 4. Plot of the estimated densities using different methods without outliers. The blue curve represents the true density, the red one the MMD density, the green one the CAVI density and the black one the EM density.

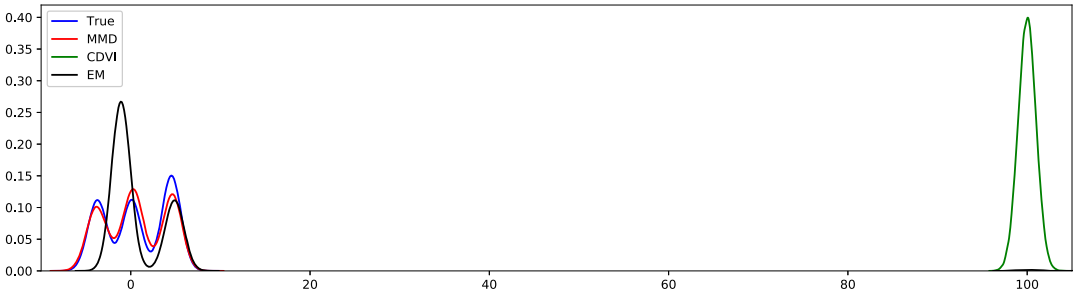

Figure 5. Plot of the estimated densities using different methods in presence of 1 outlier at 100. The blue curve represents the true density, the red one the MMD density, the green one the CAVI density and the black one the EM density. The EM estimate has a small component at 100, and CAVI only one component at 100.

Proof. First, note that $\mathbb{E}(\|k(X_i,\cdot)-\mu_{P^0}\|^2_{\mathcal{H}_k}) \leq \mathbb{E}(\|k(X_i,\cdot)\|^2_{\mathcal{H}_k})$. This formula is the RKHS version of $\mathrm{Var}(X) \leq \mathbb{E}(X^2)$ and is proven in the following way:

$$\begin{aligned}\mathbb{E}\Big(\big\|k(X_i,\cdot)-\mu_{P^0}\big\|^2_{\mathcal{H}_k}\Big) &= \mathbb{E}\Big(\big\|k(X_i,\cdot)\big\|^2_{\mathcal{H}_k}\Big) - 2\mathbb{E}\Big(\big\langle k(X_i,\cdot),\mu_{P^0}\big\rangle^2_{\mathcal{H}_k}\Big) + \mathbb{E}\Big(\|\mu_{P^0}\|^2_{\mathcal{H}_k}\Big)\\ &= \mathbb{E}\Big(\big\|k(X_i,\cdot)\big\|^2_{\mathcal{H}_k}\Big) - 2\big\langle\mathbb{E}\big(k(X_i,\cdot)\big),\mu_{P^0}\big\rangle^2_{\mathcal{H}_k} + \mathbb{E}\Big(\|\mu_{P^0}\|^2_{\mathcal{H}_k}\Big)\\ &= \mathbb{E}\Big(\big\|k(X_i,\cdot)\big\|^2_{\mathcal{H}_k}\Big) - \mathbb{E}\Big(\|\mu_{P^0}\|^2_{\mathcal{H}_k}\Big) \leq \mathbb{E}\Big(\big\|k(X_i,\cdot)\big\|^2_{\mathcal{H}_k}\Big). \qquad (4)\end{aligned}$$

Let us now prove the lemma. We have

$$\begin{aligned}\mathbb{E}\big[\mathbb{D}^2_k\big(\hat{P}_n,P^0\big)\big] &= \mathbb{E}\left\{\left\|\frac{1}{n}\sum_{i=1}^n\big[k(X_i,\cdot)-\mu_{P^0}\big]\right\|^2_{\mathcal{H}_k}\right\}\\ &= \frac{1}{n^2}\mathbb{E}\left\{\sum_{i=1}^n\|k(X_i,\cdot)-\mu_{P^0}\|^2_{\mathcal{H}_k} + 2\sum_{1\leq i<j\leq n}\big\langle k(X_i,\cdot)-\mu_{P^0},k(X_j,\cdot)-\mu_{P^0}\big\rangle_{\mathcal{H}_k}\right\}.\end{aligned}$$

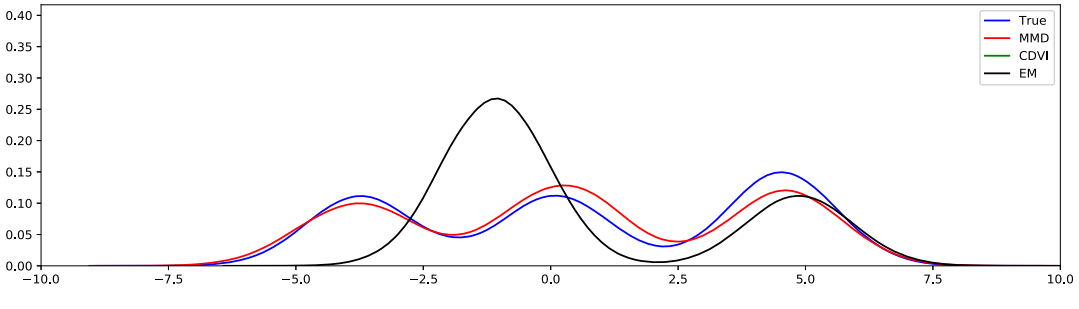

Figure 6. Zoom of Figure 5, without the component of EM at 100.

Table 2. MAE for the Gaussian mixture with/without the outlier (with the corresponding standard deviations)

Algorithm	Without the outlier	With the outlier
MMD	**0.0170 (0.0052)**	**0.0173 (0.0045)**
CAVI	0.0218 (0.0172)	0.0976 (0.0002)
EM	0.0186 (0.0147)	0.0738 (0.0186)

We upper bound the first term in the right-hand side thanks to (4), while we remark that the second term exactly matches the definition of $\varrho_{|i-j|}$. This leads to:

$$\begin{aligned}\mathbb{E}\big[\mathbb{D}_k^2\big(\hat{P}_n, P^0\big)\big] &\leq \frac{1}{n^2}\mathbb{E}\left\{\sum_{i=1}^n \big\|k(X_i,\cdot)\big\|_{\mathcal{H}_k}^2 + 2\sum_{1\leq i<j\leq n}\varrho_{|i-j|}\right\}\\ &= \frac{1}{n^2}\mathbb{E}\left\{\sum_{i=1}^n k(X_i,X_i) + 2\sum_{1\leq i<j\leq n}\varrho_{|i-j|}\right\}.\end{aligned}$$

As we assumed in all the paper that $|k(x,x')|\leq 1$ for any $(x,x')\in\mathbb{X}^2$, we obtain:

$$\mathbb{E}\big[\mathbb{D}_k^2\big(\hat{P}_n, P^0\big)\big] \leq \frac{1}{n^2}\left(n + 2\sum_{1\leq i<j\leq n}\varrho_{|i-j|}\right) = \frac{1+2\sum_{t=1}^n(1-\frac{t}{n})\varrho_t}{n}.$$

□

Note that in the i.i.d case, this leads to

$$\mathbb{E}\big[\mathbb{D}_k^2\big(\hat{P}_n, P^0\big)\big] \leq \frac{1}{n}$$

and thus

$$\mathbb{E}\big[\mathbb{D}_k\big(\hat{P}_n, P^0\big)\big] \leq \sqrt{\mathbb{E}\big[\mathbb{D}_k^2\big(\hat{P}_n, P^0\big)\big]} \leq \frac{1}{\sqrt{n}}.$$

The rate $1/\sqrt{n}$ is known to be minimax in this case: Theorem 1 in [89].

7.2. Proof of Theorem 3.1

Proof. First,

$$\begin{aligned}\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big) &\leq \mathbb{D}_k(P_{\hat{\theta}_n}, \hat{P}_n) + \mathbb{D}_k\big(\hat{P}_n, P^0\big)\\ &\leq \mathbb{D}_k(P_\theta, \hat{P}_n) + \mathbb{D}_k\big(\hat{P}_n, P^0\big)\end{aligned}$$

for any $\theta\in\Theta$, by definition of $\hat{\theta}_n$, and thus, using the triangle inequality again,

$$\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big) \leq \mathbb{D}_k\big(P_\theta, P^0\big) + 2\mathbb{D}_k\big(\hat{P}_n, P^0\big).$$

Take the expectation on both sides and note that

$$\mathbb{E}\big[\mathbb{D}_k\big(\hat{P}_n, P^0\big)\big] \leq \sqrt{\mathbb{E}\big[\mathbb{D}_k^2\big(\hat{P}_n, P^0\big)\big]} \leq \sqrt{\frac{1+2\sum_{t=1}^n(1-\frac{t}{n})\varrho_t}{n}} \leq \sqrt{\frac{1+2\sum_{t=1}^n \varrho_t}{n}}$$

where the second inequality is given by Lemma 7.1. □

7.3. Proof of Theorem 3.2

We start by reminding the following result from [15]; similar results can be found in [86] or [48].

Lemma 7.2 (Lemma 1 page 10 [15]). *For any $\delta > 0$,*

$$\mathbb{P}\left(\mathbb{D}_k\big(\hat{P}_n, P^0\big) \leq \frac{1}{\sqrt{n}}\big(1+\sqrt{\log(1/\delta)}\big)\right) \geq 1-\delta.$$

This result (that we won't use here) relies on McDiarmid inequality [74] who proposed a beautiful way to control the difference between a function of the data, $f(X_1,\dots,X_n)$, and its expectation. The idea relies on writing this function as a martingale, $f(X_1,\dots,X_n) = M_n$ where M_t, for $t \leq n$, is given by $M_t = \mathbb{E}[f(X_1,\dots,X_n)|X_1,\dots,X_t]$, and controlling the martingale increments. It appears that many inequalities can be proven by using this technique, this is discussed in details in Chapter 3 in [14]. Using this technique, Rio [82] proved a version of McDiarmid's inequality for series satisfying Assumption 3.1 (note that the paper is written in French, a more recent paper by the same author [83] in English contains this result and new ones). We start by reminding Rio's result.

Lemma 7.3 (Theorem 1 page 906 [82]). *Under Assumption 3.1, assume that $\mathcal{H}_k^n \to \mathbb{R}$ satisfies:*

$$\big|f(a_1,\dots,a_n) - f\big(a_1',\dots,a_n'\big)\big| \leq \sum_{i=1}^n \big\|a_i - a_i'\big\|_{\mathcal{H}_k}.$$

Then, for any $t > 0$,

$$\mathbb{E}\exp\big[tf(\mu_{\delta_{X_1}},\dots,\mu_{\delta_{X_1}}) - t\mathbb{E}\big[f(\mu_{\delta_{X_1}},\dots,\mu_{\delta_{X_1}})\big]\big] \leq \exp\left(\frac{t^2(1+\Gamma)^2 n}{2}\right).$$

This allows us to state our variant of Lemma 7.2.

Lemma 7.4. *Under Assumptions 2.1 and 3.1,*

$$\mathbb{P}\left(\mathbb{D}_k\big(\hat{P}_n, P^0\big) \leq \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}}\right) \geq 1-\delta.$$

Proof of Lemma 7.4. Define

$$f(a_1,\dots,a_n) = \left\|\sum_{i=1}^n (a_i - \mu_{P^0})\right\|_{\mathcal{H}_k}.$$

For any $x > 0$ and any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\big(\mathbb{D}_k\big(\hat{P}_n, P^0\big) - \mathbb{E}\big[\mathbb{D}_k\big(\hat{P}_n, P^0\big)\big] \geq x\big) \\
&\quad = \mathbb{P}\left(\frac{f(\mu_{\delta_{X_1}}, \ldots, \mu_{\delta_{X_n}})}{n} - \mathbb{E}\big[\mathbb{D}_k\big(\hat{P}_n, P^0\big)\big] \geq x\right) \\
&\quad \leq \mathbb{E}\exp\left(t\frac{f(\mu_{\delta_{X_1}}, \ldots, \mu_{\delta_{X_n}})}{n} - t\mathbb{E}\big[\mathbb{D}_k\big(\hat{P}_n, P^0\big)\big] - tx\right) \quad \text{by Markov inequality} \\
&\quad \leq \exp\left(\frac{t^2(1+\Gamma)^2}{2n} - tx\right) \quad \text{by Lemma 7.3} \\
&\quad = \exp\left(-\frac{x^2 n}{2(1+\Gamma)^2}\right)
\end{aligned}$$

where we chose $t = xn/(1+\Gamma)^2$. Put $x = (1+\Gamma)\sqrt{2\log(1/\delta)/n}$ to get:

$$\mathbb{P}\left(\mathbb{D}_k\big(\hat{P}_n, P^0\big) \leq \mathbb{E}\big[\mathbb{D}_k\big(\hat{P}_n, P^0\big)\big] + (1+\Gamma)\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}\right) \geq 1 - \delta.$$

Use Theorem 7.1 to upper bound the expectation in the right-hand side. This gives the claimed result:

$$\mathbb{P}\left(\mathbb{D}_k\big(\hat{P}_n, P^0\big) \leq \sqrt{\frac{1+2\Sigma}{n}} + (1+\Gamma)\sqrt{\frac{2\log(\frac{1}{\delta})}{n}}\right) \geq 1 - \delta. \qquad \square$$

We are now in position to prove Theorem 3.2.

Proof of Theorem 3.2. With probability $1 - \delta$, for any $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{D}_k\big(P_{\hat{\theta}_n}, P^0\big) &\leq \mathbb{D}_k(P_{\hat{\theta}_n}, \hat{P}_n) + \mathbb{D}_k\big(\hat{P}_n, P^0\big) \quad \text{(triangle inequality)} \\
&\leq \mathbb{D}_k(P_\theta, \hat{P}_n) + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}} \quad \text{(definition of } \hat{\theta}_n \text{ and Lemma 7.4)} \\
&\leq \mathbb{D}_k\big(P_\theta, P^0\big) + \mathbb{D}_k\big(\hat{P}_n, P^0\big) + \frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}} \quad \text{(triangle inequality)} \\
&\leq \mathbb{D}_k\big(P_\theta, P^0\big) + 2\frac{\sqrt{1+2\Sigma} + (1+\Gamma)\sqrt{2\log(\frac{1}{\delta})}}{\sqrt{n}} \quad \text{(Lemma 7.4)}
\end{aligned}$$

$\square$

7.4. Proof of Lemma 3.3 and of Proposition 3.5

Proof of Lemma 3.3. We have

$$\begin{aligned}
\big|\mathbb{D}_k\big(P_\theta, P^0\big) - \mathbb{D}_k(P_\theta, P_{\theta_0})\big| &\leq \mathbb{D}_k\big(P^0, P_{\theta_0}\big) \\
&= \big\|(1-\varepsilon)\mu_{P_{\theta_0}} + \varepsilon\mu_Q - \mu_{P_{\theta_0}}\big\|_{\mathcal{H}_k} \quad \text{(definition of } P^0\text{)}
\end{aligned}$$

$$
\begin{aligned}
&= \left\| \varepsilon(\mu_Q - \mu_{P_{\theta_0}}) \right\|_{\mathcal{H}_k} \\
&\leq \varepsilon\left(\|\mu_Q\|_{\mathcal{H}_k} + \|\mu_{P_{\theta_0}}\|_{\mathcal{H}_k} \right) \quad \text{(triangle inequality)} \\
&\leq 2\varepsilon.
\end{aligned}
$$

□

Proof of Proposition 3.5. Let us put

$$\tilde{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\tilde{X}_i}.$$

First, note that for any probability measure Q,

$$
\begin{aligned}
\left| \mathbb{D}_k(Q, \tilde{P}_n) - \mathbb{D}_k(Q, \hat{P}_n) \right| &\leq \mathbb{D}_k(\hat{P}_n, \tilde{P}_n) \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} \left(k(X_i, \cdot) - k(\tilde{X}_i, \cdot) \right) \right\|_{\mathcal{H}_k} \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left\| k(X_i, \cdot) - k(\tilde{X}_i, \cdot) \right\|_{\mathcal{H}_k} \\
&= \frac{1}{n} \sum_{i \in \mathcal{O}} \left\| k(X_i, \cdot) - k(\tilde{X}_i, \cdot) \right\|_{\mathcal{H}_k} \\
&\leq \frac{2|\mathcal{O}|}{n} \\
&\leq 2\varepsilon. \qquad (5)
\end{aligned}
$$

Consider $Q = P_{\tilde{\theta}_n}$. Then:

$$
\begin{aligned}
\mathbb{D}_k\left(P_{\tilde{\theta}_n}, P^0 \right) &\leq \mathbb{D}_k(P_{\tilde{\theta}_n}, \tilde{P}_n) + \mathbb{D}_k\left(\tilde{P}_n, P^0 \right) \\
&\leq \mathbb{D}_k(P_{\hat{\theta}_n}, \tilde{P}_n) + \mathbb{D}_k\left(\tilde{P}_n, P^0 \right) \quad \text{by definition of } \tilde{\theta}_n \\
&\leq \left[2\varepsilon + \mathbb{D}_k(P_{\hat{\theta}_n}, \hat{P}_n) \right] + \left[2\varepsilon + \mathbb{D}_k\left(\hat{P}_n, P^0 \right) \right]
\end{aligned}
$$

where we used (5) with $Q = P_{\hat{\theta}_n}$ and then $Q = P^0$, respectively. So:

$$
\begin{aligned}
\mathbb{D}_k\left(P_{\tilde{\theta}_n}, P^0 \right) &\leq 4\varepsilon + \mathbb{D}_k(P_{\hat{\theta}_n}, \hat{P}_n) + \mathbb{D}_k\left(\hat{P}_n, P^0 \right) \\
&\leq 4\varepsilon + \mathbb{D}_k(P_{\theta_0}, \hat{P}_n) + \mathbb{D}_k\left(\hat{P}_n, P^0 \right) \quad \text{by definition of } \hat{\theta}_n \\
&= 4\varepsilon + 2\mathbb{D}_k\left(\hat{P}_n, P^0 \right)
\end{aligned}
$$

as it is here assumed that $P^0 = P_{\theta_0}$. □

8. Conclusion

Parametric estimation with MMD provides a simple way to define universally consistent, robust estimators. In many settings, these estimators also have optimal rates of convergence. The computation of

the MMD-based estimator can generally be done through a stochastic gradient descent. We thus believe that it is a practically reasonable and nice alternative to many robust estimation procedures.

Interestingly, Proposition 4.1 provides a natural calibration to the kernel parameter, which is usually a problem in practice. However, in more general settings, the calibration of this parameter, and the choice of the kernel, remain important open questions.

The application of this method to more sophisticated models in statistics and in machine learning (time series models, regression) should be investigated in details and will be the object of future works. In particular, the coefficients ϱ_t of Definition 2.1 are new to our knowledge and it would be interesting to compare them to more weak dependence coefficients.[2]

Acknowledgements

We would like to thank Guillaume Lecué (ENSAE Paris) for his helpful comments, Mathieu Gerber (University of Bristol) who fixed a mistake in the constants in Proposition 4.1, and George Wynne (Imperial College) for his very informative comments on the coefficients ϱ_t. We also would like to thank the anonymous Referees and the Associate Editor for their insightful comments that helped to improve the structure of the paper. All the remaining mistakes are ours.

Supplementary Material

Supplement to "Finite sample properties of parametric MMD estimation: Robustness to misspecification and dependence" (DOI: 10.3150/21-BEJ1338SUPP; .pdf). The supplement [27] of this paper contains the remaining proofs.

References

[1] Alon, N., Matias, Y. and Szegedy, M. (2008). The space complexity of approximating the frequency moments *J. Comput. System Sci.* **58** 137–147.

[2] Alquier, P. (2008). Density estimation with quadratic loss: A confidence intervals method. *ESAIM Probab. Stat.* **12** 438–463. MR2437718 https://doi.org/10.1051/ps:2007050

[3] Amenta, N., Bern, M., Eppstein, D. and Teng, S.-H. (2000). Regression depth and center points. *Discrete Comput. Geom.* **23** 305–323. MR1744506 https://doi.org/10.1007/PL00009502

[4] Arlot, S., Celisse, A. and Harchaoui, Z. (2019). A kernel multiple change-point algorithm via model selection. *J. Mach. Learn. Res.* **20** Paper No. 162, 56. MR4048973

[5] Baraud, Y. and Birgé, L. (2018). Rho-estimators revisited: General theory and applications. *Ann. Statist.* **46** 3767–3804. MR3852668 https://doi.org/10.1214/17-AOS1675

[6] Baraud, Y., Birgé, L. and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.* **207** 425–517. MR3595933 https://doi.org/10.1007/s00222-016-0673-5

[7] Barron, A., Schervish, M.J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718 https://doi.org/10.1214/aos/1018031206

[8] Bernton, E., Jacob, P.E., Gerber, M. and Robert, C.P. (2017). Inference in generative models using the Wasserstein distance. Preprint. Available at arXiv:1701.05146.

[2] In a personal communication, George Wynne (Imperial College London) proved a connection to yet another notion of dependence: L^p-m-approximability, used for functional time series [55]. He proved that if $(X_t)_{t\in\mathbb{Z}}$ is L^p-m-approximable then $\sum_{t=1}^{\infty} \varrho_t < +\infty$.

[9] Biau, G., Cadre, B., Sangnier, M. and Tanielian, U. (2020). Some theoretical properties of GANs. *Ann. Statist.* **48** 1539–1566. MR4124334 https://doi.org/10.1214/19-AOS1858
[10] Bickel, P.J. (1976). Another look at robustness: A review of reviews and some new developments. *Scand. J. Stat.* **3** 145–168. MR0428589
[11] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129 https://doi.org/10.1007/BF00532480
[12] Birgé, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Annales de l'IHP Probabilités et statistiques*. **42** 273–325.
[13] Blei, D.M., Kucukelbir, A. and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773
[14] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001
[15] Briol, F.-X., Barp, A., Duncan, A.B. and Girolami, M. (2019). Statistical inference for generative models via maximum mean discrepancy. Preprint. Available at arXiv:1906.05944.
[16] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Sparse density estimation with ℓ_1 penalties. In *Learning Theory. Lecture Notes in Computer Science* **4539** 530–543. Berlin: Springer. MR2397610 https://doi.org/10.1007/978-3-540-72927-3_38
[17] Bunea, F., Tsybakov, A.B., Wegkamp, M.H. and Barbu, A. (2010). Spades and mixture models. *Ann. Statist.* **38** 2525–2558. MR2676897 https://doi.org/10.1214/09-AOS790
[18] Cai, T., Ma, Z. and Wu, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281 https://doi.org/10.1007/s00440-014-0562-z
[19] Huggins, J.H., Campbell, T., Kasprzak, M. and Broderick, T. (2018). Practical bounds on the error of Bayesian posterior approximations: A nonasymptotic approach. Preprint. Available at arXiv:1809.09505.
[20] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454
[21] Catoni, O. and Giulini, I. (2017). Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector. In *NIPS*-2017 *Workshop* (*Almost*) 50 *Shades of Bayesian Learning*: *PAC-Bayesian Trends and Insights*.
[22] Chan, T.M. (2004). An optimal randomized algorithm for maximum Tukey depth. In *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 430–436. New York: ACM. MR2291081
[23] Chen, M., Gao, C. and Ren, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607
[24] Cherapanamjeri, Y., Flammarion, N. and Bartlett, P.L. (2019). Fast mean estimation with sub-Gaussian rates. *Proceedings of the Thirty-Second Conference on Learning Theory*, *PMLR* **99** 786–806.
[25] Chérief-Abdellatif, B.-E. and Alquier, P. (2018). Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electron. J. Stat.* **12** 2995–3035. MR3855643 https://doi.org/10.1214/18-EJS1475
[26] Chérief-Abdellatif, B.-E. and Alquier, P. (2019). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Proceedings of the* 2*nd Symposium on Advances in Approximate Bayesian Inference* (*AABI*), 2020. *Proceedings of Machine Learning Research* **118** 1–21.
[27] Chérief-Abdellatif, B.-E. and Alquier, P. (2022). Supplement to "Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence." https://doi.org/10.3150/21-BEJ1338SUPP
[28] Chinot, G., Lecué, G. and Lerasle, M. (2020). Robust statistical learning with Lipschitz and convex loss functions. *Probab. Theory Related Fields* **176** 897–940. MR4087486 https://doi.org/10.1007/s00440-019-00931-3
[29] Collier, O. and Dalalyan, A.S. (2019). Multidimensional linear functional estimation in sparse Gaussian models and robust estimation of the mean. *Electron. J. Stat.* **13** 2830–2864. MR3998929 https://doi.org/10.1214/19-EJS1590
[30] Dalalyan, A.S. and Sebbar, M. (2018). Optimal Kullback-Leibler aggregation in mixture density estimation by maximum likelihood. *Math. Stat. Learn.* **1** 1–35. MR4049449 https://doi.org/10.4171/msl/1-1-1

[31] Dedecker, J., Doukhan, P., Lang, G., León, J.R., Louhichi, S. and Prieur, C. (2007). *Weak Dependence*: *With Examples and Applications*. *Lecture Notes in Statistics* **190**. New York: Springer. MR2338725
[32] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537
[33] Depersin, J. and Lecué, G. (2019). Robust sub-Gaussian estimation of a mean vector in nearly linear time. Preprint. Available at arXiv:1906.03058.
[34] Devroye, L., Lerasle, M., Lugosi, G. and Oliveira, R.I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440
[35] Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. *Springer Series in Statistics*. New York: Springer. MR1843146 https://doi.org/10.1007/978-1-4613-0125-7
[36] Diakonikolas, I., Kamath, G., Kane, D.M., Li, J., Moitra, A. and Stewart, A. (2016). Robust estimators in high dimensions without the computational intractability. In 57*th Annual IEEE Symposium on Foundations of Computer Science – FOCS* 2016 655–664. Los Alamitos, CA: IEEE Computer Soc. MR3631028 https://doi.org/10.1109/FOCS.2016.85
[37] Diakonikolas, I., Kane, D.M. and Stewart, A. (2018). List-decodable robust mean estimation and learning mixtures of spherical Gaussians. In *STOC*'18 – *Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 1047–1060. New York: ACM. MR3826316 https://doi.org/10.1145/3188745.3188758
[38] Diakonikolas, I., Kong, W. and Stewart, A. (2019). Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* 2745–2754. Philadelphia, PA: SIAM. MR3909639 https://doi.org/10.1137/1.9781611975482.170
[39] Doukhan, P. (1994). *Mixing*. *Lecture Notes in Statistics*: *Properties and Examples* **85**. New York: Springer. MR1312160 https://doi.org/10.1007/978-1-4612-2642-0
[40] Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.* **84** 313–342. MR1719345 https://doi.org/10.1016/S0304-4149(99)00055-1
[41] Du, S.S., Balakrishnan, S. and Singh, A. (2017). Computationally efficient robust estimation of sparse functionals.. *Proceedings of the* 2017 *Conference on Learning Theory*, *PMLR*. **65** 169–212.
[42] Dziugaite, G.K., Roy, D.M. and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Proc. Uncertainty in Artificial Intelligence* (*UAI*), 2015.
[43] Gao, C., Liu, J., Yao, Y. and Zhu, W. (2019). Robust estimation and generative adversarial nets. In *ICLR* 2019.
[44] Giulini, I. (2017). Robust PCA and pairs of projections in a Hilbert space. *Electron. J. Stat.* **11** 3903–3926. MR3714302 https://doi.org/10.1214/17-EJS1343
[45] Giulini, I. (2018). Robust dimension-free Gram operator estimates. *Bernoulli* **24** 3864–3923. MR3788191 https://doi.org/10.3150/17-BEJ981
[46] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680.
[47] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716
[48] Gretton, A., Fukumizu, K., Harchaoui, Z. and Sriperumbudur, B.K. (2009). A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems* 673–681.
[49] Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12** 1069–1103. MR3724979 https://doi.org/10.1214/17-BA1085
[50] Haddouche, M., Guedj, B., Rivasplata, O. and Shawe-Taylor, J. (2020). PAC-Bayes unleashed: Generalisation bounds with unbounded losses. Preprint. Available at arXiv:2006.07279.
[51] Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123 https://doi.org/10.2307/1912775
[52] Holland, M.J. (2019). PAC-Bayes under potentially heavy tails. Preprint. Available at arXiv:1905.07900.
[53] Holland, M.J. (2019). Distribution-robust mean estimation via smoothed random perturbations. Preprint. Available at arXiv:1906.10300.
[54] Hopkins, S.B. (2020). Mean estimation with sub-Gaussian rates in polynomial time. *Ann. Statist.* **48** 1193–1213. MR4102693 https://doi.org/10.1214/19-AOS1843

[55] Hörmann, S. and Kokoszka, P. (2010). Weakly dependent functional data. *Ann. Statist.* **38** 1845–1884. MR2662361 https://doi.org/10.1214/09-AOS768
[56] Hsu, D. and Sabato, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40. MR3491112
[57] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732
[58] Jerrum, M.R., Valiant, L.G. and Vazirani, V.V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** 169–188. MR0855970 https://doi.org/10.1016/0304-3975(86)90174-X
[59] Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K. and Gretton, A. (2017). A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems* 262–271.
[60] Kothari, P.K., Steinhardt, J. and Steurer, D. (2018). Robust moment estimation and improved clustering via sum of squares. In *STOC'*18 – *Proceedings of the* 50*th Annual ACM SIGACT Symposium on Theory of Computing* 1035–1046. New York: ACM. MR3826315
[61] Lai, K.A., Rao, A.B. and Vempala, S. (2016). Agnostic estimation of mean and covariance. In 57*th Annual IEEE Symposium on Foundations of Computer Science – FOCS* 2016 665–674. Los Alamitos, CA: IEEE Computer Soc. MR3631029
[62] Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Stat.* **41** 802–828. MR0267676 https://doi.org/10.1214/aoms/1177696960
[63] Le Cam, L. (1975). On local and global properties in the theory of asymptotic normality of experiments. In *Stochastic Processes and Related Topics* (*Proc. Summer Res. Inst. Statist. Inference for Stochastic Processes*, *Indiana Univ.*, *Bloomington*, *Ind.*, 1974, *Vol.* 1; *Dedicated to Jerzy Neyman*) 13–54. MR0395005
[64] LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381
[65] Lerasle, M., Szabó, Z., Mathieu, T. and Lecué, G. (2020). MONK – outlier-robust mean embedding estimation by median-of-means. In *International Conference on Machine Learning*.
[66] Lecué, G., Lerasle, M. and Mathieu, T. (2020). Robust classification via MOM minimization. *Mach. Learn.* **109** 1635–1665. MR4137195 https://doi.org/10.1007/s10994-019-05863-6
[67] Lee, J. and Raginsky, M. (2018). Minimax statistical learning with Wasserstein distances. In *Advances in Neural Information Processing Systems*.
[68] Li, Y., Swersky, K. and Zemel, R. (2015). Generative moment matching networks. In *International Conference on Machine Learning* 1718–1727.
[69] Liu, J., Huang, Y., Singh, R., Vert, J.-P. and Noble, W.S. (2019). Jointly embedding multiple single-cell omics measurements. In *BioRxiv* Cold Spring Harbor Laboratory.
[70] Liu, Q., Lee, J. and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning* 276–284.
[71] Louhichi, S. (1998). Théorèmes limites pour des suites positivement ou faiblement dépendantes. Thèse de doctorat de l'Université Paris-XI.
[72] Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.* **19** 1145–1190. MR4017683 https://doi.org/10.1007/s10208-019-09427-x
[73] Lugosi, G. and Mendelson, S. (2020). Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc.* (*JEMS*) **22** 925–965. MR4055993 https://doi.org/10.4171/jems/937
[74] McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, 1989 (*Norwich*, 1989). *London Mathematical Society Lecture Note Series* **141** 148–188. Cambridge: Cambridge Univ. Press. MR1036755
[75] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468 https://doi.org/10.3150/14-BEJ645
[76] Muandet, K., Fukumizu, K., Sriperumbudur, B. and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Found. Trends Mach. Learn.* **10** 1–141.
[77] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. MR2486041 https://doi.org/10.1137/070704277
[78] Nemirovsky, A.S. and Yudin, D.B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *Wiley-Interscience Series in Discrete Mathematics*. New York: Wiley. MR0702836

[79] O'Hagan, A., Murphy, T.B. and Gormley, I.C. (2012). Computational aspects of fitting mixture models via the expectation-maximization algorithm. *Comput. Statist. Data Anal.* **56** 3843–3864. MR2957835 https://doi.org/10.1016/j.csda.2012.05.011

[80] Park, M., Jitkrittum, W. and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* **51** 398–407.

[81] Parr, W.C. and Schucany, W.R. (1980). Minimum distance and robust estimation. *J. Amer. Statist. Assoc.* **75** 616–624. MR0590691

[82] Rio, E. (2000). Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *C. R. Acad. Sci. Paris Sér. I Math.* **330** 905–908. MR1771956 https://doi.org/10.1016/S0764-4442(00)00290-1

[83] Rio, E. (2013). On McDiarmid's concentration inequality. *Electron. Commun. Probab.* **18** no. 44, 11. MR3070910 https://doi.org/10.1214/ECP.v18-2659

[84] Rio, E. (2017). *Asymptotic Theory of Weakly Dependent Random Processes. Probability Theory and Stochastic Modelling* **80**. Berlin: Springer. MR3642873 https://doi.org/10.1007/978-3-662-54323-8

[85] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42** 43–47. MR0074711 https://doi.org/10.1073/pnas.42.1.43

[86] Song, L. (2008). Learning via Hilbert Space Embedding of Distributions Ph.D. thesis, University of Sydney.

[87] Song, L., Gretton, A., Bickson, D., Low, Y. and Guestrin, C. (2011). Kernel belief propagation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 707–715.

[88] Sriperumbudur, B.K., Gretton, A., Fukumizu, K., Schölkopf, B. and Lanckriet, G.R.G. (2010). Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.* **11** 1517–1561. MR2645460

[89] Tolstikhin, I., Sriperumbudur, B.K. and Muandet, K. (2017). Minimax estimation of kernel mean embeddings. *J. Mach. Learn. Res.* **18** Paper No. 86, 47. MR3714249

[90] Tukey, J.W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (*Vancouver, B. C.*, 1974), *Vol.* 2 523–531. MR0426989

[91] van der Vaart, A.W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

[92] Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Stat.* **28** 75–88. MR0088126 https://doi.org/10.1214/aoms/1177707038

[93] Wu, K., Ding, G.W., Huang, R. and Yu, Y. (2020). On minimax optimality of GANs for robust mean estimation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* 4541–4551.

[94] Yatracos, Y.G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.* **13** 768–774. MR0790571 https://doi.org/10.1214/aos/1176349553

[95] Zhao, S., Song, J. and Ermon, S. (2017). InfoVAE: Information Maximizing Variational Autoencoders. Preprint. Available at arXiv:1706.02262.

Received July 2020 and revised March 2021