High-Dimensional Multi-Task Averaging and Application to Kernel Mean Embedding

Hannah Marienwald* **Jean-Baptiste Fermanian**† **Gilles Blanchard**‡

*Universität Potsdam, Technische Universität Berlin, Germany.

†École Normale Supérieure de Rennes, France.

‡Université Paris-Saclay, CNRS, Inria, Laboratoire de Mathématiques d'Orsay, France.

Abstract

We propose an improved estimator for the multi-task averaging problem, whose goal is the joint estimation of the means of multiple distributions using separate, independent data sets. The naive approach is to take the empirical mean of each data set individually, whereas the proposed method exploits similarities between tasks, without any related information being known in advance. First, for each data set, similar or neighboring means are determined from the data by multiple testing. Then each naive estimator is shrunk towards the local average of its neighbors. We prove theoretically that this approach provides a reduction in mean squared error. This improvement can be significant when the dimension of the input space is large; demonstrating a "blessing of dimensionality" phenomenon. An application of this approach is the estimation of multiple kernel mean embeddings, which plays an important role in many modern applications. The theoretical results are verified on artificial and real world data.

1 INTRODUCTION

The estimation of means from i.i.d. data is arguably one of the oldest and most classical problems in statistics. In this work we consider the problem of estimating *multiple* means $\mu_1, \dots, \mu_B$ of probability distributions $\mathbb{P}_1, \dots, \mathbb{P}_B$, over a common space $\mathcal{X} = \mathbb{R}^d$ (or possibly a real Hilbert space $\mathcal{H}$). We assume that for each individual distribution $\mathbb{P}_i$, we observe an i.i.d. data set $X_\bullet^{(i)}$ of size N_i, and that these data sets have been collected independently from each other.

In the rest of the paper, we will call each such data set $X_\bullet^{(i)}$ a *bag*. Mathematically, our model is thus

$$\begin{cases} X_\bullet^{(i)} := (X_k^{(i)})_{1 \le k \le N_i} \overset{i.i.d.}{\sim} \mathbb{P}_i, \ 1 \le i \le B; \\ (X_\bullet^{(1)}, \dots, X_\bullet^{(B)}) \text{ independent,} \end{cases} \tag{1}$$

where $\mathbb{P}_1, \dots, \mathbb{P}_B$ are square integrable distributions on $\mathbb{R}^d$ which we call *tasks*, and our goal is the estimation of their means

$$\mu_i := \mathbb{E}_{X \sim \mathbb{P}_i}[X] \in \mathbb{R}^d, \ 1 \le i \le B. \tag{2}$$

Given an estimate $\widehat{\mu}_i$ of μ_i, we will be interested in its squared error $\|\widehat{\mu}_i - \mu_i\|^2$, and aim at controlling it either with high probability or on average (mean squared error, MSE):

$$\mathrm{MSE}(i, \widehat{\mu}_i) := \mathbb{E}\big[\|\widehat{\mu}_i - \mu_i\|^2\big];$$

this error can be considered either individually for each task $\mathbb{P}_i$ or averaged over all tasks.

This problem is also known as multi-task averaging (MTA) (Feldman et al., 2014), an instance of the multi-task learning (MTL) problem[1]. Prior work on MTL showed that learning multiple tasks jointly yields better performance compared to individual single task solutions (Caruana, 1997; Evgeniou et al., 2005; Feldman et al., 2014). We adapt the idea of joint estimation to the multi-task averaging problem and show that we can take advantage of some unknown *structure* in the set of tasks to improve estimation. The natural baseline for comparison is the naive estimator (NE) given by the simple empirical mean:

$$\widehat{\mu}_i^{\mathrm{NE}} := \frac{1}{N_i} \sum_{k=1}^{N_i} X_k^{(i)}; \quad \mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}) = \frac{1}{N_i} \operatorname{Tr} \Sigma_i, \tag{3}$$

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

[1] or multiple instance learning problem

where Σ_i is the covariance matrix of $\mathbb{P}_i$.

Our motivation for considering this setting is the growing number of large databases taking the above form, where independent bags, corresponding to different but conceptually similar distributions, are available; for example, one can think of i as an index for a large number of individuals, for each of which a number of observations (assumed to be sampled from an individual-specific distribution) have been collected, say medical records, or online activity by some governmental or corporate spying device.

While estimating means in such databases is of interest of its own, a particularly important motivation to consider this setting is that of Kernel Mean Embedding (KME), a technique enjoying sustained attention in the statistical and machine learning community since its introduction in the seminal paper of Smola et al. (2007); see Muandet et al. (2017) for an overview. The KME methodology is used in a large number of applications, e.g. two sample testing (Gretton et al., 2012), goodness-of-fit (Chwialkowski et al., 2016), multiple instance or distributional learning for both supervised (Muandet et al., 2012; Szabó et al., 2016) as well as unsupervised learning (Jegelka et al., 2009), to name just a few.

The core principle of KME is to represent the distribution $\mathbb{P}_Z$ of a random variable Z via the mean of $X = \phi(Z)$, where ϕ is a rich enough feature mapping from the input space $\mathcal{Z}$ to a (reproducing kernel) Hilbert space $\mathcal{H}$. In practice, it is assumed that we have an i.i.d. bag $(Z_k)_{1\leq k\leq N}$ from $\mathbb{P}$, which is used to estimate its KME. Here we are interested again in the situation where a large number of independent bags from different distributions are available, and we want to estimate their KMEs jointly. This is, therefore, an instance of the model (1), once we set $\mathcal{X} := \mathcal{H}$ and $X_k^{(i)} := \phi(Z_k^{(i)})$.

1.1 Relation to Previous Work

The fact that the naive estimator (3) can be improved upon when multiple, real-valued means are to be estimated simultaneously, has a long history in mathematical statistics. More precisely, let us introduce the following isotropic Gaussian setting:

$$\mathbb{P}_i = \mathcal{N}(\mu_i, I_d);\ N_i = N, \qquad 1 \leq i \leq B, \qquad \text{(GI)}$$

on which we will come back in the sequel.

As shown in Stein (1956), for $B = 1$ with $d \geq 3$ the naive estimator is inadmissible, i.e. there exists a strictly better estimator, with a lower MSE for any true mean vector μ_1. An explicit example of a better estimator is given by the celebrated *James-Stein* estimator (JSE) (James and Stein, 1961), which shrinks adaptively the naive estimator towards $\mathbf{0}$, or more generally, towards an a priori fixed vector ν_0.

The MTA problem was introduced by Feldman et al. (2014), who proposed an approach which regularizes the estimation such that similar tasks shall have similar means as well. In practice, however, they either assume the pairwise task similarities to be known or set them to a constant value across tasks[2], which is unrealistic in many applications. In addition to our own approach, we will also introduce a variation of theirs, suitable for the KME framework, that *estimates* the task similarity from the data. Martínez-Rego and Pontil (2013) proposed a method based on spectral clustering of the tasks and apply Feldman et al. (2014)'s method separately on each cluster, but without theoretical analysis.

Variations of the JSE can be shown to yield possible improvements over the NE in more general situations as well (see Fathi et al., 2020 for recent results in non-Gaussian settings). This has also been exploited for KME in Muandet et al. (2016), where a Stein-type estimator in kernel space was shown to generally improve over naive KME estimation. To the best of our knowledge, no shrinkage estimator for KME explicitly designed for or taking advantage of the MTA setting exists.

In the remainder of this work we proceed as follows. Section 2 introduces the basic idea of the approach and starts with a general discussion. We will expose in Section 3 a theoretical analysis proving that the presented method improves upon the naive estimation in terms of squared error, possibly by a large factor. The general theoretical results will be discussed explicitly for the Gaussian setting (Sec. 3.3), the bounded setting (Sec. 3.4) and in the KME framework (Sec. 3.5). The approach is then tested on artificial and real world data in Section 4. The supplemental material (referred to with the suffix S-) contains detailed proofs of the results, details on the different methods compared in the experiments, and additional experimental results.

2 METHOD

The basic idea of our approach is to improve the estimation of a mean of a task by basing its estimation not on its own bag alone, but concatenating the samples from all bags it is *sufficiently similar* to. Since in most practical applications task similarity is not known, we will propose a statistical test that assesses task relatedness based on the given data.

[2] Feldman et al. (2014) mention only in a footnote of their Section 5.1 the possibility of fully estimating the task similarities from the data, but didn't pursue this further.

2.1 Overview of the Approach

In the remainder of the paper we will use the notation $[\![n]\!] := \{1, \ldots, n\}$. For convenience of exposition, assume the (GI) setting. In this case, the naive estimators all have the same MSE, $\overline{\sigma}^2 := d/N$. Fix a particular task (reindexed $i = 0$) with mean μ_0 that we wish to estimate, and assume for now we are given the *side information* that for some constant $\tau > 0$, it holds $\Delta_{0i}^2 := \|\mu_0 - \mu_i\|^2 \le \tau\overline{\sigma}^2$ for some "neighbor tasks" $i \in [\![V]\!]$ (a subset of the larger set of B tasks within range $\tau\overline{\sigma}^2$ to μ_0, reindexed for convenience). Consider the estimator $\widetilde{\mu}_0$ obtained by a simple average of neighbor naive estimators, $\widetilde{\mu}_0 = \frac{1}{V+1}\sum_{i=0}^V \widehat{\mu}_i^{\mathrm{NE}}$. We can bound via usual bias-variance decomposition, independence of the bags and convexity of the squared norm:

$$\begin{aligned} \mathrm{MSE}(0, \widetilde{\mu}_0) &= \left\| \frac{1}{V+1}\sum_{i=1}^{V}(\mu_0 - \mu_i) \right\|^2 + \frac{\overline{\sigma}^2}{V+1} \\ &\le \overline{\sigma}^2 \frac{(1 + V\tau)}{V+1}. \end{aligned} \tag{4}$$

Thus, the above bound guarantees that $\widetilde{\mu}_0$ improves over $\widehat{\mu}_0^{\mathrm{NE}}$ whenever $\tau < 1$, and leads to a relative improvement of order $\max(\tau, V^{-1})$.

In practice, we *don't* have *any* prior side information on the configuration of the means. A simple idea is, therefore, to estimate the quantities Δ_{0i}^2 from the data by an estimator $\widehat{\Delta}_{0i}^2$ and select only those bags for which $\widehat{\Delta}_{0i}^2 \le \widetilde{\tau}\overline{\sigma}^2$. This is in a nutshell the principle of our proposed method.

The deceptive simplicity of the above idea might be met with some deserved skepticism. One might expect that the typical estimation error of $\widehat{\Delta}_{0i}^2$ would be of the same order as the MSE of the naive estimators. Consequently, we could at best guarantee with high probability a bound of $\Delta_{0i}^2 \lesssim \overline{\sigma}^2$ for the estimated neighbor tasks, i.e. $\tau \approx 1$, which does not lead to any substantial theoretical improvement when using (4). The reason why the above criticism is pessimistic, even in the worst case, is the role of the dimension d. From high-dimensional statistics, it is known that the rate of *testing* for $\Delta_{0i}^2 = 0$, i.e. the minimum ρ^2 such that a statistical test can detect $\Delta_{0i}^2 \ge \rho^2$ with probability close to 1, is faster than the rate of *estimation*, $\rho^2 \simeq \sqrt{d}/N = \overline{\sigma}^2/\sqrt{d}$ (see e.g. Baraud, 2002; Blanchard et al., 2018). Thus, we can reliably determine neighbor tasks with $\tau \approx 1/\sqrt{d}$. Based on (4), we can hope again for an improvement of order up to $\mathcal{O}(1/\sqrt{d})$ over NE, which is significant even for a moderately large dimension. In the rest of the paper, we develop the idea sketched here more precisely and illustrate its consequences on KME by numerical experiments. The message we want to convey is that the *curse* of higher dimensional data with its effect on MSE can be to a limit mitigated by a *relative blessing* because we can take advantage of neighboring tasks more efficiently.

2.2 Proposed Approach

Denote $\overline{\sigma}_i^2 = \mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}), i \in [\![B]\!]$. Introduce the following notation: $\Delta_{ij} := \|\mu_i - \mu_j\|$. In general, our approach assumes that we have at hand a family of tests $(T_{ij})_{1 \le i,j \le B}$ for the null hypotheses $H_{ij}^0 : \Delta_{ij}^2 > \tau\overline{\sigma}_i^2$ against the alternatives $H_{ij}^1 : \Delta_{ij}^2 \le \tau'\overline{\sigma}_i^2$, for $0 \le \tau' < \tau$. The exact form of the tests will be discussed later for specific settings.

We denote the set of detected neighbors of task $i \in [\![B]\!]$ as $V_i := \{j : T_{ij} = 1, j \in [\![B]\!]\}$; we can safely assume $T_{ii} = 1$ so that that $i \in V_i$ always holds and $|V_i| \ge 1$. We will also denote $V_i^* = V_i \backslash \{i\}$. For $\gamma \in [0, 1]$, define the modified estimator

$$\widetilde{\mu}_i := \gamma\widehat{\mu}_i^{\mathrm{NE}} + \frac{(1-\gamma)}{|V_i|}\sum_{j \in V_i} \widehat{\mu}_j^{\mathrm{NE}}, \tag{5}$$

which can be interpreted as a local shrinkage estimator pulling the naive estimator towards the simple average of its neighbors.

3 THEORETICAL RESULTS

Introduce the notation

$$\overline{\sigma}^2 := \max_{i \in [\![B]\!]} \mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}) = \max_{i \in [\![B]\!]} (\mathrm{Tr}(\Sigma_i)/N_i). \tag{6}$$

Define further

$$\begin{aligned} G(\tau) &:= \left\{(i,j) \in [\![B]\!]^2 : \Delta_{ij}^2 \le \tau\overline{\sigma}^2\right\}; \\ \overline{G}(\tau) &:= \left\{(i,j) \in [\![B]\!]^2 : \Delta_{ij}^2 \ge \tau\overline{\sigma}^2\right\}, \end{aligned}$$

and the two following events:

$$\begin{aligned} A(\tau) &:= \left\{ \max_{(i,j) \in \overline{G}(\tau)} T_{ij} = 1 \right\}; \\ B(\tau') &:= \left\{ \min_{(i,j) \in G(\tau')} T_{ij} = 0 \right\}; \end{aligned}$$

so $\mathbb{P}[A(\tau)]$ is the collective false positive rate of the tests (or family-wise error rate) while $\mathbb{P}[B(\tau')]$ is the collective false negative rate to detect $\Delta_{ij}^2 \le \tau'\overline{\sigma}^2$ (family-wise type II error rate).

3.1 A General Result under Independence of Estimators and Tests

We start with a result assuming that the tests $(T_{ij})_{(i,j) \in [\![B]\!]^2}$ and the estimators $(\widehat{\mu}_i^{\mathrm{NE}})_{i \in [\![B]\!]}$ are independent. This can be achieved, for instance, by splitting the original samples of each bag into two subsets.

Theorem 3.1. *Assume model* (1) *holds as well as* (2)*, and that* (6) *holds. Furthermore, assume that there exists a family of tests* $(T_{ij})_{(i,j)\in[\![B]\!]^2}$ *that is independent of* $(X_\bullet^{(i)})_{i\in[\![B]\!]}$*. For a fixed constant* $\tau > 0$*, consider the family of estimators* $(\widetilde{\mu}_i)_{i\in[\![B]\!]}$ *defined by* (5) *with respective parameters*

$$\gamma_i := \frac{\tau|V_i^*|}{(1+\tau)|V_i^*|+1}. \qquad (7)$$

Then, conditionally to the event $A^c(\tau)$*, it holds*

$$\forall i \in [\![B]\!] : \mathrm{MSE}(i, \widetilde{\mu}_i) \le \left(\frac{\tau|V_i^*|+1}{(1+\tau)|V_i^*|+1}\right)\overline{\sigma}^2. \qquad (8)$$

Let $\mathcal{N}$ *denote the covering number of the set of means* $\{\mu_j, j \in [\![B]\!]\}$ *by balls of radius* $\sqrt{\tau'}\overline{\sigma}/2$*. Then, conditionally to the events* $A^c(\tau)$ *and* $B^c(\tau')$ *(for* $\tau' < \tau$*), it holds*

$$\frac{1}{B}\sum_{i=1}^{B}\mathrm{MSE}(i, \widetilde{\mu}_i) \le \left(\frac{\tau}{\tau+1} + \frac{\mathcal{N}}{B}\frac{1}{(\tau+1)}\right)\overline{\sigma}^2. \qquad (9)$$

The proof can be found in the Supplemental S-1. In a nutshell, conditional to the favorable event $A^c(\tau)$, and because the tests are independent of the estimators, we can use the argument leading to (4), extended to take into account the shrinkage factor γ, and optimize the value of γ to obtain (7), (8). If $B^c(\tau')$ is satisfied as well, we can deduce (9) directly from (8).

Discussion. The above bounds are in terms of $\overline{\sigma}^2$, the maximum of the naive MSEs over bags, defined in (6). This implies a relevant comparison to the (individual or averaged) naive MSEs only if those are of the same order across tasks, i.e. $\min_{i\in[\![B]\!]}\mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}) \ge \lambda \max_{i\in[\![B]\!]}(\mathrm{Tr}(\Sigma_i)/N_i)$ for some λ close to 1. Significant departures from such an homogeneous situation will require to take into account more carefully individual task-wise MSEs.

The factor in the individual MSE bound (8) is strictly less than 1 as soon as $|V_i| > 1$. As the number of neighbors $|V_i|$ grows, the factor is larger than but approaches $\tau/(1+\tau)$. Therefore, there is a general tradeoff between τ and the number of neighbors in a neighborhood of radius $\sqrt{\tau}\overline{\sigma}$. Nevertheless, in order to aim at possibly significant improvement over naive estimation, a small value of τ should be taken.

The factor in the averaged MSE bound (9) is also always smaller than 1 (as expected from the individual MSE bound). It has a nice interpretation in terms of the ratio $\mathcal{N}/B$: if $\mathcal{N} \ll B$, the improvement factor will be very close to $\tau/(1+\tau)$. Thus, we collectively can improve over the naive estimation wrt MSE as soon as the set of means has a small covering number (at scale $\sqrt{\tau'}\overline{\sigma}/2$) in comparison to its cardinality. This condition can be met in different structural low complexity situations, e.g. clustered means, means being sparse vectors, or lying on a low-dimensional manifold. The method does not need information about said structure in advance and in this sense adapts to it.

3.2 Using the Same Data for Tests and Estimation

We now present a general result in the case where the estimators and tests are not assumed to be independent (e.g. computed from the same data.) To this end we introduce the following additional events:

$$C(\tau): \quad \left\{\max_{i\neq j}|\langle\widehat{\mu}_i^{\mathrm{NE}} - \mu_i, \widehat{\mu}_j^{\mathrm{NE}} - \mu_j\rangle| > \tau\overline{\sigma}^2\right\};$$

$$C'(\tau): \quad \left\{\max_i\|\widehat{\mu}_i^{\mathrm{NE}} - \mu_i\|^2 > \overline{\sigma}^2 + \tau\overline{\sigma}^2\right\}.$$

Theorem 3.2. *Assume that there exists a family of tests* $(T_{ij})_{(i,j)\in[\![B]\!]^2}$*. For a given* $\tau > 0$ *consider the family of estimators* $(\widetilde{\mu}_i)_{i\in[\![B]\!]}$ *defined by* (5) *with respective parameters*

$$\gamma_i := \frac{\tau}{1+\tau}. \qquad (10)$$

Then, for $\tau' \le \tau$*, with probability greater than* $1 - \mathbb{P}[A(\tau) \cup B(\tau') \cup C(\tau) \cup C'(\tau)]$*, it holds*

$$\forall i \in [\![B]\!] : \|\widetilde{\mu}_i - \mu_i\|^2 \le 2\overline{\sigma}^2\left(\tau + \frac{\tau + |V_i|^{-1}}{1+\tau}\right). \qquad (11)$$

Let $\mathcal{N}$ *denote the covering number of the set of means* $\{\mu_j, j \in [\![B]\!]\}$ *by balls of radius* $\sqrt{\tau'}\overline{\sigma}/2$*. Then, with the same probability as above, it holds*

$$\frac{1}{B}\sum_{i=1}^{B}\|\widetilde{\mu}_i - \mu_i\|^2 \le 2\overline{\sigma}^2\left(\tau + \frac{\tau}{1+\tau} + \frac{\mathcal{N}}{B}\frac{1}{1+\tau}\right). \qquad (12)$$

The interpretation of the above result is similar to that of Theorem 3.1, with the caveat that the factor in the MSE bound is not always bounded by 1 as before; but the qualitative behaviour when τ is small, which is the relevant regime, is the same as previously described.

3.3 The Gaussian Setting

In view of the previous results, the crucial point is whether there exists a family of tests such that the events $A(\tau), B(\tau'), C(\tau), C'(\tau)$ have small probability, for a value of τ significantly smaller than 1, and τ' of the same order as τ (up to an absolute numerical constant). This is what we establish now in the Gaussian setting.

Proposition 3.3. *Assume* (GI) *is satisfied. For a fixed $\alpha \in (0,1)$, define the tests*

$$T_{ij} = \mathbf{1}\Big\{ \big\| \widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}} \big\|^2 \leq (2+\tau/2)d/N \Big\}. \qquad (13)$$

Then, provided $\tau \geq 225 \max(\delta, \sqrt{\delta})$, where $\delta := (2\log B + \log \alpha^{-1})/d$, it holds $\mathbb{P}[A(\tau)] \leq \alpha$, $\mathbb{P}[B(\tau')] \leq \alpha$ with $\tau' = \tau/4$, $\mathbb{P}[C(\tau)] \leq 2\alpha$ and $\mathbb{P}[C'(\tau)] \leq \alpha$.

The above result is significant in combination with Theorems 3.1 and 3.2 when δ is small, which is the case if $\log(B)/d$ is small. The message is the following: in a high-dimensional setting, provided $B \ll e^d$, we can reach a large improvement compared to the naive estimators, if the set of means exhibits structure, as witnessed by a small covering number at scale $d^{\frac{1}{4}}\sqrt{(\log B)/N}$. The best-case scenario is when all the means are tightly clustered around a few values, so that $\mathcal{N}$ is small but B is large, then the improvement in the MSE is by a factor of order $\sqrt{(\log B)/d}$.

3.4 The Bounded Setting

The strict Gaussian setting with isotropic covariance is unrealistic for most practical applications. First, real data distributions often depart from being Gaussian; secondly, their covariance matrix is seldom a multiple of the identity, especially in high dimension. To analyze theoretically a more realistic framework, we consider in this section the case of bounded data with general covariance matrices. A particularly important motivation is the application of our results to the Kernel Mean Embedding setting using a bounded kernel, as will be discussed in more detail in the next section. We therefore consider the following bounded setting:

$$\forall i \in [\![B]\!] : N_i = N \text{ and } \big\|X_k^{(i)}\big\| \leq L, \mathbb{P}_i - \text{a.s.}, k \in [\![N]\!]. \qquad \text{(BS)}$$

(note in particular that we still assume that all bags have the same size for the theoretical results.)

In order to apply our general results of Theorems 3.1 and 3.2, we must again find suitable values of τ (as small as possible) and τ' (as close to τ as possible) so that the probability of the events $A(\tau), B(\tau'), C(\tau), C'(\tau)$ is small, in the setting (BS). In that context, the role of the dimension d will be played by the *effective dimension* $\operatorname{Tr}\Sigma/\|\Sigma\|_{op}$. Since this quantity can change from one source distribution to the the other, we will consider the minimum effective dimension across tasks:

$$d_{\text{eff}} := \min_{i\in[\![B]\!]} (\operatorname{Tr}\Sigma_i/\|\Sigma_i\|_{\text{op}}). \qquad (14)$$

We stress that the results of this section do not require that the space has finite dimension, i.e. $\mathcal{X}$ can be a Hilbert space. In this case the covariance operator of a Hilbert-valued random variable is $\Sigma = \mathbb{E}[X \otimes X] - \mathbb{E}[X] \otimes \mathbb{E}[X]$. Under (BS), the covariance operator Σ_i for task i is guaranteed to exist with $\operatorname{Tr}(\Sigma_i) \leq L$.

We will test based on the following estimate of Δ_{ij}^2:

$$U_{ij} = \frac{1}{N(N-1)} \sum_{\substack{k,\ell=1\\ k\neq\ell}}^{N} \Big(\big\langle X_k^{(i)}, X_\ell^{(i)} \big\rangle + \big\langle X_k^{(j)}, X_\ell^{(j)} \big\rangle \Big) - \frac{2}{N^2} \sum_{k,\ell=1}^{N} \big\langle X_k^{(i)}, X_\ell^{(j)} \big\rangle. \qquad (15)$$

Observe that in contrast to the statistic $\big\|\widehat{\mu}_i^{\mathrm{NE}} - \widehat{\mu}_j^{\mathrm{NE}}\big\|^2$ used in the Gaussian setting, which had a bias of $\mathrm{MSE}(i, \widehat{\mu}_i^{\mathrm{NE}}) + \mathrm{MSE}(j, \widehat{\mu}_j^{\mathrm{NE}})$, the above one is unbiased, i.e. $\mathbb{E}[U_{ij}] = \Delta_{ij}^2$. This is why the test threshold (see (18) below) does not include an offset $(2+\tau)$ like in (13). However, the discrimination power of the two types of tests has the same behavior, as established next.

Proposition 3.4. *Consider model* (1)-(2)*, the bounded setting* (BS) *and assume* (14) *holds. Define*

$$r(t) := 5\left(\sqrt{\left(\frac{1}{d_{\text{eff}}} + \frac{L}{N\overline{\sigma}}\right)t} + \frac{Lt}{N\overline{\sigma}} \right), \qquad (16)$$

and

$$\tau_{\min}(t) := r(t)\max\Big(\sqrt{2}, r(t)\Big). \qquad (17)$$

For a fixed $t \geq 1$, define the tests T_{ij} for i, j in $[\![B]\!]^2$

$$T_{ij} := \mathbf{1}\{U_{ij} < \tau\overline{\sigma}^2/2\}. \qquad (18)$$

Then, provided $\tau \geq 144\tau_{min}(t)$, it holds

$$\mathbb{P}[A(\tau) \cup B(\tau/4) \cup C(\tau/7) \cup C'(\tau/48)] \leq 14B^2e^{-t}.$$

The quantity $r(t)$ above (taking $t = \log(14B^2\alpha^{-1})$, where $1-\alpha$ is the target probability) plays a role analogous to δ in the Gaussian setting (Proposition 3.3). As the bag size N becomes sufficiently large, we expect $\overline{\sigma} = \mathcal{O}(N^{-\frac{1}{2}})$ and, therefore, $\overline{\sigma}N = \mathcal{O}(N^{\frac{1}{2}})$. Hence, provided N is large enough, the quantity $r(t)$ is mainly of the order $\sqrt{\log(B)/d_{\text{eff}}}$. Like in the Gaussian case, this factor determines the potential improvement with respect to the naive estimator, which can be very significant if the effective data dimensionality d_{eff} is large.

From a technical point of view, capturing precisely the role of the effective dimension required us to establish concentration inequalities for deviations of sums of bounded vector-valued variables improving over the classical vectorial Bernstein's inequality of Pinelis and Sakhanenko (1986). We believe this result (see Corollary S-6.3 in the Supplemental) to be of interest of its own and to have potential other applications.

3.5 The Kernel Mean Embedding Setting

We recall that the principle of KME posits a reproducing kernel k on an input space $\mathcal{Z}$, corresponding to a feature mapping $\Phi : \mathcal{Z} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space, with $k(z, z') = \langle \phi(z), \phi(z') \rangle$. The feature mapping ϕ can be extended to *probability distributions* $\mathbb{P}$ on $\mathcal{Z}$, via $\phi(\mathbb{P}) := \mathbb{E}_{Z\sim\mathbb{P}}[\phi(Z)]$, provided this expectation exists, which can be guaranteed for instance if ϕ is bounded. This gives rise to an extended kernel on probability distributions via $k(\mathbb{P}, \mathbb{Q}) := \langle \phi(\mathbb{P}), \phi(\mathbb{Q}) \rangle = \mathbb{E}_{(Z,Z')\sim\mathbb{P}\otimes\mathbb{Q}}[k(Z, Z')]$.

As explained in the introduction, if we have a large number of distributions $(\mathbb{P}_i)_{i\in[\![B]\!]}$ for each of which an independent bag $(Z_k^{(i)})_{1\le k\le N_i}$ is available, and we wish to collectively estimate their KMEs, this is an instance of the model (1)-(2) under the transformation $X_k^{(i)} := \phi(Z_k^{(i)})$. The distributions $\mathbb{P}_i$ are replaced by their image distribution through ϕ s.t. $\mu_i = \phi(\mathbb{P}_i)$ and the naive estimators are $\widehat{\mu}_i^{\mathrm{NE}} = \phi(\widehat{\mathbb{P}}_i)$, where $\widehat{\mathbb{P}}_i$ is the empirical measure associated to bag $Z_\bullet^{(i)}$. Under the common assumption that the kernel is bounded, $\sup_{z\in\mathcal{Z}} k(z, z) = \sup_{z\in\mathcal{Z}} \|\phi(z)\|^2 \le L^2$, we are precisely in the case (BS) considered in the previous section. Furthermore, once interpreted in the KME setting the inter-task distance Δ_{ij} is precisely the *maximum mean discrepancy* (MMD) between $\mathbb{P}_i$ and $\mathbb{P}_j$, and the test statistic (15) is exactly the standard MMD^2 estimate (Muandet et al., 2017), which can be evaluated simply by replacing scalar products $\langle X_k^{(i)}, X_\ell^{(j)} \rangle$ by kernel evaluations $k(Z_k^{(i)}, Z_\ell^{(j)})$.

As always for kernel-based methods, elements of the Hilbert space $\mathcal{H}$ are an abstraction which are never explicitly represented in practice; instead, norms and scalar products between elements, that can be written as linear combinations of sample points, can be computed by straightforward formulas using the kernel. In this perspective, a central object is the *inter-task Gram matrix* K defined as $K_{ij} := k(\mathbb{P}_i, \mathbb{P}_j) = \langle \mu_i, \mu_j \rangle, (i, j) \in [\![B]\!]^2$. In the framework of *inference on distributions*, as for instance support measure machines (Muandet et al., 2012), the distributions $\mathbb{P}_i$ act as (latent) training points and the matrix K as the usual kernel Gram matrix for kernel inference. In contrast to what is assumed in standard kernel inference, K is not directly observed but approximated by $\widehat{K}$ s.t. $\widehat{K}_{ij} := \langle \widehat{\mu}_i, \widehat{\mu}_j \rangle$, for some estimators $(\widehat{\mu}_i)_{i\in[\![B]\!]}$ of the true KMEs. The following elementary proposition links the quality of approximation of the means with the corresponding inter-task Gram matrix:

Proposition 3.5. *Assume the model* (1)-(2) *and the bounded setting* (BS). *Let $\widehat{\mu}_i$ be estimators of μ_i bounded by L, and the matrices K and $\widehat{K}$ defined as the Gram matrices of $(\mu_i)_{i\in[\![B]\!]}$ and $(\widehat{\mu}_i)_{i\in[\![B]\!]}$, respectively. Then*

$$\left\| \frac{1}{B}(K - \widehat{K}) \right\|_{\mathrm{Fr.}}^2 \le \frac{4L^2}{B} \sum_{i\in[\![B]\!]} \|\mu_i - \widehat{\mu}_i\|^2, \qquad (19)$$

where $\|K\|_{\mathrm{Fr.}} := \mathrm{Tr}(KK^T)^{\frac{1}{2}}$ *is the Frobenius norm.*

This result further illustrates the interest of improving the task-averaged squared error.

4 EXPERIMENTS AND EVALUATION

We validate our theoretical results in the KME setting[3] on both synthetic as well as real world data.[4] The neighboring kernel means are determined from the tests as described in (18). More specifically, in practice we use the modification that (i) we adapt the formula, i.e. U_{ij} estimated as squared MMD, for possibly unequal bag sizes, and (ii) in each test T_{ij} we replace $\overline{\sigma}^2$ by the task-dependent unbiased estimate

$$\widehat{\mathrm{MSE}}(i, \widehat{\mu}_i^{\mathrm{NE}}) := \frac{1}{2N_i^2(N_i - 1)} \cdot \sum_{k\neq\ell}^{N_i} k(Z_k^{(i)}, Z_k^{(i)}) - 2k(Z_k^{(i)}, Z_\ell^{(i)}) + k(Z_\ell^{(i)}, Z_\ell^{(i)}). \qquad (20)$$

We analyze three different variations of our method which we call similarity test based (STB) approaches. `STB-0` corresponds to (5) with $\gamma = 0$. `STB weight` uses model optimization to find a suitable value for γ, whereas `STB theory` sets γ as defined in (7). However, here we replaced τ with $c \cdot \tau$, where $c > 0$ is a multiplicative constant, to allow for more flexibility.

We compare their performances to the naive estimation, `NE`, and the regularized shrinkage estimator, `R-KMSE`, (Muandet et al., 2016) which also estimates the KME of each bag separately but shrinks it towards zero. Furthermore, we modified the multi-task averaging approach presented in Feldman et al. (2014) such that it can be used for the estimation of kernel mean embeddings. Similar to our idea, this method shrinks the estimation towards related tasks. We test two options: `MTA const` assumes constant similarity for each bag; `MTA stb` uses the proposed test from (18) to assess the bags for their similarity. See Supplemental S-7 for a detailed description of the tested methods.

In the presented results, each considered method has up to two tuning parameters that, in our experiments, are picked in order to optimize averaged test error.

[3] In the Gaussian setting, we report numerical results in the Supplemental S-9.

[4] Code is available at `https://github.com/Han1Mar/stb_kme`.

Therefore, the reported results can be understood as close to "oracle" performance – the best potential of each method when parameters are close to optimal tuning. While this can be considered unrealistic for practice, a closely related situation can occur in the setting where the user wishes to use the method on test bags of size N, and has at hand a limited number of training bags of much larger size $N' \gg N$. From each such training bag, one can subsample N points, use the method for estimation of the means of all bags of size N (incl. subsampled bags), and monitor the error with respect to the means of the full training bags (of size N', used as a ground truth proxy). This allows a reasonable calibration of the tuning parameters.

For all experiments we report the decrease in KME estimation error compared to `NE` in percent. See Supplemental S-8 for the raw errors.

4.1 Synthetic Data

The toy data consists of multiple, two-dimensional Gaussian distributed bags $Z_\bullet^{(i)}$ with fixed means but randomly rotated covariance matrices, i.e.

$$\begin{aligned} Z_\bullet^{(i)} &\sim \mathcal{N}\Big(\mathbf{0}, R(\theta_i)\Sigma R(\theta_i)^T\Big) = \mathbb{P}_i \\ \theta_i &\sim \mathcal{U}(-\pi/4, \pi/4), \end{aligned}$$

where the covariance matrix $\Sigma = \text{diag}(1, 10)$ is rotated using rotation matrix $R(\theta_i)$ according to angle θ_i. The different estimators are evaluated using the unbiased, squared MMD between the estimation $\widetilde{\mu}_i$ and μ_i as loss. Since μ_i is unknown, it must be approximated by another (naive) estimation $\widehat{\mu}_i^{\text{NE}}(Y_\bullet^{(i)})$ based on independent test bags $Y_\bullet^{(i)}$ from the same distribution as $Z_\bullet^{(i)}$, with $|Y_\bullet^{(i)}| = 1000$. The test bag $Y_\bullet^{(i)}$ has much larger size than the training bag $Z_\bullet^{(i)}$, as a consequence the estimator $\widehat{\mu}_i^{\text{NE}}(Y_\bullet^{(i)})$ has a lower MSE than all considered estimators based on $Z_\bullet^{(i)}$, and can be used as a proxy for the true μ_i.[5] In order to guarantee comparability, all methods use a Gaussian RBF with the kernel width fixed to the average feature-wise standard deviation of the data. Optimal values for the model parameter, e.g. ζ and γ for `STB weight`, are selected such that they minimize the estimation error averaged over 100 trials. Once the values for the parameters are fixed, another 200 trials of data are generated to estimate the final generalization error. Different experimental setups were tested:

(a) **Different Bag Sizes**
$B = 50$ and $N_i \in [10, 300]$ for all $i \in [\![B]\!]$,

(b) **Different Number of Bags**
$B \in [10, 300]$ and $N_i = 50$ for all $i \in [\![B]\!]$,

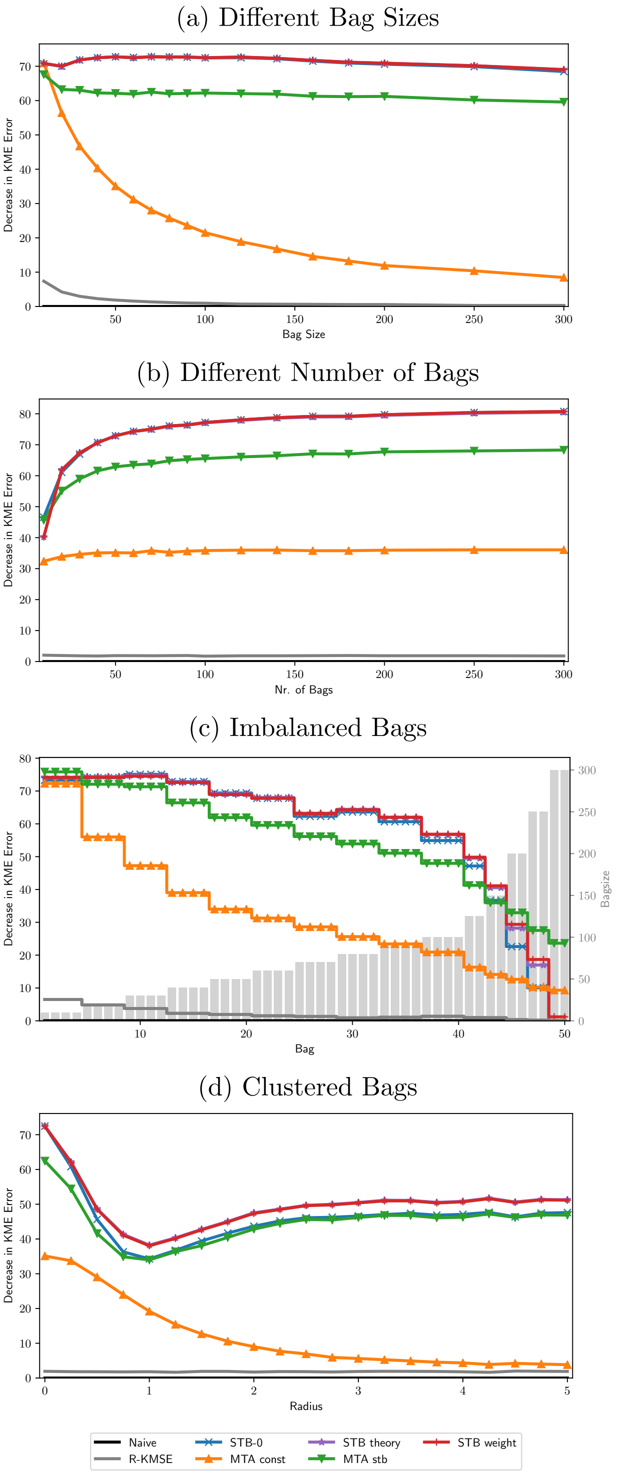
(c) **Imbalanced Bags**
$B = 50$ and $N_1 = 10, \ldots, N_{50} = 300$,

(d) **Clustered Bags**
$N_i, B = 50$ for all $i \in [\![B]\!]$ but the Gaussian distributions are no longer centered around $\mathbf{0}$. Instead, each ten bags form a cluster with the cluster centers equally spaced on a circle. The radius of the circle is varied between 0 and 5, to model different degrees of overlap between clusters.

The results for the experiments on the synthetic data can be found in Figure 1(a) to (d). The estimation of the KME becomes more accurate as the bag size increases. Nevertheless, all of the tested methods provide an increase in estimation performance over the naive estimation, although, the improvement for larger bag sizes decreases for `R-KMSE` and `MTA const`. As expected, methods that use the local neighborhood of the KME yield lower estimation error when the number of available bags increases. Interestingly, this decrease seems to converge towards a capping value, which might reflect the effective dimensionality of the data as indicated by Theorems 3.1 and 3.2 combined with Proposition 3.4. Although we assumed equal bag sizes in the theoretical results, the proposed approaches provide accurate estimations also for the imbalanced setting. Figure 1(c) shows that the improvement is most significant for bags with few samples, which is consistent with results on other multi-task learning problems (see e.g. Feldman et al., 2014). However, when the KME of a bag with many samples is shrunk towards a neighbor with few samples, the estimation can be deteriorated (compare results on (a) with those on (c) for large bag sizes). A similar effect can be seen in the results on the clustered setting. When the bags overlap ($0 <$ radius < 3), a bag from a different cluster might be considered as neighbor which leads to a stronger estimation bias and decline in accuracy. When the tasks have similar centers or are strictly separated (radius > 3), the methods show similar performance to what is shown in Figure 1(b).

To summarize, `NE` and `R-KMSE` give worst performances because they estimate the kernel means separately. Even though `MTA const` assumes all tasks to be related, it improves the estimation performance even when the bags are not similar. However, the methods that derive the task similarity from the local neighborhood achieve most accurate KME estimations in all of the tested scenarios, especially `STB weight` and `STB theory`.

[5] Additionally, the estimation of the squared loss is unbiased if the diagonal entries of the Gram matrix will be included for $Z_\bullet^{(i)}$ but excluded for $Y_\bullet^{(i)}$.


(a) Different Bag Sizes

(b) Different Number of Bags

(c) Imbalanced Bags

(d) Clustered Bags

Figure 1: Decrease in KME estimation error compared to NE in percent on experimental setups (a) to (d). Higher is better. STB methods give similar results so that their plots are overlaid.

4.2 Real World Data

We test our methods on two real world data sets. The AOD-MISR1 data set is a collection of 800 bags with each 100 samples. The samples correspond to randomly selected pixels from a MISR satellite, where each instance is formed by 12 reflectances from three MISR cameras.[6] It can be used to predict the aerosol optical depth (AOD) which poses an important problem in climate research (Wang et al., 2011).

The AOD data is standardized such that each of the features has unit standard deviation and is centered around zero. In each out of the 100 trials, we randomly subsample 20 samples from each bag, on which the KME estimation is based. This estimation is then compared to the naive estimation on the complete bag. Cross-validation, with 400 bags for training and testing, is used to optimize for the model parameters of each approach and then estimate its error. A linear kernel and a Gaussian RBF with the kernel width fixed to one are tested. The results are shown in Table 1.

Table 1: Decrease in KME estimation error compared to NE in percent on the AOD-MISR1 data with different kernels. Higher is better.

METHOD	LINEAR	RBF
R-KMSE	−5.14	8.83
MTA const	8.56	13.92
MTA stb	7.82	17.17
STB-0	0.00	1.43
STB theory	14.51	21.83
STB weight	15.30	22.73

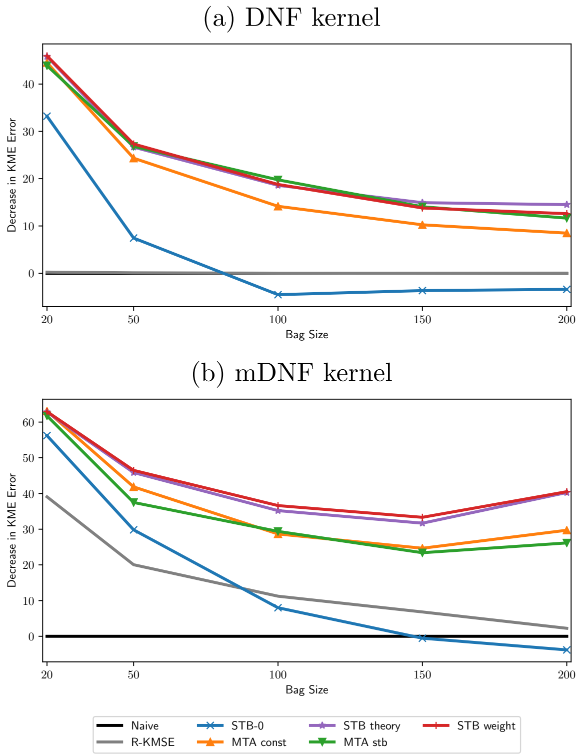
For the linear kernel, STB-0 finds no neighbors and R-KMSE is even worse than NE. STB weight and STB theory provide most accurate estimations of the KMEs.

When the RBF kernel is used, all of the methods provide a more accurate estimation of the KME than the naive approach. The estimations given by STB-0 are similar to those of NE, because STB-0 considers very few bags as neighbors. This lets us conclude that the bags are rather isolated than overlapping. MTA stb, STB weight and STB theory might give better estimations because they allow for more flexible shrinkage. Again, STB weight and STB theory are outperforming the remaining methods.

The second data set[7] consists of wine characteristics, e.g. 'acid', 'juicy', represented as 39-dimensional

[6] We only use 12 out of 16 features because the remaining four are constant per bag.

[7] https://www.kaggle.com/dbahri/wine-ratings.


(a) DNF kernel

(b) mDNF kernel

Figure 2: Decrease in KME estimation error compared to NE in percent on the wine data set for different bag sizes and kernels. Higher is better.

boolean variables of wines (used as samples) of different countries (used as bags) (Gupta et al., 2018). We selected only countries with at least 460 wines, resulting in 15 different bags. In order to get balanced bags, we randomly selected a subset of samples in case a bag had more than 460 samples. The complete bags are used to approximate their true KMEs. This proxy is then again compared with an estimated KME based only on a randomly selected subset of samples (repeated for 100 trials). Leave-one-out cross-validation is used to find optimal model parameters and estimate the estimation performance. Since this is binary data, we applied the Disjunctive Normal Form (DNF) kernel and the monotone DNF kernel (mDNF) (Polato et al., 2018). The DNF counts the number of common $\{0,1\}$-entries, whereas the mDNF kernel only counts common 1s. We normalized the DNF kernel such that it lies in $[0,1]$.

$$k_{\text{DNF}}(z,z') = -1 + 2^{(<z,z'>+<\bar{z},\bar{z}'>)/39}$$
$$k_{\text{mDNF}}(z,z') = -1 + 2^{<z,z'>},$$

where $\bar{z} = \mathbf{1} - z$. The results can be found in Figure 2.

The improvement over `NE` is most significant for small bag-sizes, as it was seen before. Neighboring bags even deteriorate the estimation by `STB-0` such that its performance becomes worse than `NE`. `R-KMSE` provides an improvement only for the mDNF kernel. `STB weight` and `STB theory` again outperform the other methods, including the ones based on `MTA`.

5 CONCLUSION

In this paper we proposed an improved estimator for the multi-task averaging problem. The estimation is improved by shrinking the naive estimation towards the average of its neighboring means. The neighbors of a task are found by multiple testing so that task similarities must not be known a priori. Provided that appropriate tests exist, we proved that the introduced shrinkage approach yields a lower squared error for each task individually and also on average. We show that there exists a family of statistical tests suitable for isotropic Gaussian distributed data or bounded data in a Hilbert space. Theoretical analysis shows that this improvement can be especially significant when the (effective) dimension of the tasks is large, using the property that the typical detection radius of the tests is much smaller than the standard estimation error in high dimension. This property is particularly important for the estimation of multiple kernel mean embeddings (KME), for which the proposed estimator and the theoretical results can naturally be translated.

We tested different variations of the presented approach on synthetic and real world data and compared its performance to other state-of-the-art methods. In all of the conducted experiments, the proposed shrinkage estimators yield the most accurate estimations.

Since the estimation of a KME is often only an intermediate step for solving a final task, as for example in distribution regression (Szabó et al., 2016), further effort must be made to assess whether the improved estimation of the KME also leads to a better final prediction performance. Furthermore, the results on the imbalanced toy data sets have shown that the shrinkage estimator particularly improves the estimation of small bags. However, when the KME of a bag with many samples is shrunk towards a neighbor with low bag size, its estimation might be distorted. Therefore, another direction for future work will be the development of a similarity test or a weighting scheme that take the bag size into account in a principled way. From a theoretical perspective, the empirical estimation of the naive MSEs should also be taken into account. We also will investigate if the improvement factor with respect to the naive estimates is optimal in a suitable minimax sense.

Acknowledgements

The research of HM was funded by the German Ministry for Education and Research as BIFOLD (01IS18025A and 01IS18037A). The research of GB has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901, and by the Agence Nationale de la Recherche (ANR, Chaire IA "BiSCottE"). GB acknowledges the support from the Franco-German University through the binational Doktorandenkolleg CDFA 01-18. GB acknowledges various inspiring and motivating discussions with A. Carpentier, U. Dogan, C. Giraud, V. Koltchinskii, G. Lugosi, A. Maurer, G. Obozinski, C. Scott.

References

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8(5):577–606.

Blanchard, G., Carpentier, A., and Gutzeit, M. (2018). Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electronic Journal of Statistics*, 12(2):3713–3735.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *Proc. of the 33rd International Conference on Machine Learning (ICML 2016)*, volume 48, pages 2606–2615.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(21):615–637.

Fathi, M., Goldstein, L., Reinert, G., and Saumard, A. (2020). Relaxing the Gaussian assumption in shrinkage and SURE in high dimension. arXiv preprint 2004.01378.

Feldman, S., Gupta, M. R., and Frigyik, B. A. (2014). Revisiting Stein's paradox: multi-task averaging. *Journal of Machine Learning Research*, 15(106):3621–3662.

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

Gupta, M. R., Bahri, D., Cotter, A., and Canini, K. (2018). Diminishing returns shape constraints for interpretability and regularization. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, pages 1–11.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379.

Jegelka, S., Gretton, A., Schölkopf, B., Sriperumbudur, B. K., and Von Luxburg, U. (2009). Generalized clustering via kernel embeddings. In *Annual Conference on Artificial Intelligence (KI 2009)*, pages 144–152. Springer.

Martínez-Rego, D. and Pontil, M. (2013). Multi-task averaging via task clustering. In *Proc. Similarity-Based Pattern Recognition - Second International Workshop, SIMBAD 2013*, pages 148–159.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2012). Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems 25 (NeurIPS 2012)*, pages 1–9.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: a review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141.

Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A., and Schölkopf, B. (2016). Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, 17(48):1–41.

Pinelis, I. and Sakhanenko, A. I. (1986). Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148.

Polato, M., Lauriola, I., and Aiolli, F. (2018). A novel Boolean kernels family for categorical data. *Entropy*, 20(6):1–14.

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

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. of the 3rd Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*, volume 1, pages 197–206.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016). Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40.

Wang, Z., Lan, L., and Vucetic, S. (2011). Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2226–2237.