
Computational Efficiency under Covariate Shift in Kernel Ridge Regression

Andrea Della Vecchia
Swiss Finance Institute (SFI)
EPFL
andrea.dellavecchia@epfl.ch

Arnaud Mavakala Watusadisi
MaLGa, DIBRIS
University of Genova
5775617@studenti.unige.it

Ernesto De Vito
MaLGa, DIMA
University of Genova
ernesto.devito@unige.it

Lorenzo Rosasco
MaLGa, DIBRIS
University of Genova
IIT
CBMM, MIT
lorenzo.rosasco@unige.it

Abstract

This paper addresses the covariate shift problem in the context of nonparametric regression within reproducing kernel Hilbert spaces (RKHSs). Covariate shift arises in supervised learning when the input distributions of the training and test data differ, presenting additional challenges for learning. Although kernel methods have optimal statistical properties, their high computational demands in terms of time and, particularly, memory, limit their scalability to large datasets. To address this limitation, the main focus of this paper is to explore the trade-off between computational efficiency and statistical accuracy under covariate shift. We investigate the use of random projections where the hypothesis space consists of a random subspace within a given RKHS. Our results show that, even in the presence of covariate shift, significant computational savings can be achieved without compromising learning performance.

1 Introduction

Classical supervised learning assumes that the training and test data distributions are identical [Vapnik \(1999\)](#). However, in practice, it is often the case that there is a significant mismatch between the two [Quiñonero-Candela et al. \(2022\)](#). This discrepancy can arise from various factors, such as inconsistencies in measurement equipment, differences in data collection domains, or variations in subject populations [Koh et al. \(2021\)](#). Among these scenarios, one particularly interesting and common case is known as covariate shift [Sugiyama & Kawanabe \(2012\)](#). Covariate shift occurs when the marginal distributions of the input covariates differ between the training and test data, while the conditional distribution of the output label given the input covariates remains unchanged [Shimodaira \(2000\)](#); [Cortes et al. \(2008\)](#). This phenomenon is observed in a variety of well-studied learning problems, including among all domain adaptation [Ben-David et al. \(2006\)](#); [Mansour et al. \(2009\)](#); [Cortes & Mohri \(2014\)](#); [Zhang et al. \(2012\)](#), active learning [Wiens \(2000\)](#); [Kanamori & Shimodaira \(2003\)](#); [Sugiyama & Ridgeway \(2006\)](#), natural language processing [Jiang & Zhai \(2007\)](#), and medical image analysis [Guan & Liu \(2021\)](#).

Despite its practical significance, covariate shift remains relatively underexplored in theoretical frameworks compared to the classical setting where no distribution mismatch is assumed. Recently, several studies have attempted to bridge this gap. A widely adopted approach to addressing covariate

shift involves correcting the learning objective by reweighting the loss function using the so-called importance weighting (IW) function. This function corresponds to the Radon-Nikodym derivative of the test marginal distribution with respect to the training marginal distribution [Huang et al. \(2006\)](#); [Sugiyama et al. \(2012\)](#); [Fang et al. \(2020\)](#). [Shimodaira \(2000\)](#) was the first to demonstrate the consistency of the importance-weighted maximum likelihood estimator, while in [Cortes et al. \(2010\)](#) the authors derived suboptimal finite-sample bounds for restricted function classes with finite pseudodimension. In the context of nonparametric regression in reproducing kernel Hilbert spaces (RKHSs) and under the assumption that the regression function belongs to the RKHS (well-specified case), [Gizewski et al. \(2022\)](#) provided optimal excess risk convergence results for the minimizer of the reweighted empirical risk when the weight function is known and uniformly bounded. In [Ma et al. \(2023\)](#), authors recently showed that the standard unweighted kernel ridge regression estimator is still minimax optimal under an appropriate choice of the regularization parameter, provided the IW function is uniformly bounded or its second moment is bounded. Similar results for nonparametric classification were obtained in [Kpotufe & Martinet \(2021\)](#). We also mention [Schmidt-Hieber & Zolotarev \(2024\)](#); [Pathak et al. \(2022\)](#) in the context of nonparametric regression for classes of functions different from RKHSs, and [Wen et al. \(2014\)](#); [Lei et al. \(2021\)](#); [Yamazaki et al. \(2007\)](#) for parametric models. Building on [Ma et al. \(2023\)](#), in [Gogolashvili et al. \(2023\)](#) the authors extended the analysis to (simplified) misspecified case, where the regression function is not assumed to lie in the RKHS itself, but its projection is. In this setting, the authors showed that, under covariate shift, the unweighted classic KRR predictor is not a consistent estimator of the projection of the regression function. In such cases, IW correction is necessary.

Kernel methods provide a robust framework for nonparametric learning, but their scalability is limited by high computational and memory costs—challenges that clearly do not disappear under covariate shift. To address this, researchers have developed more efficient strategies, ranging from improved optimization ([Johnson & Zhang, 2013](#); [Schmidt et al., 2017](#)) to randomized linear algebra techniques ([Mahoney, 2011](#); [Drineas & Mahoney, 2005](#); [Woodruff, 2014](#); [Calandriello et al., 2017](#)). These aim to reduce costs, raising a key question: do such shortcuts compromise statistical accuracy? Recent work suggests they often do not ([Rudi et al., 2015](#); [Bach, 2017](#); [Bottou & Bousquet, 2008](#); [Sun et al., 2018](#); [Rudi & Rosasco, 2017](#); [Della Vecchia et al., 2021](#)). A promising approach is to restrict the hypothesis space to a lower-dimensional (random) subspace, as in sketching ([Kpotufe & Sriperumbudur, 2019](#)) and random projection methods ([Woodruff, 2014](#)), including Nyström methods for kernels ([Smola & Schölkopf, 2000](#); [Williams & Seeger, 2001](#)). Recent results confirm that these methods can achieve both computational efficiency and statistical accuracy, for both smooth and general convex losses ([Rudi et al., 2015](#); [Bach, 2013](#); [Marteau-Ferey et al., 2019](#); [Della Vecchia et al., 2024](#)).

Although random projection techniques—such as the Nyström method—have been widely studied in standard learning settings, their use under covariate shift remains largely unexamined. A recent step in this direction was taken by [Myleiko & Solodky \(2024\)](#), building on the framework of [Gizewski et al. \(2022\)](#). The authors consider Nyström subsampling in a setting where weights are known and uniformly bounded, and derive optimal risk bounds that, in the case of Tikhonov regularization, can be recovered as a special case of our results.

Contribution. In this paper, we investigate the importance-weighting correction for kernel ridge regression under the covariate shift setting. Unlike most of the previous work in this setting, we employ random projection techniques, specifically the Nyström method, to improve the algorithm’s scalability and efficiency. Our primary goal is to explore the balance between computational efficiency and statistical accuracy in this framework. By analyzing the interplay among regularization, subspace size, and the various parameters that characterize the complexity of the problem, we identify the conditions under which the best-known statistical guarantees can be achieved while drastically reducing computational costs. Notably, we show that our algorithm achieves a remarkable result: despite Nyström approximation, it attains the best rates reported in the literature without compromising statistical accuracy, all while being significantly more computationally efficient. Regarding the theoretical aspects, while we build upon established results from the random projections literature, our novel proofs presented here address additional technical challenges arising from the mismatch between training and test distributions, as well as the potential unboundedness of the weighting function in the empirical risk minimization. Full technical details are provided in Appendix C. Finally, we validate these theoretical findings through simulations and real data experiments.

Limitations. As remarked in Section 5, Assumption 3 imposes boundedness of all moments of the weight function. While this is a relatively strong condition, especially when compared to the mere

second-moment boundedness required in Ma et al. (2023), it is essential in our setting to ensure the applicability of the Nyström method with ALS sampling in the unbounded case. Assumption 2 ensures that the projection of the regression function onto the hypothesis space \mathcal{H} exists and lies within \mathcal{H} , rather than on its boundary. This is slightly more general than the standard well-specified setting, but conceptually similar. We do not consider the fully misspecified case here, which introduces significant additional complexity (see interpolation spaces in Steinwart & Christmann (2008) for example). Finally, as mentioned in Section 8, a more refined understanding of the role of the constants in the learning rate bounds could explain why, in practice, a misalignment between training and test distributions can sometimes be benign or, conversely, highly adversarial, depending on its relation with the source condition. These directions are left for future work.

Organization. The paper is organized as follows. In Section 2, we introduce the key definitions and notations used throughout the paper, while refreshing KRR in the classical setting. Section 3 provides an overview of KRR in the covariate shift setting. In Section 4, we present empirical risk minimization on random subspaces using Nyström. Section 5 explores the interplay between computational efficiency and statistical accuracy. Section 6 extends our analysis to scenarios where the true IW function is unknown. Section 7 presents a series of simulations and real-world experiments.

2 Background

We start defining some key quantities we will need in the rest of the paper, see e.g. Caponnetto & De Vito (2007); Smale & Zhou (2007). Given a measurable space \mathcal{X} , a probability distribution μ on \mathcal{X} , a space of square-integrable functions L_μ^2 with respect to measure μ and a Reproducing Kernel Hilbert Space (RKHS) \mathcal{H} of (bounded) kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, with $K_x(\cdot) = K(x, \cdot) \in \mathcal{H}$, define

$$S_\mu : \mathcal{H} \rightarrow L_\mu^2, \quad (S_\mu f)(x) = \langle f, K_x \rangle_{\mathcal{H}} = f(x) \quad \mu - a.s. \quad \text{and} \quad S_\mu^* g = \int K_x g(x) d\mu.$$

for $f \in \mathcal{H}$, $g \in L_\mu^2$. Covariance operator $\Sigma_\mu : \mathcal{H} \rightarrow \mathcal{H}$ is defined as $\Sigma_\mu := S_\mu^* S_\mu = \mathbb{E}_\mu[K_x \otimes K_x]$. Define the sampling operator $\hat{S} : \mathcal{H} \rightarrow \mathbb{R}^n$ associated with set $\{x_1, \dots, x_n\} \in \mathcal{X}^n$, for $f \in \mathcal{H}$, as

$$(\hat{S}f)_i := f(x_i) = \langle f, K_{x_i} \rangle_{\mathcal{H}}, \quad i \in [n], \quad \text{and} \quad \hat{S}^*(\hat{y}) := \frac{1}{n} \sum_{i=1}^n y_i K_{x_i}, \quad \hat{y} \in \mathbb{R}^n.$$

2.1 Classical Setting

Classical nonparametric regression aims to predict a real-valued output $\mathcal{Y} = \mathbb{R}$ given a vector of covariates $X \in \mathcal{X}$. More formally, let $\mathcal{X} \times \mathbb{R}$ be a probability space with distribution ρ , where \mathcal{X} and \mathbb{R} are the input and output spaces, respectively. Let ρ_X denote the marginal distribution of ρ on \mathcal{X} and $\rho(\cdot|x)$ the conditional distribution on \mathbb{R} given $x \in \mathcal{X}$. For any fixed $x \in \mathcal{X}$, the optimal estimator in a mean-squared sense is given by the regression function $g^*(x) := \int y d\rho(y|x)$, i.e.

$$g^* = \arg \min_{g \in \mathbb{R}^{\mathcal{X}}} \mathcal{R}(g) = \mathbb{E}[(Y - g(X))^2].$$

The minimization problem is generally unsolvable since the distribution ρ is unknown. In practice, we only have access to a dataset of n input-output pairs $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathbb{R})$ sampled independently and identically (i.i.d.) from the joint distribution ρ , i.e. $(x_1, y_1), \dots, (x_n, y_n) \sim \rho(x, y)$.

A common approach to derive an approximation of g^* involves restricting the hypothesis space from all measurable functions to a real separable Hilbert space \mathcal{H} , and replacing the expected risk \mathcal{R} with the (regularized) empirical risk $\hat{\mathcal{R}}_\lambda : \mathcal{H} \rightarrow [0, \infty)$ defined as

$$\hat{\mathcal{R}}_\lambda(f) := \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad f \in \mathcal{H}, \quad \lambda > 0. \quad (1)$$

The (Regularized) Empirical Risk Minimization (ERM) algorithm then solves:

$$\hat{f}_\lambda := \arg \min_{f \in \mathcal{H}} \hat{\mathcal{R}}_\lambda. \quad (2)$$

When \mathcal{H} is an RKHS, as in the rest of this paper, the resulting estimator is known as the kernel ridge regression (KRR) estimator. Our primary goal is to upper bound the excess risk \mathcal{E} of \hat{f}_λ , i.e.

$$\mathcal{E}(\hat{f}_\lambda) := \mathcal{R}(\hat{f}_\lambda) - \mathcal{R}(g^*) = \|\hat{f}_\lambda - g^*\|_{\rho_X}^2.$$

Here, the equality follows from a standard result, see for example [Caponnetto & De Vito \(2007\)](#).

3 Covariate Shift Setting

The covariate shift setting introduces an additional complexity: training and test distributions may differ, but only through their marginals, while sharing the same conditional distribution:

$$\rho^{te}(x, y) = \rho(y|x)\rho_X^{te}(x), \quad \rho^{tr}(x, y) = \rho(y|x)\rho_X^{tr}(x).$$

Since the regression function g^* only depends on the conditional distribution, which is identical for both ρ^{te} and ρ^{tr} , it is unique. As in the standard setting, we are provided with n input-output pairs $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathbb{R})$ sampled i.i.d. from ρ^{tr} , i.e. $(x_1, y_1), \dots, (x_n, y_n) \sim \rho^{tr}(x, y)$.

The challenge consists in the fact that we train our model using samples from the ρ^{tr} , but we aim to evaluate its performance on new data drawn from ρ^{te} . Then, we want to upper bound the excess risk

$$\mathcal{E}(\hat{f}_\lambda) = \|\hat{f}_\lambda - g^*\|_{\rho_X^{te}}^2.$$

Note that the empirical risk computed from ρ^{tr} samples is a biased estimate of the expected risk under ρ^{te} . As a result, minimizing it may not yield a predictor that performs well on the test distribution.

3.1 Importance-Weighting (IW) Correction

The goal of importance-weighting (IW) correction is to construct an unbiased estimator of the risk with respect to test distribution ρ^{te} , while using data sampled from ρ^{tr} . The idea is to reweight ERM samples based on their *relevance* to the test distribution, ensuring good performance under ρ^{te} .

If $\rho_X^{te} \ll \rho_X^{tr}$, we can define the IW function w representing the weight assigned to a point $x \in \mathcal{X}$ as the Radon-Nikodym derivative of ρ_X^{te} with respect to ρ_X^{tr} :

$$w(x) := \frac{d\rho_X^{te}}{d\rho_X^{tr}}(x). \quad (3)$$

Points that are likely to be encountered during testing (ρ_X^{te} is large) while are rare to be sampled at training time (ρ_X^{tr} is small) are considered particularly relevant and will receive higher weights.

In the rest of the paper, we will focus on applying this framework in the context of kernel methods.

Assumption 1. \mathcal{H} is an RKHS with scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and associated kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

We define the regularized importance-weighted empirical risk, for all $f \in \mathcal{H}$, as

$$\hat{\mathcal{R}}_\lambda^w(f) := \frac{1}{n} \sum_{i=1}^n w(x_i)(y_i - \langle K_{x_i}, f \rangle)^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad (4)$$

where $\mathcal{H} \ni K_x(\cdot) = K(x, \cdot)$. For the square loss, the minimizer \hat{f}_λ^w of eq. (4) is given by:

$$\hat{f}_\lambda^w(x) = (\hat{S}^* \hat{M}_w \hat{S} + \lambda I)^{-1} \hat{S}^* \hat{M}_w \hat{y}, \quad (5)$$

where \hat{M}_w is the diagonal matrix with i -th entry $w(x_i)$. In case weights are all positive, we have

$$\hat{f}_\lambda^w(x) = \sum_{i=1}^n c_i^w K_{x_i}(x) \in \text{span}\{K_{x_1}, \dots, K_{x_n}\}, \quad c^w = (\hat{K} + n\lambda \hat{M}_{1/w})^{-1} \hat{y} \in \mathbb{R}^n, \quad (6)$$

where \hat{K} is the kernel Gram matrix, and $\hat{M}_{1/w}$ is the diagonal matrix with i -th entry $1/w(x_i)$. Since the regression function g^* may not generally belong to \mathcal{H} (i.e., the model may be misspecified), we introduce the best approximation $f_{\mathcal{H}} \in \mathcal{H}$ of g^* with respect to the $L_{\rho_X^{te}}^2$ distance.

Assumption 2. There exists an $f_{\mathcal{H}} \in \mathcal{H}$ such that

$$\mathcal{E}(f_{\mathcal{H}}) = \min_{f \in \mathcal{H}} \mathcal{E}(f) = \min_{f \in \mathcal{H}} \|f - g^*\|_{\rho_X^{te}}^2. \quad (7)$$

Note that, while the minimizer might not be unique, we select $f_{\mathcal{H}}$ as the unique minimizer with minimal norm ([De Vito et al., 2021](#)). In the following, we will evaluate the performance of our estimator relative to the best estimator in \mathcal{H} , i.e. $f_{\mathcal{H}}$.

Computations A significant limitation of the procedure outlined above is the computational cost associated with the $n \times n$ matrix inversion required to compute the estimator, see (6). This operation has a complexity of $\mathcal{O}(n^3)$ in time and $\mathcal{O}(n^2)$ in memory, making it impractical when $n > 10^5$.

4 ERM on Random Subspaces and the Nyström Method

In this paper, we consider an efficient approximation of the above procedure based on considering a subspace $\mathcal{B} \subset \mathcal{H}$ and solving the corresponding importance-weighted regularized ERM problem

$$\min_{\beta \in \mathcal{B}} \widehat{\mathcal{R}}_\lambda^w(\beta), \quad (8)$$

with $\widehat{\beta}_\lambda^w$ the unique minimizer. As clear from (6), choosing $\mathcal{B} = \mathcal{H}_n = \text{span}\{K_{x_1}, \dots, K_{x_n}\}$ is equivalent to considering the full space \mathcal{H} and yields the same solution as in (5). However, a natural alternative is to consider a smaller subspace:

$$\mathcal{B} = \mathcal{H}_m = \text{span}\{K_{\tilde{x}_1}, \dots, K_{\tilde{x}_m}\}, \quad (9)$$

where $\{\tilde{x}_1, \dots, \tilde{x}_m\} \subset \{x_1, \dots, x_n\}$ is a random subset of the input points and $m \leq n$. This is equivalent to Nyström approximation (Williams & Seeger, 2000). A basic approach is to select these points uniformly at random from the training dataset. Alternatively, we can employ more refined sampling techniques, such as using leverage scores (Drineas et al., 2012)

$$l_i(t) = (\widehat{K}(\widehat{K} + t\mathbf{I})^{-1})_{ii}, \quad i = 1, \dots, n. \quad (10)$$

Although computing leverage scores directly can be computationally expensive, approximate leverage scores (ALS) can be computed efficiently (see Definition 2 in the Appendix), as described in Rudi et al. (2018). We will use ALS sampling for all the results in the following.

Using these ideas, we define the Nyström IW-KRR problem as follows:

$$\widehat{f}_{\lambda,m}^w = \arg \min_{\beta \in \mathcal{H}_m} \frac{1}{n} \left\| \widehat{M}_w^{1/2}(\widehat{y} - \widehat{S}\beta) \right\|_2^2 + \lambda \|\beta\|_{\mathcal{H}}^2 = \arg \min_{f \in \mathcal{H}} \frac{1}{n} \left\| \widehat{M}_w^{1/2}(\widehat{y} - \widehat{S}P_m f) \right\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad (11)$$

where P_m is the orthogonal projection operator onto \mathcal{H}_m , given by $P_m = VV^*$, see Appendix A. Taking the derivative and using the first order condition (see Appendix A), the Nyström estimator can be expressed as:

$$\widehat{f}_{\lambda,m}^w = V(V^* \widehat{S}^* \widehat{M}_w \widehat{S}V + \lambda \mathbf{I})^{-1} V^* \widehat{S}^* \widehat{M}_w \widehat{y}. \quad (12)$$

Alternatively, using some linear algebra, it can also be written as:

$$\widehat{f}_\lambda^w(x) = \sum_{i=1}^m \widetilde{c}_i^w K_{\tilde{x}_i}(x), \quad \widetilde{c}^w = (\widehat{K}_{nm}^T \widehat{M}_w \widehat{K}_{nm} + n\lambda \widehat{K}_{mm})^{-1} \widehat{K}_{nm}^T \widehat{M}_w \widehat{y}, \quad (13)$$

where $\widehat{f}_\lambda^w(x) \in \text{span}\{K_{\tilde{x}_1}, \dots, K_{\tilde{x}_m}\}$, $\widetilde{c}^w \in \mathbb{R}^m$, $\widehat{K}_{nm} \in \mathbb{R}^{n \times m}$, $(\widehat{K}_{nm})_{ij} = K(x_i, \tilde{x}_j)$ and $\widehat{K}_{mm} \in \mathbb{R}^{m \times m}$, $(\widehat{K}_{mm})_{ij} = K(\tilde{x}_i, \tilde{x}_j)$ (see derivation in Appendix A).

Computations From eq. (13), it is clear that the Nyström method can offer significant computational benefits. Unlike in eq. (12), computing our projected estimator only requires $\mathcal{O}(m^3 + m^2n)$ time and $\mathcal{O}(mn)$ memory, compared to the previous $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$. When $m \ll n$ the difference between the two can be big and efficient implementations such as in Rudi et al. (2017); Meanti et al. (2020) can drastically reduce computational requirements.

5 Statistical Guarantees

In this section, we aim to derive excess risk bounds for the Nyström estimator presented in eq. (12). We begin by introducing the technical assumptions required for the subsequent analysis.

5.1 Further Assumptions

The following assumption, inspired by [Gogolashvili et al. \(2023\)](#), ensures the boundedness of the importance-weighting (IW) function or of its moments.

Assumption 3. Let $w = d\rho_X^{te}/d\rho_X^{tr}$ be the IW function. There exist constants $q \in [0, 1]$, $W > 0$ and $\sigma > 0$ such that $\forall p \in \mathbb{N}$, $p \geq 2$

$$\left(\int_{\mathcal{X}} w(x)^{\frac{p-1}{q}} d\rho_X^{te} \right)^q \leq \frac{1}{2} p! W^{p-2} \sigma^2, \quad (14)$$

where the left-hand side for $q = 0$ is defined as $\|w^{p-1}\|_{\infty, \rho_X^{te}}$, the ess sup with respect to ρ_X^{te} .

Considering the uniformly bounded case $\|w\|_{\infty} < \infty$, Assumption 3 is satisfied for $q = 0$. When w is not uniformly bounded, Assumption 3 can still be satisfied for $q \in (0, 1]$ if the moments of w are bounded. For example, it is satisfied for $q \in (0, 1]$ if $W \geq 1$, $\sigma^2 \geq 1$ and

$$2\rho_X^{te} \left(\left\{ x \in X : \frac{d\rho_X^{te}}{d\rho_X^{tr}}(x) \geq t \right\} \right) \leq \sigma^2 \exp(-W^{-1}t^{1/q}) \quad \text{for all } t > 0$$

(see Appendix A in [Gogolashvili et al. \(2023\)](#) for the detailed result). Equivalently, Assumption 3 can be stated as a condition on the Rényi divergence between ρ_X^{te} and ρ_X^{tr} ([Mansour et al., 2009](#); [Cortes et al., 2010](#)). It follows that this assumption can be interpreted as a requirement for the test distribution ρ_X^{te} not to deviate significantly from the train distribution ρ_X^{tr} , with $q \in [0, 1]$ quantifying the extent of the deviation. Introducing the parameter q is useful for presenting results such as Theorem 1 in a unified form, which can then be specialized to the two cases considered in Corollary 1: uniformly bounded ($q = 0$) and possibly unbounded ($q \neq 0$) weights. Note that in [Ma et al. \(2023\)](#), a weaker assumption requiring only the second moment to be bounded is considered. The stronger condition in eq. (14) is anyway crucial in our case to apply the Nyström method in the unbounded setting (see Appendix C and how the proof differentiates from [Rudi et al. \(2015\)](#)).

Assumption 4. The range of the output $Y \in \mathbb{R}$ is upper bounded, i.e. $Y \in [-B, B]$, $B < \infty$.

Furthermore, we make the following regularity assumption, usually known as *source condition*.

Assumption 5 (Source condition). There exist $1/2 \leq r \leq 1$ and $g \in L^2(X, \rho_X^{te})$ with $\|g\|_{\rho_X^{te}} \leq R$ for some $R > 0$ such that $f_{\mathcal{H}} = L^r g$, where $L := S_{\rho_X^{te}}^* S_{\rho_X^{te}}$ is the integral operator.

Assumption 5 and its equivalent formulations (e.g., Assumption 4 in [Rudi et al. \(2015\)](#)) are common in literature ([Smale & Zhou, 2007](#); [Caponnetto & De Vito, 2007](#)). r quantifies the smoothness of the target function $f_{\mathcal{H}}$ and the extent to which it can be well approximated by functions in \mathcal{H} . For $r = 1/2$, the assumption is always satisfied. Intuitively, a larger r implies $f_{\mathcal{H}}$ being smoother.

Finally, we impose an assumption on the capacity of our RKHS, which roughly measures the number of eigenvalues of Σ greater than λ ([Zhang, 2005](#); [Caponnetto & De Vito, 2007](#)).

Definition 1 (Effective dimension). For $\lambda > 0$, define the random variable $\mathcal{N}_x(\lambda) = \langle K_x, (\Sigma + \lambda I)^{-1} K_x \rangle_{\mathcal{H}}$ with $x \in \mathcal{X}$ distributed according to ρ , then $\mathcal{N}_{\rho}(\lambda) = \mathbb{E}_{\rho} \mathcal{N}_x(\lambda)$ is called effective dimension.

Assumption 6 (Capacity condition). $\forall \lambda > 0$, there exists $0 \leq \gamma \leq 1$, $Q > 0$ s.t. $\mathcal{N}_{\rho_X^{te}}(\lambda) < Q\lambda^{-\gamma}$.

It is known that the condition in Assumption 6 is ensured if the eigenvalues $(\eta_i)_i$ of the covariance operator Σ satisfy a polynomial decaying condition $\eta_i \sim i^{-1/\gamma}$ (see Appendix D).

5.2 Excess Risk Bounds

In this section, we present our main theoretical results. We will derive excess risk bounds for the Nyström predictor defined in eq. (12) and we will show that Nyström approximation does not affect the state-of-art rates of convergence while instead reducing both time and memory requirements.

We present now the main result of the paper.

Theorem 1. Under assumptions 1,2,3,4,5,6, for ALS sampling, let $\delta > 0$, $\left(\frac{256(W+\sigma^2)\log^2(4/\delta)}{n}\right)^{\frac{1}{\gamma(1-q)+1+q}} \leq \lambda \leq \|\Sigma\|_{op}$, and $m \geq 144T^2Q\lambda^{-\gamma}\log\frac{8n}{\delta}$, with probability greater or equal than $1 - \delta$

$$\left|\mathcal{R}(\hat{f}_{\lambda,m}^w) - \mathcal{R}(f_{\mathcal{H}})\right|^{1/2} \leq 64B \left(\frac{W}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2}{n\lambda^{\gamma(1-q)+q}}}\right) \log\left(\frac{8}{\delta}\right) + 43R\lambda^r.$$

The detailed proof of Theorem 1 can be found in Appendix C.

If the weighting function w is bounded, i.e. $\|w\|_{\infty} < \infty$, then Assumption 3 is satisfied for $q = 0$. We specify Theorem 1 for this setting against the unbounded one.

Corollary 1. Under the assumptions and conditions as in Theorem 1,

(a) with Assumption 3 satisfied by $q = 0$ and choosing $\lambda \asymp (\|w\|_{\infty}/n)^{\frac{1}{2r+\gamma}}$, with $m \gtrsim (n/\|w\|_{\infty})^{\frac{\gamma}{2r+\gamma}} \log n$, with high probability

$$\mathcal{E}(f_{\lambda,m}^w) = \|\hat{f}_{\lambda,m}^w - f_{\mathcal{H}}\|_{\rho_X^{te}}^2 \lesssim \left(\frac{\|w\|_{\infty}}{n}\right)^{\frac{2r}{2r+\gamma}}, \quad (15)$$

(b) with Assumption 3 satisfied by $q = 1$ and choosing $\lambda \asymp (n/(W + \sigma^2))^{-\frac{1}{2r+1}}$, with $m \gtrsim (n/(W + \sigma^2))^{\frac{\gamma}{2r+1}} \log n$, with high probability

$$\mathcal{E}(f_{\lambda,m}^w) = \|\hat{f}_{\lambda,m}^w - f_{\mathcal{H}}\|_{\rho_X^{te}}^2 \lesssim \left(\frac{1}{n}\right)^{\frac{2r}{2r+1}}. \quad (16)$$

The rate in eq. (15) matches the optimal convergence rate of standard kernel ridge regression (KRR) established in Caponnetto & De Vito (2007). However, it explicitly depends on $\|w\|_{\infty}$, which can become arbitrarily large as the training and test distributions diverge. This result also recovers Theorem 2.1 from Myleiko & Solodky (2024) for Tikhonov regularization in the special case $\gamma = 1$. Compared to their work, we consider ALS sampling, which enables fast rates under a suitable capacity condition. Furthermore, we extend the analysis to the case of unbounded importance weights. Specifically, eq. (16) shows that when the weighting function is unbounded (i.e., $q = 1$ in Assumption 3), the convergence rate deteriorates to $\mathcal{O}(n^{-\frac{2r}{2r+1}})$. This slower rate, which does not depend on the capacity assumption 6), is always worse than the rate in eq. (15). These findings are consistent with the results reported in Gogolashvili et al. (2023) for the full (non-projected) model. Note that the rate in eq. (16) can possibly be improved, even under our Nyström approximation, by clipping the unbounded IW function w at a threshold that depends on n (see Section 5.2.2 in Gogolashvili et al. (2023)). This idea follows Corollary 2 of Ma et al. (2023), although the result is not directly comparable due to differing assumptions.

Example 1. To illustrate the benefits of the Nyström approach, consider for example the common setting $q = 0$, $r = 1/2$ and $\gamma = 1$. From eq. (15), we achieve the optimal rate of $\mathcal{O}(n^{-1/2})$ (Caponnetto & De Vito, 2007), with $m = \mathcal{O}(\sqrt{n} \log(n))$. This results in computational costs of $\mathcal{O}(m^3 + m^2n) = \mathcal{O}(n\sqrt{n} + n^2)$ in time and $\mathcal{O}(mn) = \mathcal{O}(n\sqrt{n})$ in memory. These are significantly lower than the costs of the non-approximated method, respectively $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$.

6 Unknown Weights

Clearly, when dealing with real data, assuming knowledge of the exact weighting function w is unrealistic. Instead, if we have access to some samples from both distributions, we can attempt to estimate an approximate weighting function $v \approx w$ and control the error resulting from this approximation. Let us choose a function v such that there exists a probability distribution ρ_X^v with

$$v(x) = \frac{d\rho_X^v(x)}{d\rho_X^{tr}(x)}, \quad (17)$$

i.e. $v(x)$ is the Radon-Nikodym derivative of ρ_X^v with respect to ρ_X^{tr} . Note that, in general, $\rho_X^v \neq \rho_X^{te}$. Similarly to eq. (7) in Assumption 2, we assume that

$$f_{\mathcal{H}}^v = \arg \min_{f \in \mathcal{H}} \mathcal{E}^v(f) = \|f - g^*\|_{\rho_X^v}^2 \quad (18)$$

exists (again we consider the unique with minimal norm in case of multiple minimizers). Furthermore, we assume that v satisfies Assumptions 3, 5, 6 with constants V, η, r, γ, Q , respectively. Since v is chosen, weights $v(x)$ are known for all $x \in \mathcal{X}$, allowing us to compute the estimator

$$\hat{f}_{\lambda, m}^v = V(V^* \hat{S}^* \hat{M}_v \hat{S} V + \lambda I)^{-1} V^* \hat{S}^* \hat{M}_v \hat{y}, \quad (19)$$

where \hat{M}_v is the diagonal matrix with diagonal entries $v(x_1), \dots, v(x_n)$. Still, we are interested in the excess risk of this estimator with respect to the original ρ^{te} distribution.

Denoting $\Sigma = \Sigma_{\rho_X^{te}}$ and $\Sigma_v = \Sigma_{\rho_X^v}$, we decompose the excess risk of $\hat{f}_{\lambda, m}^v$ as:

$$\mathcal{E}(f_{\lambda, m}^v) = \|\hat{f}_{\lambda, m}^v - f_{\mathcal{H}}\|_{\rho_X^{te}}^2 \leq \|\Sigma \Sigma_{\lambda v}^{-1}\|_{op} \|\hat{f}_{\lambda, m}^v - f_{\mathcal{H}}^v\|_{\rho_X^v}^2 + \|f_{\mathcal{H}}^v - f_{\mathcal{H}}\|_{\rho_X^{te}}^2. \quad (20)$$

Well Specified Case If $g^* \in \mathcal{H}$, then clearly $g^* = f_{\mathcal{H}} = f_{\mathcal{H}}^v$ and eq. (20) simplifies to:

$$\|\hat{f}_{\lambda, m}^v - f_{\mathcal{H}}\|_{\rho_X^{te}}^2 \leq \|\Sigma \Sigma_{\lambda v}^{-1}\|_{op} \|\hat{f}_{\lambda, m}^v - f_{\mathcal{H}}^v\|_{\rho_X^v}^2. \quad (21)$$

The second term corresponds to what we have already analyzed in Theorem 1, with the distribution ρ_X^{te} and its associated weighting function w replaced by ρ_X^v and v . The first term quantifies the additional cost incurred due to the mismatch between ρ_X^{te} and ρ_X^v . Using Proposition 5 from Appendix D, we can show that if we set $v(x) \equiv 1$, i.e. $\rho_X^v \equiv \rho_X^{tr}$, then $\|\Sigma \Sigma_{\lambda v}^{-1}\|_{op} \leq \|w\|_{\infty}$. This indicates that, when w is bounded, we have a finite control on the appearing term involving the two covariance operators.

By setting $\lambda \asymp (\|w\|_{\infty} n)^{\frac{1}{2r+\gamma}}$, and ensuring $m \gtrsim n^{\frac{\gamma}{2r+\gamma}} \log n$, then, with high probability

$$\|\hat{f}_{\lambda, m}^v - f_{\mathcal{H}}\|_{\rho_X^{te}}^2 \lesssim \|w\|_{\infty} \left(\frac{1}{n}\right)^{\frac{2r}{2r+\gamma}}, \quad (22)$$

where we used the fact that $\|v\|_{\infty} = 1$. It is important to note that, when compared with the bound in eq. (15), which assumes knowledge of the true (typically unknown) weights, we achieve the same rate in n . This means that, when the model is well-specified, the classical Nyström-ERM algorithm gives the same rate as the importance-weighted variant, despite the covariate shift between train and test distributions. This result agrees with findings in Ma et al. (2023); Gogolashvili et al. (2023), where anyway random projection approximations are not involved. More than that, we emphasize that although the rate remains the same, the dependence on $\|w\|_{\infty}$, which is assumed finite but can be arbitrarily large, is worse than in eq. (15), where the true w is employed.

Misspecified Case In the misspecified case, the situation is more complex since in general $g^* \neq f_{\mathcal{H}} \neq f_{\mathcal{H}}^v$. In particular, the term $\|f_{\mathcal{H}}^v - f_{\mathcal{H}}\|_{\rho_X^{te}}^2$ in eq. (20) is not zero, and its magnitude can become arbitrarily large, depending on the severity of the mismatch between ρ_X^{te} and ρ_X^v .

7 Simulations and real data experiments

As emphasized in the introduction, the main goal of this study is to show that the Nyström method can deliver significant computational savings under covariate shift without compromising accuracy.

7.1 Simulations

We start by reproducing the experimental setting in Gogolashvili et al. (2023). We want to solve a regression problem using KRR with RBF kernel in the context of distribution shift, assuming $\mu_X^{tr} \sim \mathcal{N}(\mu_{tr}, \Sigma_{tr})$, $\rho_X^{te} \sim \mathcal{N}(\mu_{te}, \Sigma_{te})$ and $\mu_{tr} \neq \mu_{te}$, $\Sigma_{tr} \neq \Sigma_{te}$. The regression function is

$$g^*(x) := c_1 e^{-\frac{c_2}{\|x\|_2^{2k}}}, \quad k \in \mathbb{N}, c \in \mathbb{N},$$

where the parameter k controls the level of misspecification. Note that, in fact, when k increases, the regression function becomes essentially piece-wise constant, and neither constant nor discontinuous functions belong to the RKHS of the Gaussian kernel. Data samples are generated following $y_i^{tr} = g^*(x_i^{tr}) + \xi_i$, $y_i^{te} = g^*(x_i^{te})$, with $x_i^{tr} \sim \rho_X^{tr}$, $x_i^{te} \sim \rho_X^{te}$ and $\xi_i \sim \mathcal{N}(0, \varepsilon^2)$. Figure 1 shows the results in this setting for $k = 50$. In the plot on the left, we observe that the two weighted models, namely KRR with IW correction (W-KRR) and its Nyström-approximated version (W-Nyström KRR), perform similarly. As expected, the simple (unweighted) KRR model shows a performance gap. On the right, despite reaching the same error of the weighted KRR model, Nyström approximation can lead to important computational savings allowing for choosing a number of Nyström centers $m \ll n$.

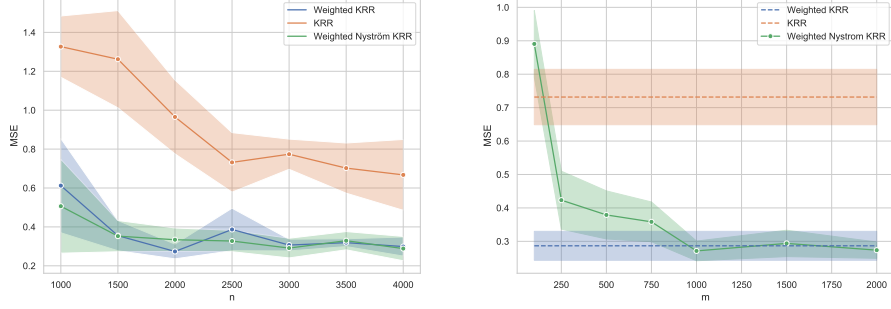


Figure 1: On the left: MSE for the different models varying the number of train samples n . The result for W-Nyström KRR model is obtained for optimal m . On the right: $n = 3000$, optimal m is selected as the smallest for which W-Nyström KRR matches the full W-KRR model ($m = 1000$ here).

7.2 Experiments on Benchmark Datasets

As regards real-world applications, we conduct experiments on commonly used benchmark datasets in the domain adaptation field (Wang & Sun, 2024; Wilson et al., 2020; He et al., 2023; Dinu et al., 2023). The size of the original datasets is reduced in case of memory issues with KRR and W-KRR when building the full Gram matrix \hat{K} (see memory bottlenecks in Section 4). In these experiments, weights are estimated using RuLSIF method (Yamada et al., 2013; Liu et al., 2013). We consider 4 real-world datasets: HHAR (Stisen et al., 2015), WISDM (Kwapisz et al., 2011), HAR70+ (Ustad et al., 2023) and HARChildren (Tørring et al., 2024). To simulate covariate shift, we use data from one user for training and evaluate the models on data from a different user. We consider RBF kernel, with the length-scale γ and regularization λ , tuned through cross-validation. For further details on the datasets and methodologies used, refer to Appendix E.

Table 1: Performances of the various methods, both in terms of MSE and training/prediction time.

	HAR70+ ($n = 20000$)			HARChildren ($n = 15000$)			HHAR ($n = 15500$)			WISDM ($n = 25000$)		
	MSE	t train (s)	t pred (s)	MSE	t train (s)	t pred (s)	MSE	t train (s)	t pred (s)	MSE	t train (s)	t pred (s)
KRR	10 ± 1	1694 ± 2	15.0 ± 0.5	26.6 ± 0.9	762 ± 12	10.2 ± 0.4	3.7 ± 0.3	876 ± 6	10.5 ± 0.9	7.8 ± 0.1	3280 ± 48	38 ± 5
W-KRR	5.0 ± 0.2	1785 ± 2	15.1 ± 0.3	13.5 ± 0.8	809 ± 26	9.0 ± 0.1	1.8 ± 0.3	1034 ± 93	9.9 ± 0.1	4.8 ± 0.2	3364 ± 30	33 ± 2
W-Ny KRR	5.1 ± 0.2	89 ± 19	1.0 ± 0.1	13.2 ± 0.6	8.0 ± 0.2	1.6 ± 0.1	1.8 ± 0.1	6.5 ± 0.4	1.1 ± 0.1	4.7 ± 0.3	9.9 ± 0.4	1.4 ± 0.1

Table 1 shows that the two methods using IW correction achieve the best and essentially equal performance. However, in terms of computational efficiency, our W-Nyström KRR method, offers significant time and memory savings. The number of Nyström points m required by W-Nyström KRR is 1100, 1800, 1400, and 1550, for HAR70+, HARChildren, HHAR, and WISDM, respectively.

8 Conclusions and Future Work

In this paper, we showed that even under covariate shift, random projection techniques —particularly the Nyström method— can significantly enhance computational efficiency without any loss in learning performance. We provide new statistical bounds for our compressed Nyström algorithm, showing that it matches the optimal statistical guarantees of the full W-KRR model. Leveraging results from random projection theory, we developed novel technical proofs to account for the mismatch between training and test distributions and the potential unboundedness of the IW function. We evaluated the effectiveness of our approach through simulations and experiments on real-world datasets.

However, several questions remain open for future investigation. Although optimal rates are achieved in the well-specified case, the misalignment between the training and test distributions relative to the target function (see the *source condition* in Assumption 5) appears to play a critical role empirically, as it can make covariate shift either benign or severely adversarial. A deeper understanding of this phenomenon may come from a more detailed analysis of the constants in the learning bounds and their influence on the overall rate (see eq. (20) and the interaction between the covariance operators of source and target distributions).

References

- Alaoui, A. and Mahoney, M. W. Fast randomized kernel ridge regression with statistical guarantees. In Advances in Neural Information Processing Systems, pp. 775–783, 2015.
- Bach, F. Sharp analysis of low-rank kernel matrix approximations. In Conference on learning theory, pp. 185–209. PMLR, 2013.
- Bach, F. On the equivalence between kernel quadrature rules and random feature expansions. The Journal of Machine Learning Research, 18(1):714–751, 2017.
- Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. Advances in neural information processing systems, 19, 2006.
- Ben-Israel, A. and Greville, T. N. Generalized inverses: theory and applications. Springer Science & Business Media, 2006.
- Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In Advances in neural information processing systems, pp. 161–168, 2008.
- Boucheron, S., Lugosi, G., and Massart, P. Concentration Inequalities: A Nonasymptotic Theory of Independence. Oxford University Press, Oxford, 2013.
- Calandriello, D., Lazaric, A., and Valko, M. Distributed adaptive sampling for kernel matrix approximation. In Artificial Intelligence and Statistics, pp. 1421–1429. PMLR, 2017.
- Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. Foundations of Computational Mathematics, 7:331–368, 2007.
- Cohen, M. B., Lee, Y. T., Musco, C., Musco, C., Peng, R., and Sidford, A. Uniform sampling for matrix approximation. In Proceedings of the 2015 conference on innovations in theoretical computer science, pp. 181–190, 2015.
- Cortes, C. and Mohri, M. Domain adaptation and sample bias correction theory and algorithm for regression. Theoretical Computer Science, 519:103–126, 2014.
- Cortes, C., Mohri, M., Riley, M., and Rostamizadeh, A. Sample selection bias correction theory. In International conference on algorithmic learning theory, pp. 38–53. Springer, 2008.
- Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. Advances in neural information processing systems, 23, 2010.
- De Vito, E., Rosasco, L., and Rudi, A. Regularization: From inverse problems to large-scale machine learning. Harmonic and Applied Analysis: From Radon Transforms to Machine Learning, pp. 245–296, 2021.
- Della Vecchia, A., Mourtada, J., De Vito, E., and Rosasco, L. Regularized erm on random subspaces. In International Conference on Artificial Intelligence and Statistics, pp. 4006–4014. PMLR, 2021.
- Della Vecchia, A., De Vito, E., Mourtada, J., and Rosasco, L. The nyström method for convex loss functions. Journal of Machine Learning Research, 25(360):1–60, 2024.
- Dinu, M.-C., Holzleitner, M., Beck, M., Nguyen, H. D., Huber, A., Eghbal-zadeh, H., Moser, B. A., Pereverzyev, S., Hochreiter, S., and Zellinger, W. Addressing parameter choice issues in unsupervised domain adaptation by aggregation. In The Eleventh International Conference on Learning Representations, 2023. URL <https://openreview.net/forum?id=M95oDwJXayG>.
- Drineas, P. and Mahoney, M. W. On the nyström method for approximating a gram matrix for improved kernel-based learning. journal of machine learning research, 6(Dec):2153–2175, 2005.
- Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. Fast approximation of matrix coherence and statistical leverage. The Journal of Machine Learning Research, 13(1):3475–3506, 2012.

- Fang, T., Lu, N., Niu, G., and Sugiyama, M. Rethinking importance weighting for deep learning under distribution shift. Advances in neural information processing systems, 33:11996–12007, 2020.
- Fujii, J., Fujii, M., Furuta, T., and Nakamoto, R. Norm inequalities equivalent to heinz inequality. Proceedings of the American Mathematical Society, 118(3):827–830, 1993.
- Gizewski, E. R., Mayer, L., Moser, B. A., Nguyen, D. H., Pereverzyev Jr, S., Pereverzyev, S. V., Shepeleva, N., and Zellinger, W. On a regularization of unsupervised domain adaptation in rkhs. Applied and Computational Harmonic Analysis, 57:201–227, 2022.
- Gogolashvili, D., Zecchin, M., Kanagawa, M., Kountouris, M., and Filippone, M. When is importance weighting correction needed for covariate shift adaptation? arXiv preprint arXiv:2303.04020, 2023.
- Guan, H. and Liu, M. Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering, 69(3):1173–1185, 2021.
- He, H., Queen, O., Koker, T., Cuevas, C., Tsiligkaridis, T., and Zitnik, M. Domain adaptation for time series under feature and label shifts. In International Conference on Machine Learning, pp. 12746–12774. PMLR, 2023.
- Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting sample selection bias by unlabeled data. Advances in neural information processing systems, 19, 2006.
- Jiang, J. and Zhai, C. Instance weighting for domain adaptation in nlp. ACL, 2007.
- Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In Advances in neural information processing systems, pp. 315–323, 2013.
- Kanamori, T. and Shimodaira, H. Active learning algorithm using the maximum weighted log-likelihood estimator. Journal of statistical planning and inference, 116(1):149–162, 2003.
- Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balasubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In International conference on machine learning, pp. 5637–5664. PMLR, 2021.
- Kpotufe, S. and Martinet, G. Marginal singularity and the benefits of labels in covariate-shift. The Annals of Statistics, 49(6):3299–3323, 2021.
- Kpotufe, S. and Sriperumbudur, B. K. Kernel sketching yields kernel jl. arXiv preprint arXiv:1908.05818, 2019.
- Kwapisz, J. R., Weiss, G. M., and Moore, S. A. Activity recognition using cell phone accelerometers. ACM SigKDD Explorations Newsletter, 12(2):74–82, 2011.
- Lei, Q., Hu, W., and Lee, J. Near-optimal linear regression under distribution shift. In International Conference on Machine Learning, pp. 6164–6174. PMLR, 2021.
- Liu, S., Yamada, M., Collier, N., and Sugiyama, M. Change-point detection in time-series data by relative density-ratio estimation. Neural Networks, 43:72–83, 2013.
- Ma, C., Pathak, R., and Wainwright, M. J. Optimally tackling covariate shift in rkhs-based nonparametric regression. The Annals of Statistics, 51(2):738–761, 2023.
- Mahoney, M. W. Randomized algorithms for matrices and data. Foundations and Trends® in Machine Learning, 3(2):123–224, 2011.
- Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. arXiv preprint arXiv:0902.3430, 2009.
- Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In Conference on Learning Theory, pp. 2294–2340. PMLR, 2019.

- Meanti, G., Carratino, L., Rosasco, L., and Rudi, A. Kernel methods through the roof: handling billions of points efficiently. Advances in Neural Information Processing Systems, 33:14410–14422, 2020.
- Myleiko, H. L. and Solodky, S. G. Regularized nystrom subsampling in covariate shift domain adaptation problems. Numerical Functional Analysis and Optimization, 45(3):165–188, 2024.
- Pathak, R., Ma, C., and Wainwright, M. A new similarity measure for covariate shift with applications to nonparametric regression. In International Conference on Machine Learning, pp. 17517–17530. PMLR, 2022.
- Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. Dataset shift in machine learning. Mit Press, 2022.
- Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In Advances in Neural Information Processing Systems 30, pp. 3215–3225, 2017.
- Rudi, A., Camoriano, R., and Rosasco, L. Less is more: Nyström computational regularization. arXiv preprint arXiv:1507.04717, 28, 2015.
- Rudi, A., Carratino, L., and Rosasco, L. Falcon: An optimal large scale kernel method. In Advances in Neural Information Processing Systems, pp. 3888–3898, 2017.
- Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. On fast leverage score sampling and optimal learning. In Advances in Neural Information Processing Systems, pp. 5672–5682, 2018.
- Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. Mathematical Programming, 162(1-2):83–112, 2017.
- Schmidt-Hieber, J. and Zolotarev, P. Local convergence rates of the nonparametric least squares estimator with applications to transfer learning. Bernoulli, 30(3):1845–1877, 2024.
- Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference, 90(2):227–244, 2000.
- Smale, S. and Zhou, D.-X. Learning theory estimates via integral operators and their approximations. Constructive approximation, 26(2):153–172, 2007.
- Smola, A. J. and Schölkopf, B. Sparse greedy matrix approximation for machine learning. In Proceedings of the seventeenth international conference on machine learning, pp. 911–918, 2000.
- Steinwart, I. and Christmann, A. Support vector machines. Springer Science & Business Media, 2008.
- Stisen, A., Blunck, H., Bhattacharya, S., Prentow, T. S., Kjærgaard, M. B., Dey, A., Sonne, T., and Jensen, M. M. Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition. In Proceedings of the 13th ACM conference on embedded networked sensor systems, pp. 127–140, 2015.
- Sugiyama, M. and Kawanabe, M. Machine learning in non-stationary environments: Introduction to covariate shift adaptation. MIT press, 2012.
- Sugiyama, M. and Ridgeway, G. Active learning in approximately linear regression based on conditional expectation of generalization error. Journal of Machine Learning Research, 7(1), 2006.
- Sugiyama, M., Suzuki, T., and Kanamori, T. Density ratio estimation in machine learning. Cambridge University Press, 2012.
- Sun, Y., Gilbert, A., and Tewari, A. But how does it work in theory? linear svm with random features. In Advances in Neural Information Processing Systems, pp. 3379–3388, 2018.
- Tørring, M. F., Logacjov, A., Brændvik, S. M., Ustad, A., Roeleveld, K., and Bardal, E. M. Validation of two novel human activity recognition models for typically developing children and children with cerebral palsy. Plos one, 19(9):e0308853, 2024.

- Ustad, A., Logacjov, A., Trollebø, S. Ø., Thingstad, P., Vereijken, B., Bach, K., and Maroni, N. S. Validation of an activity type recognition model classifying daily physical behavior in older adults: the har70+ model. Sensors, 23(5):2368, 2023.
- Vapnik, V. The Nature of Statistical Learning Theory. Springer Science & Business Media, 1999.
- Wang, X. and Sun, R. Domain adaptation of time series classification. IEEE Access, 2024.
- Wen, J., Yu, C.-N., and Greiner, R. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In International Conference on Machine Learning, pp. 631–639. PMLR, 2014.
- Wiens, D. P. Robust weights and designs for biased regression models: Least squares and generalized m-estimation. Journal of Statistical Planning and Inference, 83(2):395–412, 2000.
- Williams, C. and Seeger, M. Using the nyström method to speed up kernel machines. Advances in neural information processing systems, 13, 2000.
- Williams, C. K. and Seeger, M. Using the nyström method to speed up kernel machines. In Advances in neural information processing systems, pp. 682–688, 2001.
- Wilson, G., Doppa, J. R., and Cook, D. J. Multi-source deep domain adaptation with weak supervision for time-series sensor data. In Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 1768–1778, 2020.
- Woodruff, D. P. Sketching as a tool for numerical linear algebra. arXiv preprint arXiv:1411.4357, 2014.
- Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., and Sugiyama, M. Relative density-ratio estimation for robust distribution comparison. Neural computation, 25(5):1324–1370, 2013.
- Yamazaki, K., Kawanabe, M., Watanabe, S., Sugiyama, M., and Müller, K.-R. Asymptotic bayesian generalization error when training and test distributions are different. In Proceedings of the 24th international conference on Machine learning, pp. 1079–1086, 2007.
- Zhang, C., Zhang, L., and Ye, J. Generalization bounds for domain adaptation. Advances in neural information processing systems, 25, 2012.
- Zhang, T. Learning bounds for kernel regression using effective data dimensionality. Neural Computation, 17(9):2077–2098, 2005.

A Derivation of the estimators

We define some quantities we will need when deriving our estimators. Following Section 2, we define the sampling operator $\hat{Z}_m : \mathcal{H} \rightarrow \mathbb{R}^m$ associated with the subset $\{\tilde{x}_1, \dots, \tilde{x}_m\} \subset \{x_1, \dots, x_n\}$ as

$$\hat{Z}_m : \mathcal{H}_m \rightarrow \mathbb{R}^m, \quad (\hat{Z}_m f) = \langle f, K_{\tilde{x}_i} \rangle_{\mathcal{H}_m}, \quad (23)$$

with its adjoint

$$\hat{Z}_m^* : \mathbb{R}^m \rightarrow \mathcal{H}_m, \quad \hat{Z}_m^* \tilde{c} = \sum_{i=1}^m \tilde{c}_i K_{\tilde{x}_i}, \quad \forall \tilde{c} \in \mathbb{R}^m. \quad (24)$$

Moreover, consider the singular value decomposition (SVD) of \hat{Z}_m and \hat{Z}_m^*

$$\hat{Z}_m = UDV^*, \quad \hat{Z}_m^* = VDU^*$$

with $U : \mathbb{R}^k \rightarrow \mathbb{R}^m$, $D : \mathbb{R}^k \rightarrow \mathbb{R}^k$ the diagonal matrix of singular values sorted in non-decreasing order $D = \text{diag}(\sigma_1, \dots, \sigma_t)$ with $\sigma_1 \geq \dots \geq \sigma_t > 0$, $V : \mathbb{R}^k \rightarrow \mathcal{H}_m$, $k \leq m$ such that $U^*U = I_k$ and $V^*V = I_k$. The projection operator with range \mathcal{H}_m is given by $P_m = VV^*$. We used this fact in Section 4. We derive an expression for the KRR minimizer when considering IW correction and Nyström approximation.

Lemma 1 (Nyström IW-KRR estimator). *Given the minimization problem in (11), the unique minimizer can be written as*

$$\hat{f}_{\lambda,m}^w = V(V^* \hat{S}^* \hat{M}_w \hat{S}V + \lambda I)^{-1} V^* \hat{S}^* \hat{M}_w \hat{y} \quad (25)$$

where $\lambda > 0$, $\hat{y} = (y_1, \dots, y_n)^T$ and the matrix $\hat{M}_w = \text{diag}(w(x_1), \dots, w(x_n))$.

Proof. Eq. (11) can be rewritten as

$$\left\| \hat{M}_w^{\frac{1}{2}} (\hat{S}P_m f - \hat{y}) \right\|^2 = \left(\hat{M}_w^{\frac{1}{2}} (\hat{S}P_m f - \hat{y}) \right)^T \left(\hat{M}_w^{\frac{1}{2}} (\hat{S}P_m f - \hat{y}) \right). \quad (26)$$

Using first-order condition we have that

$$\frac{2}{n} P_m \hat{S}^* \hat{M}_w \hat{S} P_m \hat{f}_{\lambda,m}^w - \frac{2}{n} P_m \hat{S}^* \hat{M}_w \hat{y} + 2\lambda \hat{f}_{\lambda,m}^w = 0, \quad (27)$$

that is

$$(P_m \hat{S}^* \hat{M}_w \hat{S} P_m + \lambda n I) \hat{f}_{\lambda,m}^w = P_m \hat{S}^* \hat{M}_w \hat{y}. \quad (28)$$

Replacing $P_m = VV^*$ we obtain

$$V(V^* \hat{S}^* \hat{M}_w \hat{S}V + \lambda n I) V^* \hat{f}_{\lambda,m}^w = VV^* \hat{S}^* \hat{M}_w \hat{y}. \quad (29)$$

Multiplying both sides by V^* and using that $(V^* \hat{S}^* \hat{M}_w \hat{S}V + \lambda n I)$ is invertible, we obtain

$$V^* \hat{f}_{\lambda,m}^w = (V^* \hat{S}^* \hat{M}_w \hat{S}V + \lambda n I)^{-1} V^* \hat{S}^* \hat{M}_w \hat{y}. \quad (30)$$

The result is obtained multiplying both side by V and remembering that $\hat{f}_{\lambda,m}^w \in \mathcal{H}_m$. \square

We can express our estimator in an alternative form which will be useful in the actual implementation of the algorithm.

Lemma 2 (Nyström IW-KRR estimator, representer theorem form). *The above minimizer $\hat{f}_{\lambda,m}^w$ can be also written as*

$$\hat{f}_{\lambda,m}^w(x) = \sum_{i=1}^m \tilde{c}_i^w K(\tilde{x}_i, x), \quad \tilde{c}^w = (\hat{K}_{nm}^T \hat{M}_w \hat{K}_{nm} + n\lambda \hat{K}_{mm})^{-1} \hat{K}_{nm}^T \hat{M}_w \hat{y} \quad (31)$$

where $\hat{K}_{nm} = \hat{S} \hat{Z}_m^* \in \mathbb{R}^{n \times m}$, $(\hat{K}_{nm})_{ij} = K(x_i, \tilde{x}_j)$ and $\hat{K}_{mm} = \hat{Z}_m \hat{Z}_m^* \in \mathbb{R}^{m \times m}$, $(\hat{K}_{mm})_{ij} = K(\tilde{x}_i, \tilde{x}_j)$.

Proof. Eq. (25) can be rewritten as

$$\begin{aligned}\hat{f}_{\lambda,m}^w &= V(V^* \hat{S}^* \hat{M}_w \hat{S} V + \lambda n I)^{-1} V^* \hat{S}^* \hat{M}_w \hat{y} \\ &= V D U^* U D^{-1} (V^* \hat{S}^* \hat{M}_w \hat{S} V + \lambda n I)^{-1} D^{-1} U^* U D V^* \hat{S}^* \hat{M}_w \hat{y} \\ &= \hat{Z}_m^* (\hat{Z}_m \hat{S}^* \hat{M}_w \hat{S} \hat{Z}_m^* + \lambda \hat{Z}_m \hat{Z}_m^*)^\dagger \hat{Z}_m \hat{S}^* \hat{M}_w \hat{y}\end{aligned}\quad (32)$$

where we used $(FGH)^\dagger = H^\dagger G^{-1} F^\dagger$ (see the full-rank factorization of the pseudo-inverse (Ben-Israel & Greville, 2006)) with $F = UD$, $G = V^* \hat{S}^* \hat{M}_w \hat{S} V + \lambda n I$ and $H = DU^T$.

Using the definitions of \hat{K}_{mm} and \hat{K}_{nm} we have

$$(\hat{K}_{nm}^T \hat{M}_w \hat{K}_{nm} + \lambda n \hat{K}_{mm})^\dagger \hat{K}_{nm}^T \hat{M}_w \hat{y} = (\hat{Z}_m \hat{S}^* \hat{M}_w \hat{S} \hat{Z}_m^* + \lambda \hat{Z}_m \hat{Z}_m^*)^\dagger \hat{Z}_m \hat{S}^* \hat{M}_w \hat{y} \quad (33)$$

and substituting this expression above we get the result. \square

B Approximate leverage scores

Since in practice the leverage scores $l_i(\alpha)$ defined by (10) are onerous to compute, approximations $(\hat{l}_i(\alpha))_{i=1}^n$ have been considered (Drineas et al., 2012; Cohen et al., 2015; Alaoui & Mahoney, 2015). In particular, we used the following one.

Definition 2 (T -approximate leverage scores). *Let $(l_i(\alpha))_{i=1}^n$ be the leverage scores associated to the training set for a given α . Let $\delta > 0$, $t_0 > 0$ and $T \geq 1$. We say that $(\hat{l}_i(\alpha))_{i=1}^n$ are T -approximate leverage scores with confidence δ , when with probability at least $1 - \delta$,*

$$\frac{1}{T} l_i(\alpha) \leq \hat{l}_i(\alpha) \leq T l_i(\alpha), \quad \forall i \in \{1, \dots, n\}, \quad \alpha \geq t_0 \quad (34)$$

So, given T -approximate leverage score for $\alpha \geq t_0$, $\{\tilde{x}_1, \dots, \tilde{x}_m\}$ are sampled from the training set independently with replacement, and with probability to be selected given by $Q_\alpha(i) = \hat{l}_i(\alpha) / \sum_j \hat{l}_j(\alpha)$.

C Main proofs

To prove Theorem 1, we will need the following two propositions.

Proposition 1 (Empirical Effective Dimension). *Let $\hat{\mathcal{N}}_w(\lambda) = \text{Tr} \hat{\Sigma}_w \hat{\Sigma}_{\lambda,w}^{-1}$. Under assumption 3 for any $\delta > 0$ and $\left(\frac{128(W+\sigma^2)\log^2(4/\delta)}{n}\right)^{\frac{1}{1+q}} \leq \lambda \leq \|\Sigma\|$, then the following holds with probability $1 - \delta$,*

$$\frac{|\hat{\mathcal{N}}_w(\lambda) - \mathcal{N}_{\rho_X^{te}}(\lambda)|}{\mathcal{N}_{\rho_X^{te}}} \leq 2.$$

Proof. Let's call $\mathcal{N}_{\rho_X^{te}}(\lambda) = \mathcal{N}(\lambda)$ to simplify the notation. The proof partially follows the structure of Proposition 1 in Rudi et al. (2015), with some complications deriving from the presence of the (possibly unbounded) weights. Define $\hat{B}^w = \Sigma_\lambda^{-1/2} (\Sigma - \hat{\Sigma}_w) \Sigma_\lambda^{-1/2}$. Using Lemma 18 in Gogolashvili et al. (2023) we have that $\|\hat{B}^w\|_{\text{HS}} \leq 3/4$, when $n\lambda^{1+q} \geq 64(W + \sigma^2) \mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q} \log^2(\frac{2}{\delta})$. We can rewrite

$$\left| \hat{\mathcal{N}}_w(\lambda) - \mathcal{N}(\lambda) \right| = \left| \text{Tr} \left(\hat{\Sigma}_{\lambda,w}^{-1} \hat{\Sigma}_w - \Sigma \Sigma_\lambda^{-1} \right) \right| = \left| \lambda \text{Tr} \hat{\Sigma}_{\lambda,w}^{-1} (\hat{\Sigma}_w - \Sigma) \Sigma_\lambda^{-1} \right| \quad (35)$$

$$= \left| \lambda \text{Tr} \Sigma_\lambda^{-1/2} (\text{I} - \hat{B}^w)^{-1} \Sigma_\lambda^{-1/2} (\hat{\Sigma}_w - \Sigma) \Sigma_\lambda^{-1/2} \Sigma_\lambda^{-1/2} \right| \quad (36)$$

$$= \left| \lambda \text{Tr} \Sigma_\lambda^{-1/2} (\text{I} - \hat{B}^w)^{-1} \hat{B}^w \Sigma_\lambda^{-1/2} \right|. \quad (37)$$

Following [Rudi et al. \(2015\)](#) and using that, for any symmetric linear operator $X : \mathcal{H} \rightarrow \mathcal{H}$ the following identity holds

$$(\mathbf{I} - X)^{-1}X = X + X(\mathbf{I} - X)^{-1}X.$$

Applying the above identity with $X = \widehat{B}$

$$\begin{aligned} \lambda \left| \text{Tr} \Sigma_\lambda^{-1/2} (\mathbf{I} - \widehat{B}^w)^{-1} \widehat{B}^w \Sigma_\lambda^{-1/2} \right| &\leq \underbrace{\lambda \left| \text{Tr} \Sigma_\lambda^{-1/2} \widehat{B}^w \Sigma_\lambda^{-1/2} \right|}_{\mathbb{A}} \\ &\quad + \underbrace{\lambda \left| \text{Tr} \Sigma_\lambda^{-1/2} \widehat{B}^w (\mathbf{I} - \widehat{B}^w)^{-1} \widehat{B}^w \Sigma_\lambda^{-1/2} \right|}_{\mathbb{B}}. \end{aligned}$$

To find an upper bound for \mathbb{A} notice that

$$\mathbb{A} = \left| \mu - \frac{1}{n} \sum_{i=1}^n \xi_i \right|$$

with $\xi_i = \langle K_{x_i}, \lambda w(x_i) \Sigma_\lambda^{-2} K_{x_i} \rangle \in \mathbb{R}$ i.i.d. random variables with $i \in [n]$ and $\mu = \mathbb{E}[\xi_i]$. Using Lemma 18 in [Gogolashvili et al. \(2023\)](#) and a general version of Bernstein inequality requiring, instead of boundedness, only an appropriate control of moments ([Boucheron et al., 2013](#)), we have with probability greater than $1 - \delta$

$$\mathbb{A} \leq 4 \left(\frac{W}{\lambda n} + \sigma \sqrt{\frac{\mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q}}{\lambda^{1+q} n}} \right) \log \left(\frac{2}{\delta} \right).$$

As regards \mathbb{B} , write $\mathbb{B} = \|Q\|_{HS}^2$ where $Q = \lambda^{1/2} \Sigma_\lambda^{-1/2} \widehat{B}^w (\mathbf{I} - \widehat{B}^w)^{-1/2}$, moreover

$$\|Q\|_{HS}^2 \leq \left\| \lambda^{1/2} \Sigma_\lambda^{-1/2} \right\|^2 \left\| \widehat{B}^w \right\|_{HS}^2 \left\| (\mathbf{I} - \widehat{B}^w)^{-1/2} \right\|^2 \leq 4 \|\widehat{B}^w\|_{HS}^2,$$

since $\left\| (\mathbf{I} - \widehat{B}^w)^{-1/2} \right\|^2 \leq (1 - \|\widehat{B}^w\|)^{-1} \leq 4$, for $n\lambda^{1+q} \geq 64 (W + \sigma^2) \mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q} \log^2 \left(\frac{2}{\delta} \right)$.

Using again Lemma 18 in [Gogolashvili et al. \(2023\)](#) and a version of the Bernstein inequality for Hilbert space-valued random variables (see for example [Caponnetto & De Vito \(2007\)](#)) we obtain with probability greater than $1 - \delta$

$$\mathbb{B} \leq 16 \left(\frac{W}{\lambda n} + \sigma \sqrt{\frac{\mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q}}{\lambda^{1+q} n}} \right)^2 \log^2 \left(\frac{2}{\delta} \right)$$

Putting all together, with probability $1 - \delta$:

$$\left| \widehat{\mathcal{N}}_w(\lambda) - \mathcal{N}(\lambda) \right| \leq 4 \left(\frac{W}{\lambda n} + \sigma \sqrt{\frac{\mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q}}{\lambda^{1+q} n}} \right) \log \left(\frac{4}{\delta} \right) + 16 \left(\frac{W}{\lambda n} + \sigma \sqrt{\frac{\mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q}}{\lambda^{1+q} n}} \right)^2 \log^2 \left(\frac{4}{\delta} \right) \quad (38)$$

$$\leq 4 \left(\frac{W}{\lambda n} + \sigma \sqrt{\frac{1}{\lambda^{\gamma(1-q)+1+q} n}} \right) \log \left(\frac{4}{\delta} \right) + 16 \left(\frac{W}{\lambda n} + \sigma \sqrt{\frac{1}{\lambda^{\gamma(1-q)+1+q} n}} \right)^2 \log^2 \left(\frac{4}{\delta} \right) \quad (39)$$

Using that $\mathcal{N}(\lambda) \geq \|\Sigma \Sigma_\lambda^{-1}\| \geq 1/2$ if $\lambda \leq \|\Sigma\|$ we have

$$|\hat{\mathcal{N}}_w(\lambda) - \mathcal{N}(\lambda)| \quad (40)$$

$$\leq \left(4 \left(\frac{W}{\mathcal{N}_{\rho_X^{te}}(\lambda) \lambda n} + \sigma \sqrt{\frac{1}{\mathcal{N}_{\rho_X^{te}}(\lambda)^{1+q} \lambda^{1+q} n}} \right) \log \left(\frac{4}{\delta} \right) + \right. \quad (41)$$

$$\left. + 16 \left(\frac{W}{\mathcal{N}_{\rho_X^{te}}(\lambda) \lambda n} + \sigma \sqrt{\frac{1}{\mathcal{N}_{\rho_X^{te}}(\lambda)^{1+q} \lambda^{1+q} n}} \right)^2 \log^2 \left(\frac{4}{\delta} \right) \right) \mathcal{N}_{\rho_X^{te}}(\lambda) \quad (42)$$

$$\leq \left(4 \left(\frac{2W}{\lambda n} + \sigma \sqrt{\frac{4}{\lambda^{1+q} n}} \right) \log \left(\frac{4}{\delta} \right) + 16 \left(\frac{2W}{\lambda n} + \sigma \sqrt{\frac{4}{\lambda^{1+q} n}} \right)^2 \log^2 \left(\frac{4}{\delta} \right) \right) \mathcal{N}_{\rho_X^{te}}(\lambda). \quad (43)$$

Then for $\left(\frac{256(W+\sigma^2) \log^2(4/\delta)}{n} \right)^{\frac{1}{1+q}} \leq \lambda \leq \|\Sigma\|$ with probability $1 - \delta$

$$|\hat{\mathcal{N}}_w(\lambda) - \mathcal{N}(\lambda)| \leq 2\mathcal{N}(\lambda)$$

□

Proposition 2 (Nyström approximation with ALS sampling). *Let $(\hat{l}_i(t))_{i=1}^n$ be the collection of approximate leverage scores. Let $\lambda > 0$ and P_λ be defined as $P_\lambda(i) = \hat{l}_i(\lambda) / \sum_{j \in N} \hat{l}_j(\lambda)$ for any $i \in N$ with $N = \{1, \dots, n\}$. Let $\mathfrak{I} = (i_1, \dots, i_m)$ be a collection of indices independently sampled with replacement from N according to the probability distribution P_λ . Let P_m be the projection operator on the subspace $\mathcal{H}_m = \text{span}\{K_{x_j} \mid j \in J\}$ and J be the subcollection of \mathfrak{I} with all the duplicates removed. Under Assumption 3 and 6, for any $\delta > 0$ the following holds with probability $1 - 2\delta$*

$$\|(I - P_m)\Sigma_\lambda^{1/2}\| \leq 3\lambda$$

when the following conditions are satisfied:

- there exists $T \geq 1$ and $\lambda_0 > 0$ such that $(\hat{l}_i(t))_{i=1}^n$ are T -approximate leverage scores for any $t \geq \lambda_0$ (see definition 2),
- $\lambda_0 \vee \left(\frac{256(W+\sigma^2) \log^2(4/\delta)}{n} \right)^{\frac{1}{\gamma(1-q)+1+q}} \leq \lambda \leq \|\Sigma\|$,
- $m \geq 144T^2 \mathcal{N}_{\rho_X^{te}}(\lambda) \log \frac{8n}{\delta}$.

Proof. Let's call $\Sigma = \Sigma_{\rho_X^{te}}$ to simplify the notation. Define $\tau = \delta/4$. Next, define the diagonal matrix $H \in \mathbb{R}^{n \times n}$ with $(H)_{ii} = 0$ when $P_\lambda(i) = 0$ and $(H)_{ii} = \frac{nq(i)}{mP_\lambda(i)}$ when $P_\lambda(i) > 0$, where $q(i)$ is the number of times the index i is present in the collection \mathfrak{I} . We have that

$$\hat{S}_w^* H \hat{S}_w = \frac{1}{m} \sum_{i=1}^n w(x_i) \frac{q(i)}{P_\lambda(i)} K_{x_i} \otimes K_{x_i} = \frac{1}{m} \sum_{j \in J} w(x_j) \frac{q(j)}{P_\lambda(j)} K_{x_j} \otimes K_{x_j}.$$

Now, considering that $\frac{q(j)}{P_\lambda(j)} > 0$ for any $j \in J$, thus $\text{ran } \hat{S}_w^* H \hat{S}_w = \mathcal{H}_m$. Therefore, by using Prop. 3 and 7 in Rudi et al. (2015), we exploit the fact that the range of P_m is the same of $\hat{S}_w^* H \hat{S}_w$, to obtain

$$\left\| (I - P_m) \Sigma_\lambda^{1/2} \right\|^2 \leq \lambda \left\| \left(\hat{S}_w^* H \hat{S}_w + \lambda I \right)^{-1/2} \Sigma_\lambda^{1/2} \right\|^2 \leq \frac{\lambda}{1 - \beta(\lambda)},$$

with $\beta(\lambda) = \lambda_{\max} \left(\Sigma_\lambda^{-1/2} \left(\Sigma - \hat{S}_w^* H \hat{S}_w \right) \Sigma_\lambda^{-1/2} \right)$. Considering that the function $(1 - x)^{-1}$ is increasing on $-\infty < x < 1$, in order to bound $\lambda/(1 - \beta(\lambda))$ we need an upperbound for $\beta(\lambda)$. Here we split $\beta(\lambda)$ in the following way,

$$\beta(\lambda) \leq \underbrace{\lambda_{\max} \left(\Sigma_\lambda^{-1/2} \left(\Sigma - \hat{\Sigma}_w \right) \Sigma_\lambda^{-1/2} \right)}_{\beta_1(\lambda)} + \underbrace{\lambda_{\max} \left(\Sigma_\lambda^{-1/2} \left(\hat{\Sigma}_w - \hat{S}_w^* H \hat{S}_w \right) \Sigma_\lambda^{-1/2} \right)}_{\beta_2(\lambda)}.$$

β_1 can be bounded as in eq. (44).
As regards β_2 :

$$\begin{aligned}\beta_2(\lambda) &\leq \left\| \Sigma_\lambda^{-1/2} \left(\hat{\Sigma}_w - \hat{S}_w^* H \hat{S}_w \right) \Sigma_\lambda^{-1/2} \right\| \\ &\leq \left\| \Sigma_\lambda^{-1/2} \hat{\Sigma}_{w\lambda}^{1/2} \right\|^2 \left\| \hat{\Sigma}_{w\lambda}^{-1/2} \left(\hat{\Sigma}_w - \hat{S}_w^* H \hat{S}_w \right) \hat{\Sigma}_{w\lambda}^{-1/2} \right\|.\end{aligned}$$

Let

$$\beta_3(\lambda) = \left\| \hat{\Sigma}_{w\lambda}^{-1/2} \left(\hat{\Sigma}_w - \hat{S}_w^* H \hat{S}_w \right) \hat{\Sigma}_{w\lambda}^{-1/2} \right\| = \left\| \hat{\Sigma}_{w\lambda}^{-1/2} \hat{S}_w^* (I - H) \hat{S}_w \hat{\Sigma}_{w\lambda}^{-1/2} \right\|.$$

Note that $\hat{S}_w \hat{\Sigma}_{w\lambda}^{-1} \hat{S}_w^* = \hat{S}_w (\hat{S}_w^* \hat{S}_w + \lambda I)^{-1} \hat{S}_w^* = \left(\hat{K}_w + \lambda n I \right)^{-1} \hat{K}_w$ since $\hat{K}_w = n \hat{S}_w \hat{S}_w^*$, with $(\hat{K}_w)_{ij} = w(x_i)^{1/2} w(x_j)^{1/2} K(x_i, x_j)$.

Thus, if we let UDU^\top be the eigendecomposition of \hat{K}_w , we have that $\left(\hat{K}_w + \lambda n I \right)^{-1} \hat{K}_w = U(D + \lambda n I)^{-1} D U^\top$ and thus $\hat{S}_w \hat{\Sigma}_{w\lambda}^{-1} \hat{S}_w^* = U(D + \lambda n I)^{-1} D U^\top$. In particular this implies that $\hat{S}_w \hat{\Sigma}_{w\lambda}^{-1} \hat{S}_w^* = U \hat{Q}^{1/2} \hat{Q}^{1/2} U^\top$ with $\hat{Q} = (D + \lambda n I)^{-1} D$. Therefore we have

$$\beta_3(\lambda) = \left\| \hat{\Sigma}_{w\lambda}^{-1/2} \hat{S}_w^* (I - H) \hat{S}_w \hat{\Sigma}_{w\lambda}^{-1/2} \right\| = \left\| \hat{Q}^{1/2} U^\top (I - H) U \hat{Q}^{1/2} \right\|.$$

Consider the matrix $A = \hat{Q}^{1/2} U^\top$ and let a_i be the i -th column of A , and e_i be the i -th canonical basis vector for each $i \in N$. We prove that $\|a_i\|^2 = l_i(\lambda)$, the true leverage score, since

$$\|a_i\|^2 = \left\| \hat{Q}^{1/2} U^\top e_i \right\|^2 = e_i^\top U \hat{Q} U^\top e_i = \left(\left(\hat{K}_w + \lambda n I \right)^{-1} \hat{K}_w \right)_{ii} = l_i(\lambda)$$

Considering that $\sum_{k=1}^n \frac{q(k)}{\bar{P}_\lambda(k)} a_k a_k^\top = \sum_{i \in \mathcal{I}} \frac{1}{\bar{P}_\lambda(i)} a_i a_i^\top$, we have

$$\beta_3(\lambda) = \left\| A A^\top - \frac{1}{m} \sum_{i \in \mathcal{I}} \frac{1}{P_\lambda(i)} a_i a_i^\top \right\|.$$

Moreover, by the T -approximation property of the approximate leverage scores (see Def. 1 in [Rudi et al. \(2015\)](#)), we have that for all $i \in \{1, \dots, n\}$, when $\lambda \geq \lambda_0$, the following holds with probability $1 - \delta$

$$P_\lambda(i) = \frac{\hat{l}_i(\lambda)}{\sum_j \hat{l}_j(\lambda)} \geq T^{-2} \frac{l_i(\lambda)}{\sum_j l_j(\lambda)} = T^{-2} \frac{\|a_i\|^2}{\text{Tr } A A^\top}.$$

Then, we can apply Prop. 9 in [Rudi et al. \(2015\)](#), so that, after a union bound, we obtain the following inequality with probability $1 - \delta - \tau$:

$$\beta_3(\lambda) \leq \frac{2\|A\|^2 \log \frac{2n}{\tau}}{3m} + \sqrt{\frac{2\|A\|^2 T^2 \text{Tr } A A^\top \log \frac{2n}{\tau}}{m}} \leq \frac{2 \log \frac{2n}{\tau}}{3m} + \sqrt{\frac{2T^2 \hat{\mathcal{N}}_w(\lambda) \log \frac{2n}{\tau}}{m}},$$

where the last step follows from $\|A\|^2 = \left\| \left(\hat{K}_w + \lambda n I \right)^{-1} \hat{K}_w \right\| \leq 1$ and $\text{Tr}(A A^\top) = \text{Tr}(\hat{\Sigma}_{w\lambda}^{-1} \hat{\Sigma}_w) := \hat{\mathcal{N}}_w(\lambda)$. Applying proposition 1 we have that $\hat{\mathcal{N}}_w(\lambda) \leq 3\mathcal{N}_{\rho_X^{te}}(\lambda)$ with probability $1 - \tau$, when $\left(\frac{128(W + \sigma^2) \log^2(4/\delta)}{n} \right)^{\frac{1}{1+q}} \leq \lambda \leq \|\Sigma\|$. Thus, by taking a union bound again, we have

$$\beta_3(\lambda) \leq \frac{2 \log \frac{2n}{\tau}}{3m} + \sqrt{\frac{16T^2 \mathcal{N}_{\rho_X^{te}}(\lambda) \log \frac{2n}{\tau}}{m}}$$

with probability $1 - 2\tau - \delta$. The last step is to bound $\left\| \Sigma_\lambda^{-1/2} \hat{\Sigma}_{w\lambda}^{1/2} \right\|^2$, as follows

$$\left\| \Sigma_\lambda^{-1/2} \hat{\Sigma}_{w\lambda}^{1/2} \right\|^2 = \left\| \Sigma_\lambda^{-1/2} \hat{\Sigma}_{w\lambda} \Sigma_\lambda^{-1/2} \right\| = \left\| I + \Sigma_\lambda^{-1/2} \left(\hat{\Sigma}_w - \Sigma \right) \Sigma_\lambda^{-1/2} \right\| \leq 1 + \eta,$$

with $\eta = \left\| \Sigma_\lambda^{-1/2} (\hat{\Sigma}_w - \Sigma) \Sigma_\lambda^{-1/2} \right\|$. We can bound η using Lemma 18 in [Gogolashvili et al. \(2023\)](#) (see eq. (44)). Finally, by collecting the above results and taking a union bound we have

$$\beta(\lambda) \leq 4 \left(\frac{W}{\lambda n} + \sqrt{\frac{\sigma^2 \mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q}}{\lambda^{1+q} n}} \right) \log \left(\frac{2}{\tau} \right) + (1 + \eta) \left(\frac{2 \log \frac{2n}{\tau}}{3m} + \sqrt{\frac{16T^2 \mathcal{N}_{\rho_X^{te}}(\lambda) \log \frac{2n}{\tau}}{m}} \right),$$

with probability $1 - 4\tau - \delta = 1 - 2\delta$ when $\left(\frac{256(W + \sigma^2) \log^2(4/\delta)}{n} \right)^{\frac{1}{1+q}} \leq \lambda \leq \|\Sigma\|$. Note that, if we select $\left(\frac{256(W + \sigma^2) \log^2(4/\delta)}{n} \right)^{\frac{1}{\gamma(1-q)+1+q}} \leq \lambda \leq \|\Sigma\|$, and $m \geq 144T^2 \mathcal{N}_{\rho_X^{te}}(\lambda) \log \frac{8n}{\delta}$, we have $\beta(\lambda) \leq 5/6$, so that

$$\left\| (I - P_m) \Sigma^{1/2} \right\|^2 \leq 6\lambda$$

with probability $1 - 2\delta$. □

We can now proceed with the proof of Theorem 1.

Proof of Theorem 1

proof of Theorem 1. We split the excess risk as

$$\begin{aligned} \left| \mathcal{R}(\hat{f}_{\lambda,m}^w) - \mathcal{R}(f_{\mathcal{H}}) \right|^{1/2} &= \left\| \hat{f}_{\lambda,m}^w - f_{\mathcal{H}} \right\|_{\rho_X^{te}} = \left\| \Sigma^{1/2} (\hat{f}_{\lambda,m}^w - f_{\mathcal{H}}) \right\|_{\mathcal{H}} \\ &= \left\| \Sigma^{1/2} (V(V^* \hat{\Sigma}_w V + \lambda I)^{-1} V^* \hat{S}^* M_w y - f_{\mathcal{H}}) \right\|_{\mathcal{H}} \\ &\leq \underbrace{\left\| \Sigma^{1/2} V(V^* \hat{\Sigma}_w V + \lambda I)^{-1} V^* \hat{S}^* M_w (y - \hat{S} f_{\mathcal{H}}) \right\|_{\mathcal{H}}}_{\mathbb{A}} + \\ &\quad + \underbrace{\left\| \Sigma^{1/2} (I - V(V^* \hat{\Sigma}_w V + \lambda I)^{-1} V^* \hat{\Sigma}_w) f_{\mathcal{H}} \right\|_{\mathcal{H}}}_{\mathbb{B}} \end{aligned}$$

Term \mathbb{A}

$$\mathbb{A} \leq \underbrace{\left\| \Sigma^{1/2} \hat{\Sigma}_{w\lambda}^{-1/2} \right\|}_{\mathbb{A}_1} \underbrace{\left\| \hat{\Sigma}_{w\lambda}^{1/2} V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda}^{1/2} \right\|}_{\mathbb{A}_2} \underbrace{\left\| \hat{\Sigma}_{w\lambda}^{-1/2} \Sigma_\lambda^{1/2} \right\|}_{\mathbb{A}_3 = \beta} \underbrace{\left\| \Sigma_\lambda^{-1/2} \hat{S}^* M_w (y - \hat{S} f_{\mathcal{H}}) \right\|_{\mathcal{H}}}_{\mathbb{A}_4}$$

- \mathbb{A}_1 :

$$\mathbb{A}_1 \leq \underbrace{\left\| \Sigma^{1/2} \Sigma_\lambda^{-1/2} \right\|}_{\leq 1} \underbrace{\left\| \Sigma_\lambda^{1/2} \hat{\Sigma}_{w\lambda}^{-1/2} \right\|}_{\beta} \leq \beta$$

- \mathbb{A}_2 : using lemma 8 in [Rudi et al. \(2015\)](#) it's easy to show that $\left\| \hat{\Sigma}_{w\lambda}^{1/2} V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda}^{1/2} \right\|^2 = \left\| \hat{\Sigma}_{w\lambda}^{1/2} V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda}^{1/2} \right\|$, and therefore the only possible values for \mathbb{A}_2 are 0 and 1. Then

$$\mathbb{A}_2 \leq 1$$

- $\mathbb{A}_3 = \beta$: using proposition 7 in [Rudi et al. \(2015\)](#) we have

$$\beta \leq \frac{1}{1-b}$$

$$\text{if } b = \lambda_{\max} \left[\Sigma_\lambda^{-1/2} (\Sigma - \hat{\Sigma}_w) \Sigma_\lambda^{-1/2} \right] < 1.$$

Applying Lemma 18 in [Gogolashvili et al. \(2023\)](#) we get, with probability greater than $1 - \delta$

$$\lambda_{\max} \left[\Sigma_{\lambda}^{-1/2} (\Sigma - \hat{\Sigma}_w) \Sigma_{\lambda}^{-1/2} \right] \leq \left\| \Sigma_{\lambda}^{-\frac{1}{2}} (\Sigma - \hat{\Sigma}_w) \Sigma_{\lambda}^{-\frac{1}{2}} \right\|_{\text{HS}} \leq 4 \left(\frac{W}{\lambda n} + \sqrt{\frac{\sigma^2 \mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q}}{\lambda^{1+q} n}} \right) \log \left(\frac{2}{\delta} \right) \quad (44)$$

and for $n\lambda^{1+q} \geq 64 (W + \sigma^2) \mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q} \log^2 \left(\frac{2}{\delta} \right)$, with probability greater than $1 - \delta$ the above quantity is less or equal than $3/4$.

- \mathbb{A}_4 :

To control this term we use lemma 19 in [Gogolashvili et al. \(2023\)](#), where $\xi_i = \Sigma_{\lambda}^{-1/2} w(x_i) K_{x_i} y_i$. We obtain that, with probability greater or equal than $1 - \delta$

$$\left\| \frac{1}{n} \sum \xi_i - \mathbb{E}[\xi] \right\|_{\mathcal{H}} = \left\| \Sigma_{\lambda}^{-1/2} \hat{S}^* M_w(y - \hat{S} f_{\mathcal{H}}) \right\|_{\mathcal{H}} \leq 4B \left(\frac{W}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2 \mathcal{N}_{\rho_X^{te}}(\lambda)^{1-q}}{n\lambda^q}} \right) \log \left(\frac{2}{\delta} \right).$$

Term \mathbb{B}

As regards \mathbb{B} , following [Rudi et al. \(2015\)](#) we proceed as follows. Noting that $V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda} V V^* = V V^*$, we have

$$\begin{aligned} I - V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_w &= I - V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda} + \lambda V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \\ &= I - V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda} V V^* - V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda} (I - V V^*) + \lambda V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \\ &= (I - V V^*) + \lambda V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* - V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda} (I - V V^*). \end{aligned}$$

By assumption 5, we have $\left\| \Sigma_{w\lambda}^{-(r-1/2)} f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq \left\| \Sigma_{w\lambda}^{-r} f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq \left\| \Sigma_w^{-r} f_{\mathcal{H}} \right\|_{\mathcal{H}} \leq R$. Define $r' := r - 1/2$ to simplify the notation. Using the above decomposition, we can rewrite term \mathbb{B} as

$$\begin{aligned} \mathbb{B} &\leq \left\| \Sigma^{1/2} \left(I - V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_w \right) \Sigma_{w\lambda}^{r'} \right\| \left\| \Sigma_{w\lambda}^{-r'} f_{\mathcal{H}} \right\|_{\mathcal{H}} \\ &\leq R \left\| \Sigma^{1/2} \Sigma_{\lambda}^{-1/2} \right\| \left\| \Sigma_{\lambda}^{1/2} (I - V V^*) \Sigma_{\lambda}^{r'} \right\| + R \lambda \left\| \Sigma^{1/2} \hat{\Sigma}_{w\lambda}^{-1/2} \right\| \left\| \hat{\Sigma}_{w\lambda}^{1/2} V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \Sigma_{\lambda}^{r'} \right\| \\ &\quad + R \left\| \Sigma^{1/2} \hat{\Sigma}_{w\lambda}^{-1/2} \right\| \left\| \hat{\Sigma}_{w\lambda}^{1/2} V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \hat{\Sigma}_{w\lambda}^{1/2} \right\| \left\| \hat{\Sigma}_{w\lambda}^{1/2} \Sigma_{\lambda}^{-1/2} \right\| \left\| \Sigma_{\lambda}^{1/2} (I - V V^*) \Sigma_{\lambda}^{r'} \right\| \\ &\leq R(1 + \beta\theta) \underbrace{\left\| \Sigma_{\lambda}^{1/2} (I - V V^*) \Sigma_{\lambda}^{r'} \right\|}_{\mathbb{B}.1} + R\beta\lambda \underbrace{\left\| \hat{\Sigma}_{w\lambda}^{1/2} V(V^* \hat{\Sigma}_{w\lambda} V)^{-1} V^* \Sigma_{\lambda}^{r'} \right\|}_{\mathbb{B}.2}, \end{aligned}$$

$$\text{with } \theta = \left\| \hat{\Sigma}_{w\lambda}^{1/2} \Sigma_{\lambda}^{-1/2} \right\|.$$

- $\mathbb{B}.1$:

$$\mathbb{B}.1 = \left\| \Sigma_{\lambda}^{1/2} (I - V V^*)^2 \Sigma_{\lambda}^{r'} \right\| \leq \left\| \Sigma_{\lambda}^{1/2} (I - V V^*) \right\| \left\| (I - V V^*) \Sigma_{\lambda}^{r'} \right\|.$$

Since $V V^*$ is a projection operator, we have that $(I - V V^*) = (I - V V^*)^s$, for any $s > 0$, therefore, by applying Cordes inequality (see proposition 6) to $\left\| (I - V V^*) \Sigma_{\lambda}^{r'} \right\|$, we have

$$\left\| (I - V V^*) \Sigma_{\lambda}^{r'} \right\| = \left\| (I - V V^*)^{2r} \Sigma_{\lambda}^{\frac{1}{2} 2r'} \right\| \leq \left\| (I - V V^*) \Sigma_{\lambda}^{1/2} \right\|^{2r'}.$$

This term can be controlled using Proposition 2 above.

• B.2:

$$\begin{aligned}
\text{B.2} &\leq \lambda \left\| \widehat{\Sigma}_{w\lambda}^{1/2} V (V^* \widehat{\Sigma}_{w\lambda} V)^{-1} V^* \widehat{\Sigma}_{w\lambda}^{r'} \right\| \left\| \widehat{\Sigma}_{w\lambda}^{-r'} \Sigma_\lambda^{r'} \right\| \\
&\leq \lambda \left\| \widehat{\Sigma}_{w\lambda}^{1/2} V (V^* \widehat{\Sigma}_{w\lambda} V)^{-1} V^* \widehat{\Sigma}_{w\lambda}^{r'} \right\| \left\| \widehat{\Sigma}_{w\lambda}^{-1/2} \Sigma_\lambda^{1/2} \right\|^{2r'} \\
&\leq \beta^{2r'} \lambda \left\| \left(V^* \widehat{\Sigma}_{w\lambda} V \right)^{1/2} \left(V^* \widehat{\Sigma}_{w\lambda} V \right)^{-1} \left(V^* \widehat{\Sigma}_{w\lambda} V \right)^{r'} \right\| \\
&= \beta^{2r'} \lambda \left\| \left(V^* \widehat{\Sigma}_w V + \lambda I \right)^{-(1/2-r')} \right\| \leq \beta \lambda^{1/2+r'} = \beta \lambda^r,
\end{aligned}$$

where the first step is obtained multiplying and dividing by $\widehat{\Sigma}_{w\lambda}^{r'}$, the second step by applying Cordes inequality, the third step by Prop. 6 in [Rudi et al. \(2015\)](#).

Putting all together, for $\left(\frac{256(W+\sigma^2)\log^2(4/\delta)}{n} \right)^{\frac{1}{\gamma(1-q)+1+q}} \leq \lambda \leq \|\Sigma\|$, and $m \geq 144T^2Q\lambda^{-\gamma} \log \frac{8n}{\delta}$, with probability greater or equal than $1 - \delta$

$$\left| \mathcal{R}(\widehat{f}_{\lambda,m}^w) - \mathcal{R}(f_{\mathcal{H}}) \right|^{1/2} \leq 4B\beta^2 \left(\frac{W}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2}{n\lambda^{\gamma(1-q)+q}}} \right) \log \left(\frac{8}{\delta} \right) + 3R(1 + \beta\theta)\lambda^r + R\beta^2\lambda^r.$$

Under the above conditions on λ we have that $\beta \leq 4$ and $\theta \leq 2$ and then for $\left(\frac{256(W+\sigma^2)\log^2(4/\delta)}{n} \right)^{\frac{1}{\gamma(1-q)+1+q}} \leq \lambda \leq \|\Sigma\|$, and $m \geq 144T^2Q\lambda^{-\gamma} \log \frac{8n}{\delta}$, with probability greater or equal than $1 - \delta$

$$\left| \mathcal{R}(\widehat{f}_{\lambda,m}^w) - \mathcal{R}(f_{\mathcal{H}}) \right|^{1/2} \leq 64B \left(\frac{W}{n\sqrt{\lambda}} + \sqrt{\frac{\sigma^2}{n\lambda^{\gamma(1-q)+q}}} \right) \log \left(\frac{8}{\delta} \right) + 43R\lambda^r.$$

□

D Known results

In this section, we derive tight bounds for the effective dimension $\mathcal{N}_\rho(\lambda)$ defined in Definition 1 when assuming polynomial decay of the eigenvalues $\sigma_j(\Sigma)$ of the covariance operator Σ .

Proposition 3 (Polynomial eigenvalues decay, Proposition 3 in [Caponnetto & De Vito \(2007\)](#)). *If for some $\gamma \in \mathbb{R}^+$ and $1 < \beta < +\infty$*

$$\sigma_i \leq \gamma i^{-\beta}$$

then

$$\mathcal{N}_\rho(\lambda) \leq \gamma \frac{\beta}{\beta - 1} \lambda^{-1/\beta} \tag{45}$$

Proof. Since the function $\sigma/(\sigma + \lambda)$ is increasing in σ and using the spectral theorem $\Sigma = UDU^*$ combined with the fact that $\text{Tr}(UDU^*) = \text{Tr}(U(U^*D)) = \text{Tr}D$

$$\mathcal{N}_\rho(\lambda) = \text{Tr}(\Sigma(\Sigma + \lambda I)^{-1}) = \sum_{i=1}^{\infty} \frac{\sigma_i}{\sigma_i + \lambda} \leq \sum_{i=1}^{\infty} \frac{\gamma}{\gamma + i^\beta \lambda} \tag{46}$$

The function $\gamma/(\gamma + x^\beta \lambda)$ is positive and decreasing, so

$$\begin{aligned}
\mathcal{N}_\rho(\lambda) &\leq \int_0^\infty \frac{\gamma}{\gamma + x^\beta \lambda} dx \\
&= \alpha^{-1/\beta} \int_0^\infty \frac{\gamma}{\gamma + \tau^\beta} d\tau \\
&\leq \gamma \frac{\beta}{\beta - 1} \lambda^{-1/\beta}
\end{aligned} \tag{47}$$

since $\int_0^\infty (\gamma + \tau^\beta)^{-1} \leq \beta/(\beta - 1)$. □

Further improvements can be found assuming exponential decay of the eigenvalues $\sigma_j(\Sigma)$ of Σ .

Proposition 4 (Exponential eigenvalues decay, Proposition 5 in [Della Vecchia et al. \(2021\)](#)). *If for some $\gamma, \beta \in \mathbb{R}^+$ $\sigma_i \leq \gamma e^{-\beta i}$ then*

$$\mathcal{N}_\rho(\lambda) \leq \frac{\log(1 + \gamma/\lambda)}{\beta}. \quad (48)$$

Proof.

$$\mathcal{N}_\rho(\lambda) = \sum_{i=1}^{\infty} \frac{\sigma_i}{\sigma_i + \lambda} = \sum_{i=1}^{\infty} \frac{1}{1 + \lambda/\sigma_i} \leq \sum_{i=1}^{\infty} \frac{1}{1 + \lambda' e^{\beta i}} \leq \int_0^{+\infty} \frac{1}{1 + \lambda' e^{\beta x}} dx, \quad (49)$$

where $\lambda' = \lambda/\gamma$. Using the change of variables $t = e^{\beta x}$ we get

$$\begin{aligned} (49) &= \frac{1}{\beta} \int_1^{+\infty} \frac{1}{1 + \lambda' t} \frac{1}{t} dt = \frac{1}{\beta} \int_1^{+\infty} \left[\frac{1}{t} - \frac{\lambda'}{1 + \lambda' t} \right] dt = \frac{1}{\beta} \left[\log t - \log(1 + \lambda' t) \right]_1^{+\infty} \\ &= \frac{1}{\beta} \left[\log \left(\frac{t}{1 + \lambda' t} \right) \right]_1^{+\infty} = \frac{1}{\beta} \left[\log(1/\lambda') + \log(1 + \lambda') \right]. \end{aligned} \quad (50)$$

So we finally obtain

$$\mathcal{N}_\rho(\lambda) \leq \frac{1}{\beta} \left[\log(\gamma/\lambda) + \log(1 + \lambda/\gamma) \right] = \frac{\log(1 + \gamma/\lambda)}{\beta}. \quad (51)$$

□

Next proposition gives some relations between weights and covariance operators.

Proposition 5. *Suppose that ρ_X^{te} is absolutely continuous with respect to ρ_X^v , and the Radon-Nikodym derivative $d\rho_X^{te}/d\rho_X^v$ is bounded by $G \in \mathbb{R}$. Then*

$$\|\Sigma \Sigma_v^{-1}\| \leq G,$$

see [Gogolashvili et al. \(2023\)](#).

Moreover, if $w(x) < \infty$ and $v(x) > 0$ for all $x \in \mathcal{X}$, with w and v defined as in eq. (3) and (17) respectively, then

$$\|\Sigma \Sigma_v^{-1}\| \leq \left\| \frac{w}{v} \right\|_\infty.$$

Proof. For operators A and B on \mathcal{H} , denote by $A \geq B$ that $A - B$ is a non-negative operator. Let $G := \|d\rho_X^{te}/d\rho_X^v\|_\infty$. For all $f \in \mathcal{H}$, we have

$$\begin{aligned} \langle f, \Sigma f \rangle_{\mathcal{H}} &= \left\langle f, \int K_x f(x) d\rho_X^{te}(x) \right\rangle_{\mathcal{H}} = \int \langle f, K_x \rangle_{\mathcal{H}} f(x) d\rho_X^{te}(x) = \int f^2(x) d\rho_X^{te}(x) \\ &= \int f^2(x) \frac{d\rho_X^{te}}{d\rho_X^v}(x) d\rho_X^v(x) \leq G \int f^2(x) d\rho_X^v(x) = G \langle f, \Sigma_v f \rangle_{\mathcal{H}}. \end{aligned}$$

Therefore, we have

$$\Sigma \preceq G \Sigma_v \preceq G (\Sigma_v + \lambda \mathbb{I}) \implies \Sigma (\Sigma_v + \lambda \mathbb{I})^{-1} \preceq G \mathbb{I},$$

where $\mathbb{I} : \mathcal{H} \mapsto \mathcal{H}$ is the identity operator. This implies $\|\Sigma (\Sigma_v + \lambda \mathbb{I})^{-1}\| \leq G$, which proves the first assertion. The second part comes simply from the fact that

$$\langle f, \Sigma f \rangle_{\mathcal{H}} = \int f^2(x) \frac{d\rho_X^{te}}{d\rho_X^v}(x) \frac{d\rho_X^{tr}}{d\rho_X^v}(x) d\rho_X^v(x) = \int f^2(x) w(x) \frac{1}{v(x)} d\rho_X^v(x) \leq \left\| \frac{w}{v} \right\|_\infty \langle f, \Sigma_v f \rangle_{\mathcal{H}}.$$

□

Proposition 6 (Cordes Inequality [Fujii et al. \(1993\)](#)). *Let A, B two positive semidefinite bounded linear operators on a separable Hilbert space. Then*

$$\|A^s B^s\| \leq \|AB\|^s \quad \text{when } 0 \leq s \leq 1.$$

E Experimental Details and Datasets

This section includes some additional details on the experimental setting and a detailed summary of all the datasets presented.

The four datasets consist of data collected from multiple users using wearable sensors, such as accelerometers and gyroscopes. We allocate 1% of the test dataset for weight estimation, reserving the remaining 99% for predictions. Weights are estimated using Relative Unconstrained Least-Squares Importance Fitting (RuLSIF) method, which minimizes the relative density-ratio divergence between distributions (Yamada et al., 2013; Liu et al., 2013). The experiments were conducted in Python on a 2018 MacBook Pro with a 2.3 GHz Intel Core i5 Quad-Core processor, 16GB of RAM, and no GPU.

We give here a brief description of the four datasets.

- The HHAR dataset (Stisen et al., 2015), accessible on DR-NTU (Data), dataset DOI: <https://doi.org/10.21979/N9/OWDFXO>, version used: 3.0 (May 27, 2022), license: CC BY-NC 4.0, authors of the dataset version: Mohamed Ragab and Emadeldeen Eldele. It comprises 13,062,475 samples with 10 features for human activity recognition using smartphone and smartwatch sensors. It explores the effect of sensor heterogeneity on activity recognition algorithms, aiming to classify physical activities. Data were collected from accelerometer and gyroscope sensors during activities like Biking, Sitting, Standing, Walking, Stair Up, and Stair Down. Nine users participated, using 4 smartwatches (2 LG, 2 Samsung Galaxy) and 8 smartphones (2 each of Samsung Galaxy S3 mini, Samsung Galaxy S3, LG Nexus 4, and Samsung Galaxy S+). Features include accelerometer readings (x, y, z), user ID, device, model, and activity labels. Non-essential columns—index, arrival time, and creation time—were removed. The categorical columns (device and model) were converted to numerical values. User A has 1,218,871 samples designated as training data, from which 15,500 samples were randomly selected, while user H provides the test data.
- The HAR70+ dataset (Ustad et al., 2023), available from the UCI Machine Learning Repository, dataset DOI: <https://doi.org/10.24432/C5CW3D>, version used: Latest version available via the Norwegian Centre for Research Data (NSD), license: CC BY 4.0 (Creative Commons Attribution 4.0 International), dataset authors: Aleksej Logacjov, Astrid Ustad. It contains 2,259,597 samples with 6 features for activity classification. Data were recorded from 18 older adults (ages 70–95) wearing two Axivity AX3 accelerometers at 50 Hz for 40 minutes in a semi-structured, free-living setting. Five participants used walking aids. Sensors were placed on the right thigh and lower back. Activities such as walking, shuffling, ascending/descending stairs, standing, sitting, and lying were annotated frame-by-frame using chest-mounted camera video. For training and testing, we used data from two files: 518.csv, containing 141,714 samples with 6 features, and 516.csv, containing 138,278 samples with 6 features. From 518.csv, 20,000 samples were randomly selected for training. To introduce covariate shift and study feature selection, the timestamp and thigh accelerations in the x and z directions were excluded.
- The HARChildren dataset (Tørring et al., 2024), available on DR-NTU (Data), dataset DOI: <https://doi.org/10.18710/EPCXCC>, version used: published on August 30, 2024, license: CC0 1.0, dataset authors: Marte Fosslaten Tørring et al. It contains over 5 million samples with 8 features for activity classification. It includes data from 63 typically developing (TD) children and 16 children with Cerebral Palsy (CP), classified at levels I and II of the Gross Motor Function Classification System (GMFCS). These children wore two accelerometers, on the lower back and thigh. The features initially included a timestamp and accelerometer readings (x, y, z), but the timestamp was removed during preprocessing. The recorded activities include walking, running, shuffling, ascending and descending stairs, standing, sitting, lying, bending, cycling seated, cycling standing, and jumping. For training and testing, two files were used: 004.csv, which contains 354,139 samples, and 010.csv, which contains 238,574 samples. From 010.csv, 15,000 samples were randomly selected for training, while 004.csv was used for testing.
- The WISDM dataset (Kwapisz et al., 2011), available on DR-NTU (Data), dataset DOI: <https://doi.org/10.21979/N9/KJWE5B>, version used: 1.0 (May 27, 2022), license: CC BY-NC 4.0, dataset authors: Mohamed Ragab and Emadeldeen Eldele. It contains 1,098,209

samples with 5 features from accelerometer and gyroscope data collected at 20 Hz using smartphones and smartwatches. Data were recorded from 36 users performing 18 activities for 3 minutes each. Features include a user ID, activity code (label), and sensor readings (x, y, z). User 12 has 32,641 samples designated as training data, from which 25,000 samples were randomly selected, while user 19 provides the test data.

A common way to specify the grid of possible values of λ is to consider a geometric series.

Remark 1 (Geometric grid). *Let λ_{min} and λ_{max} the smallest and largest values of the regularization parameter we wish to consider and let*

$$b = \left(\frac{\lambda_{max}}{\lambda_{min}} \right)^{1/(Q-1)}.$$

We generate a geometric grid of Q values of the regularization if for $q = 1, \dots, Q$, we let

$$\lambda_q = b^{q-1} \lambda_{min},$$

so that $\lambda_1 = \lambda_{min}$ and $\lambda_Q = \lambda_{max}$.

We used $\lambda_1 = 10^{-4}$, $\lambda_Q = 1$ and $Q = 10$ to generate the geometric grid. We also propose a method for generating a geometric sequence of γ values, derived from distinct formulas for even and odd indices.

Remark 2. *The following values represent the parameter γ for the Gaussian kernel:*

$$\gamma_k = \begin{cases} 10^{-3+\frac{k-1}{2}}, & k \text{ odd}, \\ 5 \cdot 10^{-3+\frac{k-2}{2}}, & k \text{ even}, \end{cases} \quad k = 1, 2, \dots, 6.$$

We optimize the hyperparameters using hold-out cross-validation, partitioning the dataset into 70% for training and 30% for validation. For each combination of hyperparameters λ and γ , we train the model on X_{train} and y_{train} and validate it on X_{val} and y_{val} .

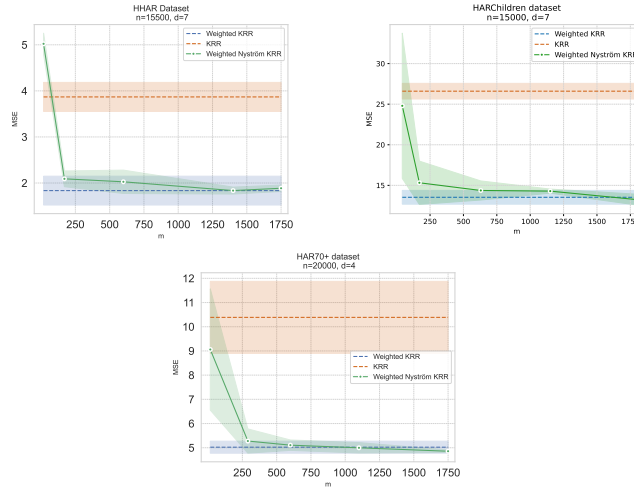


Figure 2: The graphs above illustrate the results summarized in Table 1 for HHAR, HAR70+ and HARChildren datasets.

As regards the simulations, we report here the parameters used to generate the results in Figure 1: $k = 50$ (that we can consider misspecified), $\mu_{tr} = (0.7, 0.7)$, $\sigma_{tr}^2 = \text{diag}(0.7, 0.7)$, $\mu_{te} = (1.8, 1.8)$, $\sigma_{te}^2 = \text{diag}(0.5, 0.5)$, $\epsilon^2 = 0.2$, $c_1 = c_2 = 10$.

Table 2 shows that even when using uniform sampling, the two methods with importance weighting (IW) correction achieve the best and essentially equal performance. However, the number of Nyström points m required by W-Nyström KRR is 5500, 6500, 1750, and 4500 for HAR70+, HARChildren, HHAR, and WISDM respectively, which represents an increase compared to the ALS method.

Table 2: Performances of the various methods, both in terms of MSE and training/prediction time using uniform sampling.

	HAR70+ ($n = 20000$)			HARChildren ($n = 15000$)			HHAR ($n = 15500$)			WISDM ($n = 25000$)		
	MSE	t train (s)	t pred (s)	MSE	t train (s)	t pred (s)	MSE	t train (s)	t pred (s)	MSE	t train (s)	t pred (s)
KRR	10 ± 1	1694 ± 2	15.0 ± 0.5	26.6 ± 0.9	762 ± 12	10.2 ± 0.4	3.7 ± 0.3	876 ± 6	10.5 ± 0.9	7.8 ± 0.1	3280 ± 48	38 ± 5
W-KRR	5.0 ± 0.2	1785 ± 2	15.1 ± 0.3	13.5 ± 0.8	809 ± 26	9.0 ± 0.1	1.8 ± 0.3	1034 ± 93	9.9 ± 0.1	4.8 ± 0.2	3364 ± 30	33 ± 2
W-Ny KRR	5.0 ± 0.1	67 ± 0.8	3.8 ± 0.1	12.9 ± 0.2	57 ± 0.2	6.2 ± 1.0	1.79 ± 0.01	5.9 ± 0.1	1.3 ± 0.2	4.8 ± 0.1	32 ± 0.5	7.2 ± 0.1