LEARNING MULTI-INDEX MODELS WITH HYPER-KERNEL RIDGE REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks excel in high-dimensional problems, outperforming models such as kernel methods, which suffer from the curse of dimensionality. However, the theoretical foundations of this success remain poorly understood. We follow the idea that the compositional structure of the learning task is the key factor determining when deep networks outperform other approaches. Taking a step towards formalizing this idea, we consider a simple compositional model, namely the multi-index model (MIM). In this context, we introduce and study hyper-kernel ridge regression (HKRR), an approach blending neural networks and kernel methods. Our main contribution is a sample complexity result demonstrating that HKRR can adaptively learn MIM, overcoming the curse of dimensionality. Further, we exploit the kernel nature of the estimator to develop ad hoc optimization approaches. Indeed, we contrast alternating minimization and alternating gradient methods both theoretically and numerically. These numerical results complement and reinforce our theoretical findings.

1 INTRODUCTION

The search for principles underlying the success of deep networks in learning from high-dimensional problems has been the subject of much interest. At least two ideas have a long history. The first is **invariance**. Deep architectures emerge from the need to derive models insensitive to transformations that are uninformative for the task at hand. Computational primitives such as filtering and pooling at different scales can be understood as implementing these ideas. This perspective traces back to early work in computer vision (Fukushima, 1980; LeCun et al., 1989), itself motivated by ideas in neuroscience (Hubel & Wiesel, 1962; Riesenhuber & Poggio, 1999), and we refer to Serre et al. (2007); Mallat (2012) for examples of more recent contributions in this line of work. A second idea is **compositionality**. High-dimensional data often have a hierarchical structure where parts at different scales interact. Language provides a natural example, with its structure in letters, syllables, words, and sentences. Deep architectures can then be designed to exploit this structure. These ideas, which go back at least to Bienenstock & Geman (1998); Yuille & Kersten (2006), provide another perspective on algorithmic developments such as convolutions (LeCun et al., 2002) and attention mechanisms (Vaswani et al., 2017). Ultimately, the relevance of either one of these principles relies on their ability to reduce the need for data, thus translating into more successful learning schemes. The study of sample complexity in statistical learning theory provides a framework within which this intuition can be formalized and tested (Vapnik, 2013).

Classic sample complexity results highlight the role of data dimension and function smoothness. In the absence of any assumption, no sample complexity results can be derived (Vapnik, 2013; Devroye et al., 2013). Assuming the task of interest is described by Lipschitz functions leads to sample complexity scaling exponentially with the dimension of the input data—the so-called curse of dimensionality (Donoho et al., 2000). Such dependence can be alleviated if further smoothness is assumed, yielding sample complexity that depends exponentially on the ratio of dimension to smoothness (Stone, 1982). Both these classes of problems (Lipschitz and smooth Sobolev functions) can be learned by a variety of learning approaches, including kernel methods and neural networks, hence not explaining the better performance of the latter on high-dimensional problems. Starting from the seminal work in Barron (2002), this observation has led to investigating how to characterize the class of problems where deep networks excel; see, e.g., Poggio et al. (2017) for a recent account.

Circling back to the initial discussion, the role of invariance in sample complexity has been discussed in Poggio et al. (2017) and analyzed, for example, in Mei et al. (2021) in the context of group transformations. However, invariance alone seems insufficient to account for the striking empirical performance observed in practice. A functional viewpoint on compositionality was proposed in Mhaskar et al. (2017) and further developed in Dahmen

(2025) from the perspective of approximation theory. Sample complexity bounds were derived in Schmidt-Hieber (2020); Kohler & Langer (2021), laying the groundwork for a theoretical understanding of compositional structure. It is within this line of work that our contribution is situated. Our study is further motivated by the work of Radhakrishnan et al. (2022), which points to a simpler compositional structure and proposes a kernel-based approach to learn it, called Recursive Feature Machine (RFM), drawing on ideas from sufficient dimensionality reduction (Fukumizu et al., 2009). As we discuss next, we propose an alternative approach within the same context.

The approach we study blends ideas from kernel methods and neural networks. It draws inspiration from Poggio & Girosi (1990), where an extension of radial basis function networks (RBF), called hyper-RBF, was proposed. Instead of a single kernel and its RKHS, we consider a family of kernels and their corresponding RKHSs. Each kernel is obtained by composing a fixed common kernel with a linear transformation that maps inputs to a lower-dimensional space. A solution is then obtained through regularized empirical risk minimization with least squares. For any fixed transformation, the approach reduces to kernel ridge regression (KRR). But now, rather than being fixed, the best transformation is learned during training. The resulting method is called hyper-kernel ridge regression (HKRR), and reduces to hyper-RBF when radial kernels are used. It can be seen as a special form of neural network or as a kernel method augmented with a built-in linear representation learning step. In particular, classic approximation schemes for kernel methods, such as Nyström approximations (Rudi et al., 2015), can be exploited.

HKRR provides a natural framework for learning multi-index models (MIMs), given their structure as the composition of a linear transformation and a smooth nonlinear function. Our main contribution is the characterization of the sample complexity of HKRR for learning MIMs. We show that in this case, the dependence is exponential in the ratio between the true transformation dimension and the smoothness, and only polynomial in the input data dimension. We further show that the transformation dimension does not need to be known a priori but can be tuned by hold-out cross-validation, preserving the same sample complexity up to logarithmic factors. We complete our statistical analysis by showing that the HKRR estimator can be compressed using Nyström subsampling (Rudi et al., 2015), without degrading the sample complexity. The proofs largely draw on techniques developed for kernel methods, extended to handle the compositional nature of hyper-kernels. A second contribution is to investigate the solution of the HKRR optimization problem both theoretically and numerically. In particular, we contrast two different approaches. The first leverages the connection to KRR and alternates closed-form updates for the estimator (given a transformation) with transformation updates via gradient descent (VarPro), in the spirit of variable projection methods (Golub & Pereyra, 1973). The second strategy alternates gradient descent (AGD) steps, akin to the PALM algorithm in Bolte et al. (2014). The HKRR optimization problem is non-convex, but both strategies can be shown to converge to a critical point. Numerically, however, the AGD approach appears more stable and ultimately outperforms VarPro. We attribute this behavior to the nonlocal nature of the latter, as we illustrate numerically. Overall, our results show that HKRR can be viewed as a useful augmentation of kernel methods, while providing a sound algorithmic approach to study simple compositionality and representation learning models.

Some notation and background are given in Section 2. Section 3 introduces the HKRR problem and two algorithms, VarPro and AGD. Section 4 presents the sample complexity of HKRR and the convergence analysis of both algorithms. Experimental results and conclusions are reported in Section 5 and Section 6.

2 BACKGROUND

In this section, we collect some basic definitions and notation.

Statistical learning and sample complexity. Let ρ be a joint probability distribution on $X \times Y \subset \mathbb{R}^D \times \mathbb{R}$. The learning problem with the square loss consists in minimizing, over all measurable functions, the expected risk

$$\mathcal{R}(f) = \mathbb{E}[(f(x) - y)^2]$$

given $(x_i, y_i)_{i=1}^m \overset{\text{i.i.d.}}{\sim} \rho^m$. The quality of a learning solution $\hat{f}$ is measured by the excess risk $\mathcal{R}(\hat{f}) - \mathcal{R}(f_*)$, where f_* denotes a risk minimizer. For the square loss, a minimizer is the so-called regression function defined as $f_*(x) = \mathbb{E}[y \mid x]$ almost surely. The sample complexity of a learning algorithm is the number of samples required by a corresponding empirical solution to achieve a prescribed accuracy with a prescribed confidence level. More precisely, given $\epsilon > 0$ and $\delta \in [0, 1]$, we say that a procedure outputting solutions $\hat{f}$ given m points has sample complexity $m(\epsilon, \delta)$, if for all $m \geq m(\epsilon, \delta)$, $\mathcal{R}(\hat{f}) - \mathcal{R}(f_*) \leq \epsilon$ with probability at least $1 - \delta$. Here, ϵ and $1 - \delta$ are the accuracy level and the confidence level, respectively. The function $m(\epsilon, \delta)$ can typically be inverted to express the results in terms of error bounds. Given m and δ, an error bound is a function $\epsilon(m, \delta)$ such that $\mathcal{R}(\hat{f}) - \mathcal{R}(f_*) \leq \epsilon(m, \delta)$, with probability at least $1 - \delta$. In the following, we will take this latter point of view and review classical algorithms relevant to our study.

ERM and kernel methods. Fixed a hypothesis space $\mathcal{H}$ of measurable functions $f : X \to \mathbb{R}$, the empirical risk minimization (ERM) over $\mathcal{H}$ is given by $\hat{f} = \arg\min_{f\in\mathcal{H}} \widehat{\mathcal{R}}(f) = \arg\min_{f\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\big(f(x_i) - y_i\big)^2$. In particular, kernel methods consider $\mathcal{H}$ to be an RKHS, that is a Hilbert space of functions with a reproducing kernel $k : X \times X \to \mathbb{R}$ satisfying $k_x = k(\cdot, x) \in \mathcal{H}$, and $f(x) = \langle f, k_x\rangle_{\mathcal{H}}$, for all $x \in X$ and $f \in \mathcal{H}$ (Aronszajn, 1950). Further, KRR corresponds to minimizing the regularized empirical risk

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

By the representer theorem (Schölkopf et al., 2001), $\hat{f}_\lambda = \sum_{i=1}^{m}\alpha_i\, k(x_i, \cdot)$, so that KRR reduces to a finite-dimensional problem

$$\alpha^* := \arg\min_{\alpha\in\mathbb{R}^m} \frac{1}{m}\|\hat{K}\alpha - \mathbf{y}\|^2 + \lambda\alpha^T\hat{K}\alpha, \tag{2}$$

where $\mathbf{y} = (y_i)_{i=1}^m \in \mathbb{R}^m$ and $\hat{K} \in \mathbb{R}^{m\times m}$ with $(\hat{K})_{i,j} = k(x_i, x_j)$ is the empirical kernel matrix. More efficient computations are possible using Nyström approximation, considering $\tilde{m} < m$ inducing points $(\tilde{x}_i)_{i=1}^{\tilde{m}} \subset (x_i)_{i=1}^m$ and a subspace of functions of the form $f(\cdot) = \sum_{i=1}^{\tilde{m}}\tilde{\alpha}_i\, k(\tilde{x}_i, \cdot)$. The Nyström KRR is then given by

$$\tilde{\alpha}^* := \arg\min_{\tilde{\alpha}\in\mathbb{R}^{\tilde{m}}} \frac{1}{m}\|\hat{K}_{m\tilde{m}}\tilde{\alpha} - \mathbf{y}\|^2 + \lambda\tilde{\alpha}^T\hat{K}_{\tilde{m}\tilde{m}}\tilde{\alpha}, \tag{3}$$

where $(\hat{K}_{m\tilde{m}})_{i,j} = k(x_i, \tilde{x}_j)$ and $(\hat{K}_{\tilde{m}\tilde{m}})_{i,j} = k(\tilde{x}_i, \tilde{x}_j)$.

It is useful to contrast kernel methods with classic one-hidden-layer neural networks.

Remark 1 (Neural and RBF networks)**.** *One-hidden-layer neural networks consider functions of the form $f(x) = \sum_{j=1}^{u} c_i\sigma(w_j^\top x + b_j)$, where $\sigma : \mathbb{R} \to \mathbb{R}$ is a nonlinearity, e.g. the ReLU $\sigma(z) = \max\{0, z\}$, and $c_j, b_j \in \mathbb{R}, w_j \in \mathbb{R}^D$, $j = 1, \dots, u$ are parameters to be determined. Each term $\sigma(w_j^\top x + b_j)$ is called a neuron, u is the number of neurons/units, and $(w_j, b_j)_j$ are called hidden weights. Radial basis function (RBF) networks consider functions of the form $f(x) = \sum_{j=1}^{u} c_i\phi(\|w_j - x\|)$, where $\phi : \mathbb{R} \to \mathbb{R}$ is a nonlinearity, e.g. the Gaussian $\phi(z) = e^{-z}$, and again $c_j \in \mathbb{R}, w_j \in \mathbb{R}^D$, $j = 1, \dots, u$ are parameters to be determined.*

3 Hyper-kernel ridge regression

In this section, we describe HKRR, an approach blending ideas from kernel methods and neural networks. HKRR is based on regularized ERM like KRR, but considers a class of functions defined by a family of parameterized kernels, rather than one fixed kernel. Similar to neural networks, the solution is a linear combination of nonlinearities with parameters to be determined during training. The corresponding optimization problem is nonconvex, but its structure suggests ad-hoc gradient approaches.

3.1 Hyper-kernel ridge regression

Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a fixed "mother" reproducing kernel with associated RKHS $\mathcal{H}_k$. Define $\mathcal{B} = \{B \in \mathbb{R}^{d\times D} : \|B\|_\infty \le 1\}$, where $d < D$ and $\|B\|_\infty := \sup_{\|x\|\le 1}\|Bx\|$. A hyper-kernel $k_B : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is defined by the composition of the kernel k with a linear map $B \in \mathcal{B}$, namely,

$$k_B(x, x') = k\big(Bx, Bx'\big), \quad x, x' \in \mathbb{R}^D.$$

The RKHS with reproducing kernel k_B is denoted by $\mathcal{H}_B$ for each $B \in \mathcal{B}$. Note that our definition of the hyper-RKHS differs from that in Liu et al. (2021). The intuition is that each map B provides a low-dimensional linear representation of the data, while the mother kernel k defines a space of nonlinear functions on this reduced space. Considering hyper-kernels allows us to learn an estimator that composes an optimal linear representation and a corresponding nonlinear function. HKRR achieves this by solving the regularized ERM problem

$$\min_{B\in\mathcal{B}}\min_{f\in\mathcal{H}_B} \widehat{\mathcal{R}}_\lambda(f), \qquad \widehat{\mathcal{R}}_\lambda(f) = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 + \lambda\|f\|_{\mathcal{H}_B}^2. \tag{4}$$

For any given $B \in \mathcal{B}$, the inner optimization over $\mathcal{H}_B$ is a standard KRR problem (equation 1) with kernel k_B. As discussed in Section 2, the problem is strongly convex and admits a unique minimizer $\hat{f}_\lambda^B = \arg\min_{f\in\mathcal{H}_B} \widehat{\mathcal{R}}_\lambda(f)$,

which can be computed using the representer theorem (equation 2), and more efficiently via a Nyström approximation (equation 3). The outer optimization problem over $\mathcal{B}$ is non-convex and corresponds to

$$\operatorname*{arg\,min}_{B\in\mathcal{B}} \hat{H}_\lambda(B), \qquad \hat{H}_\lambda(B) := \min_{f\in\mathcal{H}_B} \widehat{\mathcal{R}}_\lambda(f) = \frac{1}{m}\sum_{i=1}^{m}\left(\hat{f}_\lambda^B(x_i) - y_i\right)^2 + \lambda\|\hat{f}_\lambda^B\|_{\mathcal{H}_B}^2.$$

If $\hat{B}_d$ is a solution of the above problem, then the HKRR estimator is $\hat{f}_\lambda^{\hat{B}_d}$ and relies on the choice of the mother kernel k, the regularization parameter λ, and the dimension d of the linear maps in $\mathcal{B}$. In Section 5, we will investigate how these choices influence the corresponding learning performances. We first discuss the practical computation of the HKRR estimator.

3.2 Computing an HKRR solution

As already mentioned, the representer theorem (2) allows us to reduce problem (4) to a finite-dimensional optimization. In practice, we adopt the Nyström approximation as equation 3, where $(\tilde{x}_i)_{i=1}^{\tilde{m}}$ are sampled uniformly without replacement from the training set. This procedure is referred to as the plain Nyström method (Rudi et al., 2015), and leads to

$$\min_{B\in\mathcal{B}}\min_{\alpha\in\mathbb{R}^{\tilde{m}}} \hat{\mathcal{L}}(B,\alpha), \quad \hat{\mathcal{L}}(B,\alpha) = \frac{1}{m}\left\|\hat{K}_{m\tilde{m}}^B\alpha - \mathbf{y}\right\|^2 + \lambda\,\alpha^\top \hat{K}_{\tilde{m}\tilde{m}}^B\alpha, \tag{5}$$

where $\hat{K}_{m\tilde{m}}^B \in \mathbb{R}^{m\times\tilde{m}}$ and $\hat{K}_{\tilde{m}\tilde{m}}^B \in \mathbb{R}^{\tilde{m}\times\tilde{m}}$ are defined by $(\hat{K}_{m\tilde{m}}^B)_{i,j} = k(Bx_i, B\tilde{x}_j), (\hat{K}_{\tilde{m}\tilde{m}}^B)_{i,j} = k(B\tilde{x}_i, B\tilde{x}_j)$. Let $\hat{f}_\lambda^{\hat{B}_{d,\tilde{m}}}$ denote the solution of equation 5, with the index $\tilde{m}$ highlighting the use of the Nyström approximation.

We next discuss some aspects of HKRR optimization and refer to Appendix C, and in particular to Lemma 8, for details. We begin by noting that, for each $B \in \mathcal{B}$, the inner minimization admits an explicit solution

$$\alpha(B) = \left((\hat{K}_{m\tilde{m}}^B)^\top \hat{K}_{m\tilde{m}}^B + \lambda m\,\hat{K}_{\tilde{m}\tilde{m}}^B\right)^{-1}(\hat{K}_{m\tilde{m}}^B)^\top\mathbf{y}. \tag{6}$$

Plugging this expression into equation 3, we have for each $B \in \mathcal{B}$ that

$$\hat{H}_\lambda(B) = \min_{\alpha\in\mathbb{R}^{\tilde{m}}} \hat{\mathcal{L}}(B,\alpha) = \frac{1}{m}\mathbf{y}^\top\mathbf{y} - \frac{1}{m}\mathbf{y}^\top\hat{K}_{m\tilde{m}}^B\left((\hat{K}_{m\tilde{m}}^B)^\top\hat{K}_{m\tilde{m}}^B + \lambda m\,\hat{K}_{\tilde{m}\tilde{m}}^B\right)^{-1}(\hat{K}_{m\tilde{m}}^B)^\top\mathbf{y}. \tag{7}$$

If the kernel is smooth, then $\hat{H}_\lambda$ is differentiable, and so is $\hat{\mathcal{L}}(B,\alpha)$ for any $\alpha \in \mathbb{R}^{\tilde{m}}$, and hence $\hat{\mathcal{L}}$ itself. However, $\hat{H}_\lambda$ is not convex, and neither is $\hat{\mathcal{L}}$. We will see that, if the mother kernel k is analytic, then $\hat{H}_\lambda$ satisfies the Kurdyka–Łojasiewicz property Attouch et al. (2013), which will allow the derivation of some optimization guarantees; see Section 4.4 (Theorem 4).

Given the above discussion, we next propose two methods to compute an (Nyström) HKRR solution. The first method is **Variable Projection (VarPro)**, see Algorithm 1. It exploits the closed-form solution to update α (see equation 6), while applying gradient descent steps on B to minimize $\hat{H}(B)$, see equation 7. This approach is well known in the optimization literature (Golub & Pereyra, 1973; 2003), and allows the use of other optimization schemes such as L-BFGS (Poon & Peyré, 2023). We note that this idea has also been adapted in a related, though slightly different, setting in Follain & Bach (2024), introducing BKerNN. The second method is **Alternating Gradient Descent (AGD)**, see Algorithm 2. It is based only on gradient information and successively updates B and α through gradient descent steps. This algorithm is similar to PALM (for Proximal Alternating Linearized Minimization) introduced in Bolte et al. (2014), allowing multiple steps in α to improve its performance. Such alternating gradient schemes have already been applied in the literature in nonconvex settings, for example, for matrix factorization or two-layer neural networks (Lu et al., 2019; Ward & Kolda, 2023). In Appendix C, we provide further details on the above methods, including line search strategies for automatically tuning the learning rates s_α and s_B, and how to handle the constraint on matrix B. Some convergence results are provided in Theorem 4, while empirical performances are investigated in Section 5. We end this section discussing some comparison with other works in the literature.

3.3 Related approaches

In this section, we discuss the connection to some approaches that directly influence our study. An inspiration for the HKRR approach is Hyper-RBF networks proposed in Poggio & Girosi (1990), from which the term

Algorithm 1: VarPro (informal)

Require: B^0, $s_B > 0$
1: $\alpha^0 = \arg\min_{\alpha \in \mathbb{R}^{\tilde{m}}} \hat{\mathcal{L}}(B^0, \alpha)$
2: **for** $i = 0, 1, \ldots$ **do**
3: $\quad B^{i+1} = B^i - s_B \nabla_B \hat{\mathcal{L}}(B^i, \alpha^i)$
4: $\quad \alpha^{i+1} = \arg\min_{\alpha \in \mathbb{R}^{\tilde{m}}} \hat{\mathcal{L}}(B^{i+1}, \alpha)$
5: **end for**
6: **return** (B^{i+1}, α^{i+1})

Algorithm 2: AGD (informal)

Require: B^0, α^0, $s_\alpha > 0$, $s_B > 0$, $n_\alpha \in \mathbb{N}^*$
1: **for** $i = 0, 1, \ldots$ **do**
2: $\quad B^{i+1} = B^i - s_B \nabla_B \hat{\mathcal{L}}(B^i, \alpha^i)$
3: $\quad \alpha^{i,0} = \alpha^i$
4: $\quad$ **for** $j = 0, 1, \ldots, n_\alpha - 1$ **do**
5: $\quad\quad \alpha^{i,j+1} = \alpha^{i,j} - s_\alpha \nabla_\alpha \hat{\mathcal{L}}(B^{i+1}, \alpha^{i,j})$
6: $\quad$ **end for**
7: $\quad \alpha^{i+1} = \alpha^{i,n_\alpha}$
8: **end for**
9: **return** (B^{i+1}, α^{i+1})

"hyper" is borrowed. Hyper-RBF networks extend standard RBF networks, considering functions of the form $f(x) = \sum_{i=1}^m \phi\big(\|B(x - w_i)\|\big)$, with B a linear transformation to be learned. In Poggio & Girosi (1990), neither the representer theorem nor Nyström inducing points were considered, and the *centers* $w_1, \ldots, w_m$, together with the coefficients $\alpha_1, \ldots, \alpha_m$ and the matrix B, were optimized using stochastic gradient with no optimization guarantees. In comparison, we consider a more general class of hyper-kernels; we do not optimize the centers, but use Nyström inducing points; and finally, we consider different gradient-based methods for which convergence guarantees are provided. Another inspiration for our work is the recursive feature machine (RFM) proposed in Radhakrishnan et al. (2024b;a); see also Zhu et al. (2025). RFM is based on hyper-RBF kernels and defines an estimator similar, though with a slightly different form. Indeed, noting that $\|Bx\| = \big(x^\top M x\big)^{1/2}$, with $M = B^\top B$, RFM considers functions of the form $f(x) = \sum_{i=1}^m k\big((x - x_i)^\top M (x - x_i)\big)$, where k is a radial basis function that is also a reproducing kernel. The centers $x_1, \ldots, x_m$ are taken to be the input data points, as in kernel methods and HKRR. The coefficients are computed for an initial B (or rather M) via KRR using a closed-form expression. The key feature of RFM lies in the computation of M, which is given by the average gradient outer product (AGOP) operator (Xia et al., 2002): $M = \frac{1}{n} \sum_{i=1}^n \nabla f(x_i) \nabla f(x_i)^\top$, with f a KRR solution. KRR and AGOP computations are then alternated. Aside from the more specific nature of the hyper-kernels considered, RFM is close to our VarPro algorithm. The gradient step update of B in VarPro is replaced by the AGOP update. The AGOP operator has a long history in statistics in the context of sufficient dimension reduction (Samarov, 1993; Hristache et al., 2001). However, unlike VarPro, the RFM iteration does not currently have an ERM and hence an optimization interpretation. Finally, HKRR was also considered in Chen et al. (2023), developing ideas of Fukumizu et al. (2009). We will discuss this more in the next section.

4 Theoretical results of HKRR for learning MIMs

In this section, we present a bound on the excess risk of HKRR for learning MIMs (Theorem 1), derive the convergence rate of the Nyström approximation for HKRR in Theorem 2, and provide a theoretical analysis of adaptively estimating the unknown latent dimension d_* and the regularization parameter λ via cross-validation (Theorem 3). The convergence analysis of AGD and VarPro is also established in Theorem 4.

4.1 Learning MIMs with HKRR: excess risk bound

Consider MIMs, where the regression function takes the form

$$f_*(x) = g_*(B_* x), \qquad \rho_X\text{-a.e. } x \in X, \tag{8}$$

with B_* a $d_* \times D$ matrix such that $d_* < D$, $\|B_*\|_\infty \leq 1$, and g_* a measurable function defined on $\mathbb{R}^{d_*}$. Estimating MIMs is challenging due to both the nonlinearity of the function g_* and the difficulty of determining the linear map B_*. The following assumptions are needed to derive the rate of excess risk.

Assumption 1. *We assume that:*

1.1 (Bounded data). The input space X is a closed subset of $\mathbb{R}^D$, with $\|x\| \leq 1$ and $|y| \leq M$ for some $M > 0$.

1.2 (Smoothness). For some integer $r \geq 1$, the mother kernel satisfies $k \in C^r(\mathbb{R}^{d_*} \times \mathbb{R}^{d_*})$.

1.3 (Source condition). For some $d_* < D$, there exists $B_* \in \mathbb{R}^{d_* \times D}$ with $\|B_*\|_\infty \leq 1$, such that f_* lies in $\mathrm{Range}\big(L_{k_{B_*}}^{\theta/2}\big)$ for some $\theta \in (0,1]$. Here, $L_{k_{B_*}} : L_2(X, \rho_X) \to L_2(X, \rho_X)$ is an integral operator given by $(L_{k_{B_*}} f)(x) = \int_X k_{B_*}(x, x')\, f(x')\, \mathrm{d}\rho_X(x')$.

The condition that the input space X is contained in the unit ball can always be enforced for bounded inputs by rescaling. The boundedness of the outputs is also a standard assumption. The setting we consider is in the field of classical distribution-free non-parametric learning (Györfi et al., 2002). This contrasts with the stricter distributional assumptions adopted in other works, see, e.g., Mousavi-Hosseini et al. (2022); Bietti et al. (2025). The smoothness of the kernel k provides a sufficient condition to control the covering numbers (see Assumption 4). The Matérn kernel is an example satisfying this assumption (Williams & Rasmussen, 2006). The source condition is well studied in classical kernel methods (see, e.g., Cucker & Zhou (2007); De Vito et al. (2021)). It states the relationship between the target f_* and the space determined by the integral operator defined by the kernel k_{B_*}. The parameter θ controls the smoothness of f_*. A larger θ implies smoother functions and a smaller function space, and therefore a better approximation rate.

The following theorem establishes the excess risk rate of HKRR defined in equation 4. Here, the dimension d_* is assumed to be known a priori, while the adaptive result for unknown d_* is stated in Theorem 3. The proof is provided in Appendix A.7.

Theorem 1. *Suppose Assumption 1 holds. Let $0 < \delta < 2/e$, $\zeta < r/(d_* + r)$ and $\lambda = \lambda_m = m^{-\zeta}$. Then with probability at least $1 - \delta$, there holds*

$$\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq C_1 D d_* \log^2(2/\delta) m^{-\theta\zeta}$$

for all $m \geq m_\delta$, where m_δ is independent of D, d_* and C_1 is a constant independent of D, d_* and δ.

Remark 2. *Explicit expressions for m_δ (see equation 28 with $s^* = d_*/r$) and for C_1 (see equation 31) are given in the proofs. For sufficiently large m_δ, the factor Dd_* can be improved to $(Dd_*)^{1/(s^*+1)}$ (Remark 7). Moreover, Theorem 6 in Appendix A yields a weaker bound, valid for all $m \geq 1$, of order $m^{-r\theta/(1+\theta)(d_*+r)}$.*

Remark 3 (Beating the curse of dimensionality)**.** *The minimax excess risk for estimating an r-smooth function $f : \mathbb{R}^D \to \mathbb{R}$ from m samples scales as $m^{-2r/(2r+D)}$ (Györfi et al., 2002), which deteriorates exponentially with the input dimension D. In contrast, HKRR for MIM achieves a rate that depends exponentially only on the true transformation dimension $d_* \ll D$ and only polynomially on D, thereby mitigating the curse of dimensionality.*

Remark 4 (Suboptimal rate)**.** *Theorem 1 yields an excess risk bound of order $m^{-2r/(2r+2d_*)}$, which introduces an extra factor of 2 in the d_*-term compared with the conjectured optimal rate. This suboptimality likely arises from relying on L_∞-based covering number bounds over $\bigcup_B \mathcal{H}_B$. Sharper analysis based on L_2-norm estimates or local Rademacher complexity (Bartlett et al., 2005) is left for future work.*

4.2 Nyström approximation

Recall that the solution of the Nyström problem in equation 5 is denoted by $\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}$. To describe the relationship between $\tilde{m}$ and m, we define a random variable $\mathcal{N}_{B_*,x}(\lambda) = \langle k_{B_*x}, (\Sigma_{B_*} + \lambda I)^{-1} k_{B_*x} \rangle_{\mathcal{H}_{B_*}}$ for $\lambda > 0$ with $x \sim \rho_X$, where Σ_{B_*} is the covariance operator of k_{B_*} (see equation 9), and set $\mathcal{N}_{B_*,\infty}(\lambda) = \sup_{x \in X} \mathcal{N}_{B_*,x}(\lambda)$.

The following result shows that, with $\tilde{m} < m$ points, the plain Nyström estimator can achieve the same excess risk rate as in Theorem 1, up to constants. The proof is given in Appendix B.

Theorem 2. *Under the assumptions of Theorem 1, with probability at least $1 - \delta$,*

$$\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \mathcal{R}(f_*) \leq C_2 D d_* \log^2(2/\delta) m^{-\theta\zeta},$$

where C_2 is given in equation 34, provided $\tilde{m} \geq 67 \log\frac{4\kappa}{\lambda\delta} \vee 5\mathcal{N}_{B_*,\infty}(\lambda) \log\frac{4\kappa}{\lambda\delta}$ for $\kappa = \sup_x k(x,x)$.

Remark 5. *Since $\mathcal{N}_{B_*,\infty}(\lambda) \leq \kappa/\lambda$ for all $\lambda > 0$ (Caponnetto & De Vito, 2007; Rudi et al., 2015), under the assumptions of Theorem 1 we have $m > \tilde{m} \sim m^\zeta$, where ζ can be chosen arbitrarily close to $r/(d_* + r) < 1$.*

Remark 6. *We also provide the rate of the approximate leverage score (ALS) Nyström (Rudi et al., 2015) with varying numbers of subsampling points; see Appendix B for details. In fact, ALS requires fewer samples than the plain Nyström method since $\mathcal{N}_{B_*}(\lambda) \leq \mathcal{N}_{B_*,\infty}(\lambda)$, where $\mathcal{N}_{B_*}(\lambda) = \mathbb{E}_{x\sim\rho_X}[\mathcal{N}_{B_*,x}(\lambda)]$.*

4.3 ADAPTIVITY

The latent dimension d_* is unknown in practice. To obtain adaptive guarantees, d is tuned over $\{1, \dots, D\}$. Given $N \in \mathbb{N}, \lambda_1, \lambda_N > 0$ and $Q = (\lambda_N/\lambda_1)^{1/(N-1)}$, the regularization parameter λ is chosen from the geometric grid $\Lambda = \{\lambda_j = \lambda_1 Q^{j-1}\}_{j=1}^N$ assuming that the interval $[\lambda_1, \lambda_N]$ contains the optimal λ. Let $\Gamma = \big\{(d, \lambda) \mid d \in \{1, \dots, D\},\ \lambda \in \Lambda\big\}$, so that $|\Gamma| = DN$. Let $\{(x_i', y_i')\}_{i=1}^{m'} \sim \rho^{m'}$ be an independent validation set. We select

$$(\hat{d}, \hat{\lambda}) = \operatorname*{arg\,min}_{(d,\lambda)\in\Gamma} \frac{1}{m'} \sum_{i=1}^{m'} \Big(T_M \hat{f}_\lambda^{\hat{B}_d}(x_i') - y_i' \Big)^2.$$

Here, T_M is a truncation operator given by $T_M f(x) = \operatorname{sign}\big(f(x)\big)\ \min\big\{|f(x)|,\ M\big\}$, which handles the unboundedness of functions obtained by HKRR. The resulting estimator is denoted by $\hat{f}_{\hat{\lambda}}^{\hat{B}_{\hat{d}}}$. The next theorem states that it achieves the same rate (up to constants) as the estimator in Theorem 1. The idea of its proof is classical; see, e.g., Devroye et al. (2013); Chirinos-Rodríguez et al. (2024), and it is given explicitly in Appendix B.

Theorem 3. *For $\delta \in (0, 1)$ and a suitable $q \in [1, Q]$, the following holds with probability at least $1 - \delta$ that*

$$\mathcal{R}(T_M \hat{f}_{\hat{\lambda}}^{\hat{B}_{\hat{d}}}) - \mathcal{R}(f_*) \le 2q^\theta C_1 D d_* \log^2(2/\delta) m^{-\theta\zeta} + \frac{52M^2}{m'} \log \frac{2DN}{\delta}.$$

The above theorem shows how to choose hyperparameters adaptively and optimally. Furthermore, our experiments (Figure 2) highlight the impact of different choices of d and reveal an interesting phenomenon: overparameterizing d can sometimes yield better results. This observation suggests the conjecture that $\hat{d} > d_*$.

4.4 OPTIMIZATION GUARANTEES

We next study the convergence properties of Algorithms 1 and 2 introduced in Section 3.2; see Appendix C (Theorems 8 and 9) for further details. The proofs rely on the Kurdyka-Łojasiewicz property (Attouch et al., 2013), which in turn requires the kernel k to be analytic.

Theorem 4 (Convergence of AGD and VarPro (informal))**.** *Let k be an analytic kernel. Suppose that the sequences $\big(B^i\big)_{i\in\mathbb{N}}$ and $\big(\alpha^i\big)_{i\in\mathbb{N}}$ generated by Algorithm 1 or 2 are such that the minimal eigenvalue $\lambda_{min}\left(\hat{K}_{\tilde{m}\tilde{m}}^{B^i}\right) \ge \sigma$ for some $\sigma > 0$. Then, the sequence $\big(B^i, \alpha^i\big)_{i\in\mathbb{N}}$ converges to a critical point of $\Psi : B, \alpha \mapsto \mathcal{L}(B, \alpha) + i_{\mathcal{B}}(B)$ as i goes to infinity, and both sequences have finite length. In addition, there exists a constant $C > 0$ such that after N iterations, either $\big(B^N, \alpha^N\big)$ is a critical point of Ψ or*

$$\min_{0\le i\le N} \left\| \nabla_B \hat{\mathcal{L}}\left(B^i, \alpha^i\right) \right\|^2 \le \frac{C}{N}.$$

The above result ensures that both methods converge to some critical point of the objective function as long as the sequence $\big(B^i\big)_{i\in\mathbb{N}}$ does not shrink the minimal eigenvalue of the kernel matrix $\hat{K}_{\tilde{m}\tilde{m}}^{B^i}$. In other words, we require the data points to be linearly independent under the hyper-kernel k_{B^i} at each iteration. This assumption guarantees that the sequence $\big(\alpha^i\big)_{i\in\mathbb{N}}$ is well defined and bounded, which allows us to analyze the algorithms using Kurdyka-Łojasiewicz property.

4.5 COMPARISON WITH OTHER WORKS

Hyper-kernel RKHSs have been studied for dimension reduction, see e.g., Fukumizu et al. (2009); Fukumizu & Leng (2014); Chen et al. (2023). In Fukumizu et al. (2009), they studied the conditional cross-covariance operator between input and output RKHSs. It was shown that the operator equals to one induced by hyper-kernel input RKHS and output RKHS when B spans a central mean subspace (Chiaromonte & Cook, 2002). They further connected the operator to the expected risk and yield an ERM framework. Building on this idea, Chen et al. (2023) proposed a related HKRR method and proved that it can recover the true subspace dimension asymptotically.

The explicit excess risk rates for learning MIMs in different approaches are also studied in the literature. For example, Klock et al. (2021) used k-nearest neighbors and piecewise polynomials to learn the link function and employed the response-conditional least squares (RCLS) algorithm to estimate the latent matrix via inverse regression. Their generalization bound is $O(m^{-2r/(2r+d_*)})$, plus the error from learning the latent matrix. By contrast,

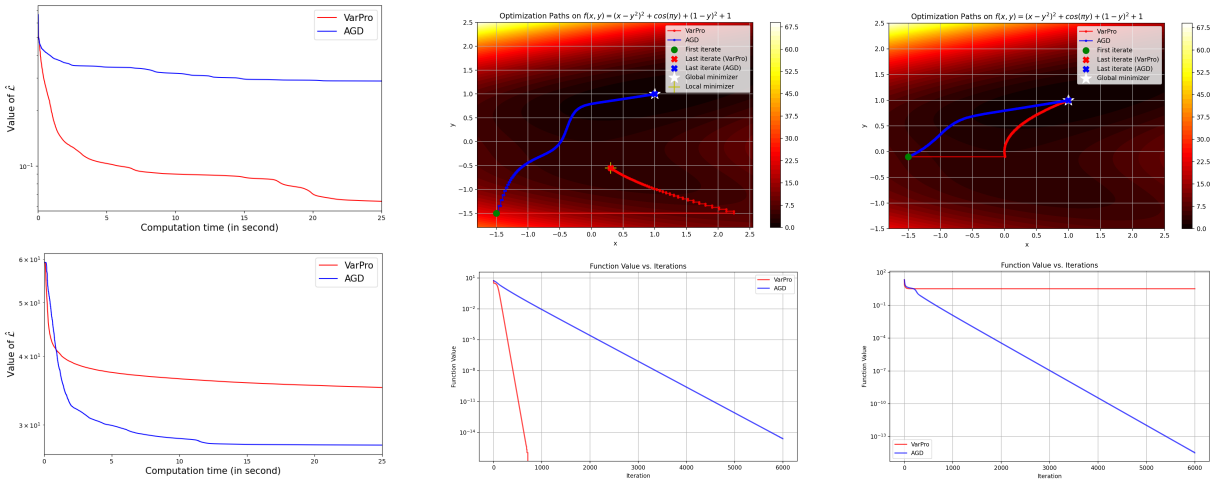
our approach achieves $O(m^{-r/(r+d_*)})$ and provides two alternating minimization algorithms with both theoretical guarantees and empirical validation. Bach (2017) established generalization bounds for MIMs with Lipschitz property. He considered hypothesis spaces as neural networks with finite variation norm and activation $\sigma(x) = (x)_+^\alpha$, $\alpha > 0$. For ReLU ($\alpha = 1$), the rate is $O((\log D)^{2/d_*+3} m^{-1/(d_*+3)} \log m)$. Our results emphasize the blessing of smoothness: the rate improves with r, and even for $r = 1$ we obtain a sharper bound. Finally, Schmidt-Hieber (2020) analyzed more general compositional functions in Hölder spaces. He showed that deep neural networks with bounded parameters and depth $L = \log_2 m$ achieve $O(m^{-2s/(2s+d_*)})$ for MIMs, though without optimization analysis; see also Kohler & Langer (2021); Kohler et al. (2022).

The computational complexities of gradient-based algorithms have also been studied for learning the single-index model ($d_* = 1$) and MIM recently. A quantity characterizing complexity is the information exponent (Arous et al., 2021), see, e.g., online SGD (Arous et al., 2021), GD (Ba et al., 2022; Moniri et al., 2023), and SGD (Mousavi-Hosseini et al., 2022; Damian et al., 2023). An alternative notion of complexity is given by the leap exponent (Abbe et al., 2023; Dandi et al., 2023; Bietti et al., 2025). For a broader discussion of MIMs, see the survey Bruna & Hsu (2025).

5 Numerical Experiments

In this section, we study the performance of the methods introduced in Section 3.2 on simulated datasets, using the Gaussian kernel $k : x, x' \mapsto \exp\left(-\gamma \|x - x'\|^2\right)$. Details on the experimental setup can be found in Appendix D.1.

Non-convexity of HKRR. Given the nonconvex nature of the objective function minimized in HKRR, the performance of first-order methods such as VarPro (Algorithm 1) and AGD (Algorithm 2) can be severely impacted by a poor initialization. In particular, VarPro directly exploits the structure of the problem in equation 5 by computing a closed-form solution at each iteration. This leads to a **faster convergence** than AGD, especially for few Nyström centers, since only a matrix of size $\tilde{m} \times \tilde{m}$ must be inverted. However, this advantage comes at a cost: because VarPro optimizes solely over B—the variable responsible for the non-convexity—it is prone to being **trapped in local minima** and is thus highly sensitive to initialization. By contrast, AGD explores the landscape of $\hat{\mathcal{L}}$ jointly in both B and α, which can help it escape critical points where VarPro stagnates. This behavior is illustrated in Figure 1a: in one scenario (top graph), both methods converge ultimately to the same solution, with VarPro reaching it more quickly; in the other scenario (bottom graph), AGD manages to escape a critical point in which VarPro remains stuck. A simple two-dimensional problem illustrates the above intuition, see Figure 1b, Figure 1c, and Appendix D.2 for further details.



(a) (b) (c)

Figure 1: Comparison between VarPro (red) and AGD (blue). (a) Training losses across time for two random initializations of B^0. (b) Two-dimensional toy example with initialization $(-1.5, -1.5)$: AGD escapes a local minimum where VarPro remains stuck. (c) Initialization $(-1.5, -0.1)$: both methods converge to minima, with VarPro being significantly faster. See Appendix D.2 for additional details.

Initialization and hyper-parameter tuning. To avoid the effect of poor initialization of B, the proposed strategy is to sample several matrices from $\mathcal{B}$, with 10 matrices sampled in the presented experiments. The initialization B^0 is then selected by cross-validation. This involves computing the coefficients that minimize $\hat{\mathcal{L}}(B, \alpha)$ and testing each pair of matrix and coefficients on a validation set. Since $\hat{\mathcal{L}}$ involves a regularization parameter λ, it must be initialized either by coupled cross-validation with B^0 or arbitrarily. For the Gaussian kernel used in these experiments, an additional scaling parameter γ must also be tuned. We adopt the well-known heuristic $\gamma = \frac{1}{2\tilde{\mu}^2}$, where $\tilde{\mu} = \text{median}\{\|B(x_i - x_j)\| : i \neq j\}$ is computed separately for each sampled matrix B.

On the role of the latent dimension. Beyond the conventional hyper-parameters of KRR, HKRR introduces the latent dimension d_*. Since this value is unknown in practice, it is crucial to understand how its estimate d affects performance. Figure 2 shows that underestimating d ($d < d_*$) severely reduces accuracy, while **overparameterization** is more robust: choosing $d > d_*$ often even outperforms the true value d_*. However, with a limited computational budget, very large d may degrade approximation quality. For larger datasets, setting $d = 20$ consistently yields better results than $d = d_*$, whereas setting $d = D = 50$ is inefficient due to the fixed budget.

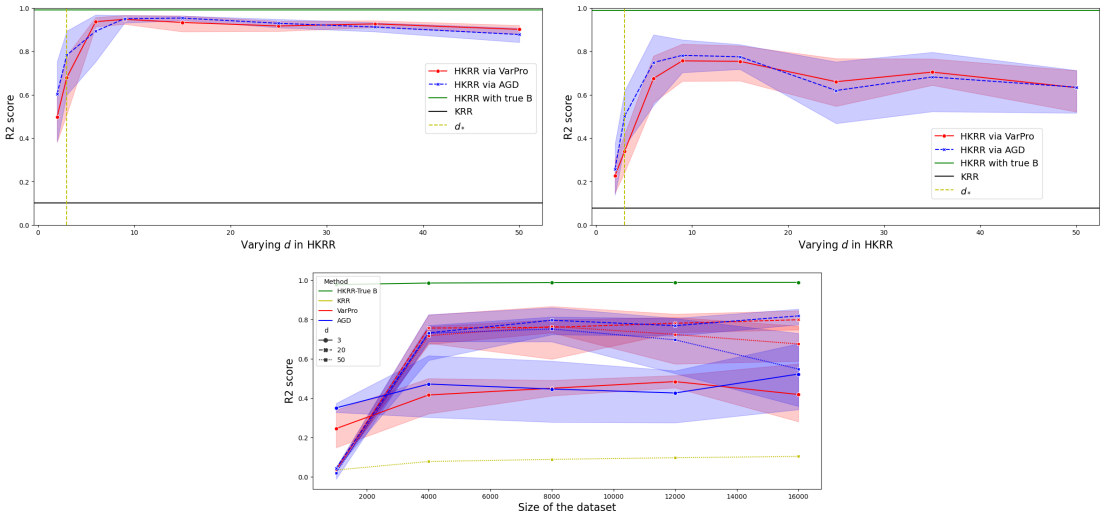


Figure 2: R2 score on test sets for B and α learned by VarPro (red) and AGD (blue). Top: performance w.r.t. the parameter d for Dataset 1 (left) and Dataset 2 (right), with true latent dimension $d_* = 3$, $D = 50$. Bottom: performance for $d \in \{3, 20, 50\}$ as the training size increases for Dataset 1. See Appendix D.1 for further details.

6 Conclusion

In this work, we investigated hyper-kernel ridge regression as a step towards exploring the compositional principle underlying deep learning. HKRR is an approach combining ideas from kernel methods and neural networks, related to recently proposed methods such as RFM. Our main contribution is the analysis of the sample complexity of HKRR when learning MIMs. Unlike standard KRR, HKRR can adapt to the MIM structure to escape the curse of dimensionality. From an algorithmic perspective, we exploit the structure of HKRR to analyze two approaches, VarPro and AGD, drawing ideas from convex optimization and for which local convergence guarantees can be established. Numerical results illustrate and corroborate our findings. Altogether, these results suggest that HKRR can be regarded as a useful augmentation of kernel methods, and point to new directions for developing efficient algorithms that bridge kernel and neural network approaches.

A natural direction for future work is to refine our analysis to obtain sharper bounds. It would be especially interesting to consider more general forms of compositional functions beyond MIMs, and see if kernel methods and neural network ideas can again be combined to provably learn such models.

REFERENCES

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2552–2623. PMLR, 2023.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3): 337–404, 1950.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical programming*, 137(1):91–129, 2013.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 2002.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Carl Bernd and Stephani Irmtraud. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, 1990.

Elie Bienenstock and Stuart Geman. Compositionality in neural systems. In *The handbook of brain theory and neural networks*, pp. 223–226. 1998.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow part i: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, 2025.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.

Joan Bruna and Daniel Hsu. Survey on algorithms for multi-index models. *arXiv preprint arXiv:2504.05426*, 2025.

Luca Calatroni and Antonin Chambolle. Backtracking strategies for accelerated descent methods with smooth composite objectives. *SIAM journal on optimization*, 29(3):1772–1798, 2019.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 8(02):161–183, 2010.

Yunlu Chen, Yang Li, Keli Liu, and Feng Ruan. Kernel learning in ridge regression "automatically" yields exact low rank solution. *arXiv preprint arXiv:2310.11736*, 2023.

Francesca Chiaromonte and R Dennis Cook. Sufficient dimension reduction and graphics in regression. *Annals of the Institute of Statistical Mathematics*, 54:768–795, 2002.

Jonathan Chirinos-Rodríguez, Ernesto De Vito, Cesare Molinari, Lorenzo Rosasco, and Silvia Villa. On learning the optimal regularization parameter in inverse problems. *Inverse Problems*, 40(12):125004, 2024.

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Wolfgang Dahmen. Compositional sparsity, approximation classes, and parametric transport equations. *Constructive Approximation*, pp. 1–65, 2025.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36:752–784, 2023.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

Ernesto De Vito, Lorenzo Rosasco, and Alessandro Rudi. Regularization: From inverse problems to large-scale machine learning. *Harmonic and Applied Analysis: From Radon Transforms to Machine Learning*, pp. 245–296, 2021.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

David L Donoho et al. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, 1(2000):32, 2000.

Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

Bertille Follain and Francis Bach. Enhanced feature learning via regularisation: Integrating neural networks and kernel methods. *arXiv preprint arXiv:2407.17280*, 2024.

Kenji Fukumizu and Chenlei Leng. Gradient-based kernel dimension reduction for regression. *Journal of the American Statistical Association*, 109(505):359–370, 2014.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, pp. 1871–1905, 2009.

Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

G. H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10(2):413–20, 04 1973.

Gene Golub and Victor Pereyra. Separable nonlinear least squares: the variable projection method and itsapplications. *Inverse problems*, 19(2):R1, 2003.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pp. 595–623, 2001.

David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

T Klock, A Lanteri, and S Vigogna. Estimating multi-index models with response-conditional least squares. *Electronic Journal of Statistics*, 15(1):589–629, 2021.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

Michael Kohler, Adam Krzyżak, and Sophie Langer. Estimation of a function of low local dimensionality by deep neural networks. *IEEE transactions on information theory*, 68(6):4032–4042, 2022.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1 (4):541–551, 1989.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Fanghui Liu, Lei Shi, Xiaolin Huang, Jie Yang, and Johan AK Suykens. Generalization properties of hyper-rkhs and its applications. *Journal of Machine Learning Research*, 22(140):1–38, 2021.

Songtao Lu, Mingyi Hong, and Zhengdao Wang. Pa-gd: On the convergence of perturbed alternating gradient descent to second-order stationary points for structured nonconvex optimization. In *International Conference on Machine Learning*, pp. 4134–4143. PMLR, 2019.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In *Conference on Learning Theory*, pp. 3351–3418. PMLR, 2021.

Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. When and why are deep networks better than shallow ones? In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with sgd. *arXiv preprint arXiv:2209.14863*, 2022.

T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497, 1990.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Clarice Poon and Gabriel Peyré. Smooth over-parameterized solvers for non-smooth structured optimization. *Mathematical programming*, 201(1):897–952, 2023.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Feature learning in neural networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022.

Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 383(6690):1461–1467, 2024a.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Dmitriy Drusvyatskiy. Linear recursive feature machines provably recover low-rank matrices. *arXiv preprint arXiv:2401.04553*, 2024b.

Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in neural information processing systems*, 28, 2015.

Alexander M Samarov. Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, 88(423):836–847, 1993.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.

Thomas Serre, Aude Oliva, and Tomaso Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429, 2007.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 1040–1053, 1982.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Rachel Ward and Tamara Kolda. Convergence of alternating gradient descent for matrix factorization. *Advances in Neural Information Processing Systems*, 36:22369–22382, 2023.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Yingcun Xia, Howell Tong, Wai Keung Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):363–410, 2002.

Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006.

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Libin Zhu, Damek Davis, Dmitriy Drusvyatskiy, and Maryam Fazel. Iteratively reweighted kernel machines efficiently learn sparse functions. *arXiv preprint arXiv:2505.08277*, 2025.

A PRELIMINARY LEMMAS AND BASIC ERROR BOUNDS FOR THEOREM 1

In this appendix, we prove Theorem 1 and some accompanying results. Many of these results hold true under weaker conditions than Assumption 1, so we treat these results and conditions separately. The proof of Theorem 1 is given in Subsection A.7.

In the following, if S is a compact space, the Banach space of continuous functions on S endowed with the sup norm $\|\cdot\|_\infty$ is denoted by $C(S)$. We also need to recall some basic quantities and fact associated to every RKHS.

A.1 RKHS AND RELATED OPERATORS

We recall that if k is continuous and bounded, then the following operators are well defined, bounded, and positive:

a) The integral operator $L_k : L_2(X, \rho_X) \to L_2(X, \rho_X)$

$$L_k(g)(x) = \int_X k(x, x')g(x') \, \mathrm{d}\rho_X(x'), \quad g \in L_2(X, \rho_X).$$

b) The covariance operator $\Sigma : \mathcal{H} \to \mathcal{H}$

$$\Sigma f = \int_X \langle f, k_x \rangle_{\mathcal{H}} \, k_x \, \mathrm{d}\rho_X(x) = \left[\int_X (k_x \otimes k_x) \, \mathrm{d}\rho_X(x) \right] f, \quad f \in \mathcal{H}, \tag{9}$$

where for all $x \in X$, $k_x := k(x, \cdot) \in \mathcal{H}$, and $(k_x \otimes k_x) : \mathcal{H} \to \mathcal{H}$ is the positive rank-one operator

$$(k_x \otimes k_x)(f) = \langle f, k_x \rangle_{\mathcal{H}} \, k_x.$$

Moreover, the relationship between the $L_2(\rho_X)$ norm and the RKHS norm is given by: for $g \in \mathcal{H}$,

$$\|g\|_{\rho_X}^2 = \|\Sigma^{\frac{1}{2}} g\|_{\mathcal{H}}^2, \tag{10}$$

where $\Sigma^{\frac{1}{2}}$ is the square root of the positive operator Σ, defined via spectral calculus.

A.2 COVERING NUMBER OF COMPOSITE CLASSES AND HYPER RKHS

Let V be a vector space endowed with a norm $\|\cdot\|_V$. The ball of radius R and centered at the origin is denoted by $\mathbb{B}_{V,R} = \{f \in V : \|f\|_V \leq R\}$. Given a subset $\mathcal{G} \subset V$ with compact closure, for all $\epsilon > 0$, $\mathcal{N}_V(\mathcal{G}, \epsilon)$ is the covering number of $\mathcal{G}$, defined as the minimal $J \in \mathbb{N}$ such that there exist $g_1, \ldots, g_J \in \mathcal{G}$ satisfying $\mathcal{G} = \bigcup_{j=1}^J \{g \in \mathcal{G} : \|g - g_j\|_V \leq \epsilon\}$.

If $\mathcal{H}$ is an RKHS on S with a continuous kernel, then $\mathcal{H}$ is a subspace of $C(S)$, and its ball of radius R is compact in $C(S)$ (Cucker & Zhou, 2007). We denote by $\mathcal{N}(\mathbb{B}_{\mathcal{H},R}, \epsilon)$ the corresponding covering number, omitting the index $C(S)$ for simplicity. We need the following condition on the data space.

Assumption 2. *The input space X is a compact subset of $\mathbb{R}^D$ such that $\sup_{x \in X} |x| \leq 1$, and for some $M > 0$, $|y| \leq M$.*

The assumption that X is bounded is needed to control the covering number, see Lemma 1. The assumption that the outputs are bounded implies $|f_*(x) = \int_Y y \, \mathrm{d}\rho(y|x)| \leq M$.

Given an integer d, we recall that $\mathcal{B} = \{B \in \mathbb{R}^{d \times D} : \|B\|_\infty \leq 1\}$, so that

$$\Omega = \{Bx \in \mathbb{R}^d \mid x \in X, B \in \mathcal{B}\} \subset \mathbb{R}^d$$

is compact, since the map $(x, B) \mapsto Bx$ is continuous and $X \times \mathcal{B}$ is compact.

We impose the following condition on the mother RKHS.

Assumption 3. *The mother space $\mathcal{H}$ is an RKHS on Ω with a continuous kernel k, and for all $g \in \mathcal{H}$,*

$$|g(x') - g(x)| \leq C_{\mathcal{H}} \, \|g\|_{\mathcal{H}} \, \|x' - x\|, \qquad x, x' \in \Omega,$$

for some constant $C_{\mathcal{H}} > 0$.

If k is defined on an open set $U \supset \Omega \times \Omega$ and $k \in C^1(U)$, then Assumption 3 always holds. The above assumption states that the elements of $\mathcal{H}$ are Lipschitz functions with a Lipschitz constant that is uniform on any ball of $\mathcal{H}$. Furthermore, for all $g \in \mathcal{H}$,

$$\|g\|_\infty \leq \kappa^{\frac{1}{2}} \|g\|_{\mathcal{H}},$$

where $\kappa = \sup_{x \in \Omega} k(x, x)$, which is finite since Ω is compact.

Recall that, for any $B \in \mathcal{B}$, the hypothesis space $\mathcal{H}_B$ is the RKHS with reproducing kernel

$$k_B(x, x') = k\big(Bx, Bx'\big), \quad x, x' \in X,$$

which is continuous and bounded by κ. Hence $\mathcal{H}_B \subset C(X)$ and, for all $f \in \mathcal{H}_B$,

$$\|f\|_\infty \leq \kappa^{\frac{1}{2}} \|f\|_B, \tag{11}$$

where $\|f\|_B = \|f\|_{\mathcal{H}_B}$ and $\mathbb{B}_{B,R} = \mathbb{B}_{\mathcal{H}_B,R}$.

Moreover, it holds that

$$\mathcal{H}_B = \{f : X \to \mathbb{R} \mid f = g \circ B \text{ for some } g \in \mathcal{H}\},$$

and

$$\|f\|_B = \min\{\|g\|_{\mathcal{H}} \mid f = g \circ B,\, g \in \mathcal{H}\}.$$

Since the minimum is achieved, for every $f \in \mathbb{B}_{B,R}$ there exists $g \in \mathbb{B}_{\mathcal{H},R}$ such that $f = g \circ B$.

The following lemma provides a bound on the covering number of $\bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R} \subset C(X)$.

Lemma 1. *Assume 2 and 3. Fix $\epsilon > 0$ and $R > 0$. Then*

$$\mathcal{N}\left(\bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}, \epsilon\right) \leq \left(\frac{6C_{\mathcal{H}} R}{\epsilon}\right)^{Dd} \mathcal{N}\big(\mathbb{B}_{\mathcal{H},R}, \tfrac{\epsilon}{2}\big)\,. \tag{12}$$

Proof. Let g_j be the covering centers of $\mathbb{B}_{\mathcal{H},R}$ with radius $\epsilon/2$, and let B_i be the covering centers of $\mathcal{B}$ with radius $\epsilon/(2C_{\mathcal{H}}R)$ (where $\mathcal{B}$ is regarded as a compact subset of the space V of $d \times D$ matrices endowed with the uniform norm). Then for any $g \in \mathbb{B}_{\mathcal{H},R}$ and $B \in \mathcal{B}$, there exist $g_j \in \mathbb{B}_{\mathcal{H},R}$ and $B_i \in \mathcal{B}$ such that $\|g - g_j\|_\infty \leq \epsilon/2$ and $\|B - B_i\|_\infty \leq \epsilon/(2C_{\mathcal{H}}R)$.

If we denote $f_\ell = g_j \circ B_i$, then $f_\ell \in \bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}$ because $g_j \in \mathbb{B}_{\mathcal{H},R}$ and $B_i \in \mathcal{B}$. For $f = g \circ B$ we have

$$\begin{aligned}
\|f - f_\ell\|_\infty &= \|g \circ B - g_j \circ B_i\|_\infty \\
&\leq \|g \circ B - g_j \circ B\|_\infty + \|g_j \circ B - g_j \circ B_i\|_\infty \\
&\leq \sup_{x \in X} |g(Bx) - g_j(Bx)| + \sup_{x \in X} |g_j(Bx) - g_j(B_i x)| \\
&\leq \sup_{x' \in \Omega} |g(x') - g_j(x')| + C_{\mathcal{H}} \|g_j\|_{\mathcal{H}} \sup_{x \in X} \|(B - B_i)x\| \\
&\leq \|g - g_j\|_\infty + C_{\mathcal{H}} \|g_j\|_{\mathcal{H}} \|B - B_i\|_\infty \sup_{x \in X} \|x\| \\
&\leq \frac{\epsilon}{2} + C_{\mathcal{H}} R \frac{\epsilon}{2C_{\mathcal{H}} R} \sup_{x \in X} \|x\| = \epsilon.
\end{aligned}$$

Here, we used the property $\|x\| \leq 1$, and the fact that g_j is Lipschitz with constant $C_{\mathcal{H}}\|g_j\|_{\mathcal{H}}$.

Therefore, we obtain an ϵ-cover of $\bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}$ with centers f_ℓ, induced by an $\epsilon/2$-cover of $\mathbb{B}_{\mathcal{H},R}$ and an $\epsilon/(2C_{\mathcal{H}}R)$-cover of $\mathcal{B}$. By the metric entropy of $\mathcal{B}$, we have

$$\begin{aligned}
\mathcal{N}\left(\bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}, \epsilon\right) &\leq \mathcal{N}_{M_{dD}}\left(\mathcal{B}, \tfrac{\epsilon}{2C_{\mathcal{H}}R}\right) \mathcal{N}\big(\mathbb{B}_{\mathcal{H},R}, \tfrac{\epsilon}{2}\big) \\
&\leq \left(\frac{6C_{\mathcal{H}}R}{\epsilon}\right)^{Dd} \mathcal{N}\big(\mathbb{B}_{\mathcal{H},R}, \tfrac{\epsilon}{2}\big)\,,
\end{aligned}$$

where the last inequality follows from the classical bound

$$\mathcal{N}_{M_{dD}}(\mathcal{B}, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^{Dd},$$

see (Zhang, 2023, Thm. 5.3). □

A.3 ERROR DECOMPOSITION

We recall that in the multi-index model

$$f_*(x) = g_*(B_* x), \qquad \rho_X\text{-a.e. } x \in X,$$

for some $d_* \times D$ matrix B_* with $\|B_*\|_\infty \leq 1$ and some measurable function g_*, which we can assume to be defined on Ω.

Note that if $f_* \in \mathcal{H}_{B_*}$, then

$$\mathcal{R}(f_*) = \min_{f: X \to \mathbb{R}} \mathcal{R}(f) \leq \inf_{B \in \mathcal{B}_*} \inf_{f \in \mathcal{H}_B} \mathcal{R}(f) = \mathcal{R}(f_*),$$

so that

$$\inf_{B \in \mathcal{B}_*} \inf_{f \in \mathcal{H}_B} \mathcal{R}(f) = \mathcal{R}(f_*), \tag{13}$$

indicating that HKRR provides a suitable criterion for the MIM.

In the following, we set $\mathcal{B}_* = \mathcal{B}$ with $d = d_*$, and recall that

$$\begin{aligned} \hat{f}_\lambda^B &= \arg\min_{f \in \mathcal{H}_B} \widehat{\mathcal{R}}_\lambda(f), \quad B \in \mathcal{B}_*, \\ \hat{B}_{d_*} &\in \arg\min_{B \in \mathcal{B}_*} \widehat{\mathcal{R}}_\lambda(\hat{f}_\lambda^B), \\ f_\lambda^{B_*} &= \arg\min_{f \in \mathcal{H}_{B_*}} \mathcal{R}_\lambda(f). \end{aligned}$$

Note that both $\hat{f}_\lambda^B$ and $f_\lambda^{B_*}$ exist and are unique. For simplicity, we assume that $\hat{B}_{d_*}$ also exists; otherwise, it suffices to consider an ϵ-minimizer.

We now state the following error decomposition for the excess risk of $\hat{f}_\lambda^{\hat{B}_{d_*}}$, where the main challenge lies in identifying a suitable intermediate term to incorporate, since there are many possible choices of f_λ^B and $\hat{f}_\lambda^B$ corresponding to different B.

Lemma 2. *Fix $R > 0$. Then*

$$\begin{aligned} \mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq & \underbrace{\sup_{f \in \bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}} \Big((\mathcal{R}(f) - \mathcal{R}(f_*)) - (\widehat{\mathcal{R}}(f) - \widehat{\mathcal{R}}(f_*)) \Big)}_{I} \\ & + \underbrace{\widehat{\mathcal{R}}(f_\lambda^{B_*}) - \widehat{\mathcal{R}}(f_*) - (\mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*))}_{II} + \underbrace{\|f_\lambda^{B_*} - f_*\|_{\rho_X}^2 + \lambda \|f_\lambda^{B_*}\|_{B_*}^2}_{III} \end{aligned} \tag{14}$$

for all training sets such that $\|\hat{f}_\lambda^{\hat{B}_{d_*}}\|_{\hat{B}_{d_*}} \leq R$.

Proof. The excess risk can be rewritten and decomposed as

$$\begin{aligned} & \mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \\ & = \mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \widehat{\mathcal{R}}(\hat{f}_\lambda^{\hat{B}_{d_*}}) + \widehat{\mathcal{R}}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \widehat{\mathcal{R}}(f_\lambda^{B_*}) + \widehat{\mathcal{R}}(f_\lambda^{B_*}) - \mathcal{R}(f_\lambda^{B_*}) + \mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*) \\ & \leq \mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \widehat{\mathcal{R}}(\hat{f}_\lambda^{\hat{B}_{d_*}}) + \widehat{\mathcal{R}}_\lambda(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \widehat{\mathcal{R}}_\lambda(f_\lambda^{B_*}) \\ & \quad + \widehat{\mathcal{R}}(f_\lambda^{B_*}) - \mathcal{R}(f_\lambda^{B_*}) + \lambda \|f_\lambda^{B_*}\|_{B_*}^2 + \mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*) \\ & \leq (\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \widehat{\mathcal{R}}(\hat{f}_\lambda^{\hat{B}_{d_*}})) + (\widehat{\mathcal{R}}(f_\lambda^{B_*}) - \mathcal{R}(f_\lambda^{B_*})) + \lambda \|f_\lambda^{B_*}\|_{B_*}^2 + \mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*) \\ & = \Big\{ \mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) - (\widehat{\mathcal{R}}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \widehat{\mathcal{R}}(f_*)) \Big\} + \Big\{ \widehat{\mathcal{R}}(f_\lambda^{B_*}) - \widehat{\mathcal{R}}(f_*) - (\mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*)) \Big\} \\ & \quad + \Big\{ \|f_\lambda^{B_*} - f_*\|_{\rho_X}^2 + \lambda \|f_\lambda^{B_*}\|_{B_*}^2 \Big\} \\ & \leq \sup_{f \in \bigcup_{B \in \mathcal{B}} \mathcal{H}_{R,B}} \Big\{ (\mathcal{R}(f) - \mathcal{R}(f_*)) - (\widehat{\mathcal{R}}(f) - \widehat{\mathcal{R}}(f_*)) \Big\} \\ & \quad + \Big\{ \widehat{\mathcal{R}}(f_\lambda^{B_*}) - \widehat{\mathcal{R}}(f_*) - (\mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*)) \Big\} + \Big\{ \|f_\lambda^{B_*} - f_*\|_{\rho_X}^2 + \lambda \|f_\lambda^{B_*}\|_{B_*}^2 \Big\}. \end{aligned} \tag{15}$$

Inequality in equation 15 follows from the fact that, by definition of $\hat{B}_{d_*}$, $\widehat{\mathcal{R}}_\lambda(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \widehat{\mathcal{R}}_\lambda(f_\lambda^{B_*}) \leq 0$. □

Note that by the definition of $\hat{f}_\lambda^{\hat{B}_{d_*}}$, $\widehat{\mathcal{R}}_\lambda(\hat{f}_\lambda^{\hat{B}_{d_*}}) \leq \widehat{\mathcal{R}}_\lambda(0)$, so that, under Assumption 2,

$$\|\hat{f}_\lambda^{\hat{B}_{d_*}}\|_{\hat{B}_{d_*}} \leq \frac{M}{\sqrt{\lambda}}, \tag{16}$$

hence we can always choose $R = M/\sqrt{\lambda}$.

The first two components in equation 14 are estimation errors, the first of which is typically more challenging to control since it depends on the complexity of the hypothesis space. The final component is the approximation error $\mathcal{A}(\lambda) = \inf_{f \in \mathcal{H}_{B_*}} \mathcal{R}(f) - \mathcal{R}(f_*) + \lambda \|f\|_{B_*}^2$, a quantity that has been extensively studied in the classical KRR (Cucker & Zhou, 2007; De Vito et al., 2021).

In what follows, we will concentrate on analyzing these estimation errors.

A.4 Estimation error I

The proof of this lemma follows a similar approach to that in Cucker & Zhou (2007), which is based on the following condition.

Assumption 4. *The covering number of $\mathbb{B}_{\mathcal{H},R}$ satisfies*

$$\log \mathcal{N}(\mathbb{B}_{\mathcal{H},R}, \epsilon) \leq c_1 (R/\epsilon)^{s^*}$$

for some $c_1 > 0, s^* > 0$.

Assumption 4 describes the complexity of hypothesis space using the concept of covering number, which is commonly used in literature (Cucker & Zhou, 2007; Zhou, 2002). There are different metrics to measure the complexity of RHKSs. Covering number quantifies the compactness of a space by measuring how many subsets with a fixed radius are needed to cover it. While entropy number (Steinwart & Christmann, 2008; Bernd & Irmtraud, 1990) represents the inverse concept by fixing the number of balls and determining the smallest radius needed to achieve that coverage. Eigenvalue decay (Caponnetto & De Vito, 2007), on the other hand, describes the smoothness or compactness of the space through the rate at which eigenvalues of covariance operators diminish. These measures are deeply interrelated, with covering numbers and entropy offering geometric and growth-based perspectives, while eigenvalue decay provides a spectral view of the hypothesis space. (Steinwart & Christmann, 2008, Chapter 5) provides a more detailed discussion there, see also Assumption 6.

Lemma 3. *Assume 2, 3 and 4. Fix $f_0 \in \bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}$ and $\delta \in (0,1)$, the following holds with confidence at least $1 - \frac{\delta}{2}$,*

$$\sup_{f \in \bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}} \left(\mathcal{R}(f) - \mathcal{R}(f_*) - (\widehat{\mathcal{R}}(f) - \widehat{\mathcal{R}}(f_*)) \right) \leq \frac{1}{2} \left(\mathcal{R}(f_0) - \mathcal{R}(f_*) \right) + \\ + C_3 \max\{1, R^2\} D d_* \max\left\{1, \log \frac{2}{\delta}\right\} \left(\frac{1}{m}\right)^{\frac{1}{s^*+1}},$$

where

$$C_3 = 360 \max\{1, M + \kappa\} \left(1 + c_1 + \log(3 C_{\mathcal{H}})\right). \tag{17}$$

Remark 7. *By inspecting the proof, it holds that for m large enough ($m > m_0$ where m_0 is given by equation 21), the factor $D d_*$ can be replaced by $(D d_*)^{\frac{1}{s^*+1}}$.*

Proof. Without loss of generality, we can assume that $R \geq 1$. Choose a function class

$$\mathcal{F} = \{F(x,y) | F(x,y) = (f(x) - y)^2 - (f_*(x) - y)^2, f \in \bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}\}.$$

Then $\mathbb{E}(F) = \mathcal{R}(f) - \mathcal{R}(f_*)$ and $\frac{1}{n} \sum_{i=1}^n F(x_i, y_i) = \widehat{\mathcal{R}}(f) - \widehat{\mathcal{R}}(f_*)$. By equation 11

$$\|f\|_\infty \leq \kappa \|f\|_B \leq \kappa R,$$

and $|f_*(x)| \leq M$, then

$$\begin{aligned}|F(z)| &= |(f(x) - f_*(x))(f(x) + f_*(x) - 2y)| \\ &\leq (\kappa R + M)(\kappa R + 3M),\end{aligned}$$

$|F(z) - \mathbb{E}(F)| \leq 2(\kappa R + M)(\kappa R + 3M)$ and $\mathbb{E}(F^2) \leq \|f - f_*\|_{\rho_X}^2 (\kappa R + M)(\kappa R + 3M) = (\kappa R + M)(\kappa R + 3M)\mathbb{E}(F)$. For $f_1, f_2 \in \bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}$, we have

$$|F_1(x, y) - F_2(x, y)| \leq 2(M + \kappa R)\|f_1 - f_2\|_\infty.$$

It follows that a $\frac{\epsilon}{2(M+\kappa R)}-$cover of $\bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}$ yields an $\epsilon-$cover of $\mathcal{F}$, that is,

$$\mathcal{N}(\mathcal{F}, \epsilon) \leq \mathcal{N}\left(\bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}, \frac{\epsilon}{2(M + \kappa R)}\right).$$

By taking $\alpha = \frac{1}{4}$ of Lemma 3.19 in Cucker & Zhou (2007), then with probability at least

$$1 - \mathcal{N}\left(\bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}, \frac{\epsilon}{8(M + \kappa R)}\right) \exp\left\{-\frac{3m\epsilon}{160(\kappa R + M)(\kappa R + 3M)}\right\} \tag{18}$$

there holds, for any $f_0 \in \bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}$

$$\begin{aligned}&\sup_{f \in \bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}} \left(\mathcal{R}(f) - \mathcal{R}(f_*) - (\widehat{\mathcal{R}}(f) - \widehat{\mathcal{R}}(f_*))\right) \leq \sqrt{\epsilon}\sqrt{\mathcal{R}(f_0) - \mathcal{R}(f_*) + \epsilon} \\ &\leq \frac{1}{2}(\mathcal{R}(f_0) - \mathcal{R}(f_*)) + \epsilon.\end{aligned}$$

Fixed $\delta \in (0, 1)$, we choose ϵ such that

$$\mathcal{N}\left(\bigcup_{B \in \mathcal{B}} \mathbb{B}_{B,R}, \frac{\epsilon}{8(M + \kappa R)}\right) \exp\left\{-\frac{3m\epsilon}{160(\kappa R + M)(\kappa R + 3M)}\right\} \leq \delta/2.$$

By equation 18 and Lemma 1, we need to solve

$$\left(\frac{48(M + \kappa R)C_{\mathcal{H}} R}{\epsilon}\right)^{Dd_*} \mathcal{N}\left(\mathbb{B}_{\mathcal{H},R}, \frac{\epsilon}{16(M + \kappa R)}\right) \cdot \exp\left\{-\frac{3m\epsilon}{160(\kappa R + M)(\kappa R + 3M)}\right\} \leq \frac{\delta}{2}. \tag{19}$$

Let $x = 16R(M + \kappa R)/\epsilon > 0$. Set

$$A = Dd_* \log(3C_{\mathcal{H}}) + \log\frac{2}{\delta}, \quad B = \frac{3R}{10(\kappa R + 3M)},$$

taking into account condition (4), the above inequality becomes

$$Dd_* \log x + c_1 x^{s^*} - Bmx^{-1} + A \leq 0.$$

Since $\log x \leq x^{s^*}$, the above inequality is satisfied if

$$(Dd_* + c_1)x^{s^*} - Bmx^{-1} + A \leq 0,$$

which is equivalent to

$$x^{s^*+1} + ax - mb \leq 0, \quad a = \frac{A}{Dd_* + c_1},\ b = \frac{B}{Dd_* + c_1}.$$

The function $\varphi(x) = x^{s^*+1} + ax$ is continuous and strictly increasing on $(0, +\infty)$, tends to 0 as $x \to 0^+$, and diverges to $+\infty$ as $x \to +\infty$. Hence there is a unique $x_m \in (0, +\infty)$ such that $\varphi(x_m) = mb$, and the above inequality is satisfied for all $0 < x < x_m$.

Since $\varphi(1) = 1 + a$, it holds that

$$\begin{cases} x_m \geq 1 & \text{if } 1 + a \leq mb, \\ x_m < 1 & \text{if } 1 + a > mb. \end{cases} \tag{20}$$

If $1 + a \leq mb$, since $s^* + 1 > 1$, for all $x \geq 1$ we have

$$\varphi(x) \leq (1 + a)\, x^{s^*+1},$$

so that

$$x_m \geq \left(\frac{mb}{1+a}\right)^{\frac{1}{s^*+1}} \geq 1.$$

If $1 + a > mb$, then for all $x \leq 1$,

$$\varphi(x) \leq (1 + a)\, x,$$

so that

$$1 > x_m \geq \frac{mb}{1+a}.$$

Hence inequity 19 is satisfied if

$$\epsilon \geq 16R(M + \kappa R)\left(\frac{10(\kappa R + 3M)}{3R}\right)^{t_m}\left(Dd_* + c_1 + Dd_* \log(3C_{\mathcal{H}}) + \log\frac{2}{\delta}\right)^{t_m}\left(\frac{1}{m}\right)^{t_m},$$

where

$$t_m = \begin{cases} 1, & m < m_0, \\ \frac{1}{s^*+1}, & m \geq m_0, \end{cases} \qquad m_0 = \frac{3R}{10(\kappa R + 3M)\left(Dd_* + c_1 + Dd_* \log(3C_{\mathcal{H}}) + \log\frac{2}{\delta}\right)}. \tag{21}$$

Taking into account that $R \geq 1$, the above inequality is implied by

$$\epsilon \geq 160R^2(M + \kappa)^{t_m+1}\left(Dd_* + c_1 + Dd_* \log(3C_{\mathcal{H}}) + \log\frac{2}{\delta}\right)^{t_m}\left(\frac{1}{m}\right)^{t_m}.$$

Since $a + b \leq 2ab$ for all $a, b \geq 1$ and $Dd_* \geq 1$ then

$$Dd_* + c_1 + Dd_* \log(3C_{\mathcal{H}}) + \log\frac{2}{\delta} \leq 2\, Dd_*\,(1 + c_1 + \log(3C_{\mathcal{H}}) \max\{1, \log\frac{2}{\delta}\},$$

so that bound in equation 17 is a consequence of the fact that $\frac{1}{1+s^*} \leq t_m \leq 1$.

□

This lemma shows that the largest error can be bounded in terms of $\mathcal{R}(f) - \mathcal{R}(f_*)$ for any $f \in \bigcup_{B\in\mathcal{B}} \mathbb{B}_{B,R}$. In particular, by taking $f = \hat{f}_\lambda^{\hat{B}_{d_*}}$, the excess risk of HKRR appears in the upper bound, which is essential for the full excess risk analysis. Moreover, the radius R in equation 17 depends on assumptions about the hypothesis space and may vary across different settings. Specifically, under the assumptions of Theorem 5 and Theorem 6 we can set $R = M/\sqrt{\lambda}$, while under the assumptions of Theorem 1, R can be chosen as $R \simeq \sqrt{\mathcal{A}(\lambda)/\lambda} + 1$ (see Lemma 6).

A.5 Estimation error II

Note that to bound the item $\widehat{\mathcal{R}}(f_\lambda^{B_*}) - \widehat{\mathcal{R}}(f_*) - (\mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*))$, the primary error arises from the difference between $\widehat{\mathcal{R}}$ and $\mathcal{R}$, which reflects the discrepancy between integration and discretization.

Recall that the approximation error is defined by

$$\mathcal{A}(\lambda) = \inf_{f\in\mathcal{H}_{B_*}} \mathcal{R}(f) - \mathcal{R}(f_*) + \lambda\|f\|_{B^*}^2 = \mathcal{R}(f_\lambda^{B^*}) + \lambda\|f_\lambda^{B^*}\|_{B_*}^2$$

The following is a restatement of a result from Cucker & Zhou (2007). We provide a proof for the sake of completeness.

Lemma 4. *Assume* 2*, the following holds with probability at least* $1 - \delta/2$*,*

$$\begin{aligned}&\widehat{\mathcal{R}}(f_\lambda^{B_*}) - \widehat{\mathcal{R}}(f_*) - (\mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*)) \\ &\leq \left(\frac{14\kappa^2 \log(2/\delta)}{3m\lambda} + 1\right)\mathcal{A}(\lambda) + \frac{42M^2 \log(2/\delta)}{m}.\end{aligned}$$

Proof. Consider a random variable ξ with $f_\lambda^{B_*} \in \mathcal{H}_{B_*}$, $\|f_\lambda^{B_*}\|_\infty \le R'$ as

$$\xi(x,y) = (f_\lambda^{B_*}(x) - y)^2 - (f_*(x) - y)^2.$$

then $|\xi| \le (R' + 3M)^2 = C_4^{'}$, $|\xi - \mathbb{E}(\xi)| \le 2C_4^{'}$, $\mathbb{E}(\xi) = \mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*) \ge 0$ and $\mathbb{E}(\xi^2) \le C_4^{'}\mathbb{E}(\xi)$. Then by Bernstein's inequality, we have

$$\widehat{\mathcal{R}}(f_\lambda^{B_*}) - \widehat{\mathcal{R}}(f_*) - (\mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*)) \le \epsilon$$

holds with confidence $1 - \delta/2$ with

$$\frac{\delta}{2} = \exp\left\{-\frac{m\epsilon^2}{2C_4^{'}\mathbb{E}(\xi) + \frac{4}{3}C_4^{'}\epsilon}\right\}.$$

Solving the quadratic equation for ϵ tells us with confidence at least $1 - \delta/2$

$$\begin{aligned}
&\widehat{\mathcal{R}}(f_\lambda^{B_*}) - \widehat{\mathcal{R}}(f_*) - (\mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*)) \\
&\le \frac{\frac{2}{3}C_4^{'}\log\frac{2}{\delta} + \sqrt{\frac{4}{9}(C_4^{'})^2(\log\frac{2}{\delta})^2 + 2mC_4^{'}\log\frac{2}{\delta}\mathbb{E}(\xi)}}{m} \\
&\le \frac{4C_4^{'}\log\frac{2}{\delta}}{3m} + \sqrt{\frac{2C_4^{'}\log\frac{2}{\delta}\mathbb{E}(\xi)}{m}}.
\end{aligned}$$

Applying the elementary inequality with the dual number p' and p

$$ab \le \frac{1}{p}a^p + \frac{1}{p'}b^{p'} \quad \forall a, b > 0$$

to $p' = p = 2$, $a = \left(\frac{2C_4^{'}\log(2/\delta)}{m}\right)^{1/2}$, and $b = (\mathbb{E}(\xi))^{1/2}$, we get

$$\sqrt{\frac{2C_4^{'}\log\frac{2}{\delta}\mathbb{E}(\xi)}{m}} \le \frac{C_4^{'}\log\frac{2}{\delta}}{m} + \frac{1}{2}\mathbb{E}(\xi).$$

Hence, with confidence at least $1 - \delta/2$, we have

$$\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbb{E}(\xi) \le \frac{4C_4^{'}\log\frac{2}{\delta}}{3m} + \frac{C_4^{'}\log\frac{2}{\delta}}{m} + \mathbb{E}(\xi).$$

Note that for all $\lambda > 0$,

$$\|f_\lambda^{B_*}\|_{B_*} \le \sqrt{\mathcal{A}(\lambda)/\lambda} \quad \text{and} \quad \|f_\lambda^{B_*}\|_\infty \le \kappa\sqrt{\mathcal{A}(\lambda)/\lambda}.$$

In fact, since f_* is a minimizer of $\mathcal{R}(f)$, we know that

$$\lambda\|f_\lambda^{B_*}\|_{B_*}^2 \le \mathcal{R}(f_\lambda^{B_*}) - \mathcal{R}(f_*) + \lambda\|f_\lambda^{B_*}\|_{B_*}^2 = \mathcal{A}(\lambda).$$

And the second follows from $\|f_\lambda^{B_*}\|_\infty \le \kappa\|f_\lambda^{B_*}\|_{B_*}$. Thus, by taking $R' = \kappa\sqrt{\mathcal{A}(\lambda)/\lambda}$, it follows that $C_4^{'} \le 2\kappa^2\mathcal{A}(\lambda)/\lambda + 18M^2$ and

$$\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mathbb{E}(\xi) \le \left(\frac{14\kappa^2\log(2/\delta)}{3m\lambda} + 1\right)\mathcal{A}(\lambda) + \frac{42M^2\log(2/\delta)}{m}. \tag{22}$$

□

This inequality offers a tighter bound with respect to the sample size m compared to classical concentration inequalities. This explains why we add the intermediate terms $\widehat{\mathcal{R}}(f_*)$ and $\mathcal{R}(f_*)$ in the error decomposition step (15), which allows for a more refined bound on the variance of the random variables.

A.6 BASIC ERROR BOUND

As a consequence of the above results and the trivial bound (16), we obtain our first main result.

Theorem 5. *Assume 2, 3, and 4. Let $\delta > 0$. Then, with confidence at least $1 - \delta$,*

$$\begin{aligned}\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq 2C_3 \max\left\{1, \frac{M^2}{\lambda}\right\} \left(\frac{1}{m}\right)^{\frac{1}{s^*+1}} \\ + \left(4 + \frac{28\kappa^2 \log(2/\delta)}{3m\lambda}\right) \mathcal{A}(\lambda) + \frac{84M^2 \log(2/\delta)}{m},\end{aligned} \tag{23}$$

where C_3 is given by equation 17.

Proof. By equation 16, bound in equation 14 holds with $R = M/\sqrt{\lambda}$. Taking $f = \hat{f}_\lambda^{\hat{B}_{d_*}}$ on the right-hand side of Lemma 3 bounds I, Lemma 4 bounds II, and the definition of $\mathcal{A}(\lambda)$ yields III. Hence, with confidence at least $1 - \delta$,

$$\begin{aligned}\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq 2C_3 \max\left\{1, \tfrac{M^2}{\lambda}\right\} Dd_* \max\left\{1, \log\frac{2}{\delta}\right\} \left(\frac{1}{m}\right)^{\frac{1}{s^*+1}} \\ + 2\left(\frac{14\kappa^2 \log(2/\delta)}{3m\lambda} + 1\right) \mathcal{A}(\lambda) + \frac{84M^2 \log(2/\delta)}{m} + 2\mathcal{A}(\lambda).\end{aligned}$$

□

Note that in the above error bound, $\mathcal{A}(\lambda)$ is the approximation error, which depends on both the hypothesis space $\mathcal{H}_{B_*}$ and the properties of the target function f_*. It decreases as λ increases. By contrast, s^* describes the complexity of the ambient hypothesis space $\mathcal{H}$ (not $\mathcal{H}_{B_*}$); it is typically determined by the intrinsic input dimension d_* (rather than the ambient dimension D) and the regularity of $\mathcal{H}$.

The approximation error under so-called source conditions has been extensively studied in the context of classical kernel methods. The following is a standard formulation, where $L_{k_{B_*}}$ denotes the integral operator

$$L_{k_{B_*}} : L^2(X, \rho_X) \to L^2(X, \rho_X), \qquad (L_{k_{B_*}} f)(x) = \int_X k_{B_*}(x, x')\, f(x')\, \mathrm{d}\rho_X(x'), \tag{24}$$

which is positive, so that for any $\theta > 0$, the fractional power $L_{k_{B_*}}^\theta$ is well defined by spectral calculus.

Assumption 5. *There exists $\theta \in (0, 1]$ such that $f_* \in \mathrm{Range}\big(L_{k_{B_*}}^{\theta/2}\big)$.*

Assumption 5 states that f_* is not arbitrary, but belongs to a smoother subspace determined by the integral operator $L_{k_{B_*}}$ associated with the kernel k_{B_*}. The parameter θ quantifies the smoothness of f_*: larger values of θ correspond to smoother target functions, smaller hypothesis spaces, and therefore better approximation rates.

Under the above assumption, we have the following classical result (Cucker & Zhou, 2007; De Vito et al., 2021).

Lemma 5. *Under Assumption 5,*

$$\mathcal{A}(\lambda) = \|f_\lambda^{B_*} - f_*\|_{\rho_X}^2 + \lambda \|f_\lambda^{B_*}\|_{B_*}^2 \;\leq\; \lambda^\theta \|L_{B_*}^{-\frac{\theta}{2}} f_*\|_{\rho_X}^2. \tag{25}$$

We are now ready to state a main result, whose rate is slower than Theorem 1. To simplify the statement, we introduce the following smoothness assumption on the mother kernel.

Assumption 6. *The mother kernel k is defined on an open set $U \supset \mathcal{H}_{R_0} \times \mathcal{H}_{R_0}$ for some $R_0 > 1$, and $k \in C^r(U)$ for some $r \in \mathbb{N}$ with $r \geq 1$.*

Remark 8. *By Assumption 2 and the definition of $\mathcal{B}$, we have*

$$\Omega \subset \mathbb{B}_{\mathcal{H},1} \subset \mathbb{B}_{\mathcal{H},R_0},$$

so that we can apply (Steinwart & Christmann, 2008, Th. 6.26).

Remark 9. *As shown in (Steinwart & Christmann, 2008, Th. 6.26 and the subsequent remark), under Assumption 6, s^* in Assumption 4 equals to $s^* = d_*/r$. Moreover, since $r \geq 1$, it also implies Assumption 3.*

Examples of kernels satisfying these assumptions include the Matérn kernel with parameter 2.5, polynomial kernels of degree greater than 2, and others. The following theorem is an immediate consequence of the error decomposition and the results above, and its proof is therefore omitted.

Theorem 6. *Assume 2, 5 and 6. Fix $0 < \delta < 1$, with probability at least $1 - \delta$*

$$\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq 2C_3 \max\left\{1, \frac{M^2}{\lambda}\right\} Dd_* \max\{1, \log\frac{2}{\delta}\} \left(\frac{1}{m}\right)^{\frac{r}{d_*+r}} + \\ + \frac{28\kappa^2 \log(2/\delta)}{3m\lambda^{1-\theta}} + \frac{84M^2\log(2/\delta)}{m} + 4\lambda^\theta, \tag{26}$$

where C_3 is given by equation 17. By taking $\lambda = M^2 m^{-\frac{1}{(1+\theta)(1+d_*/r)}}$, there holds

$$\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq C_4 \left(Dd_*\right) \log\frac{2}{\delta} \left(\frac{1}{m}\right)^{\frac{r\theta}{(1+\theta)(d_*+r)}}$$

with $C_4 = M^2\left(C_3 + 28M^2\kappa^2 + 88\right)$.

Classical results with the same assumptions on hypothesis space are of order $m^{n\theta/(1+\theta)(D+n)}$ (Cucker & Zhou, 2007). When the input dimension D is exceptionally large, as is often the case with the rise of big data, the rate is adversely impacted by the curse of dimensionality. However, as noted in Theorem 6, the exponential dependence on m is governed by d rather than D, with the dependence on D being polynomial, which helps mitigate the curse of dimensionality. Moreover, when s is sufficiently large and $\theta = 1$, the rate in Theorem 6 asymptotically reduces to $O(m^{-1/2})$. Although this is slower than the $O(m^{-1})$ rate of Theorem 1 under the additional sample size condition, the proof techniques are essentially the same.

A.7 Refined error bound: proof of Theorem 1

The following lemma provides a more refined bound than the bound in equation 16 under stricter assumptions on the sample size. A detailed proof is available in (Cucker & Zhou, 2007, Lemma 8.19).

Lemma 6. *Under Assumptions 2, 3, and 6, suppose $\zeta < 1/(1+s^*)$ with $s^* = d_*/r$ and choose $\lambda_m = m^{-\zeta}$. Fix $0 < \delta < 1$. Then, with confidence at least with confidence $1 - 3\delta/(1/(1+s^*) - \zeta)$, we have*

$$\|\hat{f}_\lambda^{\hat{B}_{d_*}}\|_{\hat{B}_{d_*}} \leq c_2\sqrt{\log(2/\delta)}(\sqrt{\mathcal{A}(\lambda_m)/\lambda_m} + 1) = R^* \tag{27}$$

for all $m \geq m_\delta$. Here $c_2 > 0$ is a constant depending only on s^*, ζ, κ, and M, and

$$m_\delta = \max\left\{(108/c_1)^{1/s^*}(\log(2/\delta))^{1+1/s^*}, (1/2c_3)^{2/(\zeta - 1/(1+s^*))}\right\} \tag{28}$$

with $c_3 = (2\kappa+5)(108c_1)^{1/(1+s^*)}$.

Now we are ready to prove Theorem 1.

Proof of Theorem 1. Assumption 1 states that Assumptions 2, 3, and 6 hold. Recall that Assumption 6 implies Assumption 4 with $s_* = d_*/r$. We can now apply the error decomposition in Lemma 2, together with the corresponding bounds for each term, to derive the excess risk.

Take $f_0 = \hat{f}_\lambda^{\hat{B}_{d_*}}$ on the right-hand side of the inequality in Lemma 3 with $R = R^*$ given in equation 27 to bound item I, combine with Lemma 4 to bound item II, and use Lemma 5 for item III. Then, with confidence at least $1 - \delta$,

$$\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq 2C_3 Dd_* c_2^2 \log^2\left(\tfrac{2}{\delta}\right)\left(\frac{\mathcal{A}(\lambda)}{\lambda} + 1\right)\left(\frac{1}{m}\right)^{\frac{1}{s^*+1}} \\ + 2\left(\frac{14\kappa^2\log(2/\delta)}{3m\lambda} + 1\right)\mathcal{A}(\lambda) + \frac{84M^2\log(2/\delta)}{m} + 2\mathcal{A}(\lambda), \tag{29}$$

for all $m \geq m_\delta$. Moreover, since $\mathcal{A}(\lambda) \leq \lambda^\theta$ by Lemma 5 and $\lambda = m^{-\zeta}$, we obtain

$$\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq 28\kappa^2\log(2/\delta)\, m^{-1+(1-\theta)\zeta} + \frac{84M^2\log(2/\delta)}{m} + 4m^{-\theta\zeta} \\ + 2C_3 Dd_* c_2^2 \log^2(2/\delta)\, m^{-\frac{r}{d_*+r}+(1-\theta)\zeta}. \tag{30}$$

Since $\delta < 2/e$, we have

$$1 \leq \log(2/\delta) \leq \log^2(2/\delta).$$

Therefore, the dominant terms with respect to m in equation 30 are the third and the last ones, because $-1 > -1 + (1-\theta)\zeta > -\frac{r}{d_*+r} + (1-\theta)\zeta$. If $\zeta < r/(d_* + r)$, then the rate becomes

$$\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*}}) - \mathcal{R}(f_*) \leq C_1 D d_* \log^2(2/\delta)\, m^{-\theta\zeta},$$

where

$$C_1 = 28\kappa^2 + 84M^2 + 4 + 2C_3 c_2^2. \tag{31}$$

This completes the proof. □

B Proofs of other excess risk bounds in Section 4

In this part, we will give the proofs of excess risk rates for Theorems 2 and Theorem 3.

B.1 Rate of Nyström approximation: proof of Theorem 2

Before proving Theorem 2, we introduce an alternative strategy for selecting the Nyström points based on approximate leverage scores (ALS), referred to as the ALS Nyström approximation. This method, together with some necessary definitions, will be included in the next theorem. The leverage scores associated to points $(x_i)_{i=1}^m$ are

$$(\ell_i(t))_{i=1}^m, \quad \ell_i(t) = \left(\hat{K}^B(\hat{K}^B + t\,m\,I)^{-1}\right)_{ii}, \quad i \in \{1, \ldots, m\}, \quad t > 0,$$

where $(\hat{K}^B)_{ij} = k(Bx_i, Bx_j)$. Computing these scores exactly can be challenging in practice; thus, one may consider approximations $(\hat{\ell}_i(t))_{i=1}^m$ (Drineas et al., 2012; Rudi et al., 2015). Given $t_0 > 0$, $T \geq 1$ and confidence level $\delta > 0$, we say that $(\hat{\ell}_i(t))_{i=1}^m$ are (T, t_0)-approximate leverage scores with probability at least $1 - \delta$, if

$$\frac{1}{T}\ell_i(t) \leq \hat{\ell}_i(t) \leq T\ell_i(t), \quad t \geq t_0, \quad i = 1, \ldots, m.$$

The ALS sampling selects the Nyström points $(\tilde{x}_i)_{i=1}^{\tilde{m}}$ independently with replacement from the training set, where each x_i is selected with probability $p_t(i) = \hat{\ell}_i(t)/\sum_j \hat{\ell}_j(t)$.

Theorem 7 (Extension of Theorem 2)**.** *Under the same assumptions as Theorem 1, the following holds with probability at least* $1 - \delta$*,*

$$\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \mathcal{R}(f_*) \leq C_2 D d_* \log^2(2/\delta)\,(m)^{-\theta\zeta}$$

with C_2 *given explicitly in equation 34 under conditions:*

1. *for plain Nyström,* $\tilde{m} \geq 67 \log \frac{4\kappa}{\lambda\delta} \vee 5\mathcal{N}_{B_*,\infty}(\lambda) \log \frac{4\kappa}{\lambda\delta}$.
2. *for ALS Nyström and* (T, t_0)*-approximate leverage scores with subsampling probabilities* p_t*,*

$$m \geq 1655\kappa + 223\kappa \log \frac{2\kappa}{\delta},$$
$$t_0 \vee \frac{19\kappa}{m} \log \frac{2m}{\delta} \leq \lambda \leq \left\|\Sigma_{B_*}\right\|,$$
$$\tilde{m} \geq (334 \vee 78\,T^2\,\mathcal{N}_{B_*}(\lambda)) \log \frac{8m}{\delta}.$$

Proof of Theorem 2(Theorem 7). Let $P_{B,\tilde{m}} : \mathcal{H}_B \to \mathcal{H}_B$ denote the orthogonal projection from $\mathcal{H}_{B,m}$ onto the subspace $\mathcal{H}_{B,\tilde{m}} \subset \mathcal{H}_{B,m} \subset \mathcal{H}_B$. Recall that $\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}} = \arg\min_{B\in\mathcal{B}} \min_{f\in\mathcal{H}_{B,\tilde{m}}} \widehat{R}_\lambda(f)$. We decompose the

excess risk as follows:

$$
\begin{aligned}
&\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \mathcal{R}(f_*) \\
&= \mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \widehat{\mathcal{R}}_\lambda(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) + \widehat{\mathcal{R}}_\lambda(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \widehat{\mathcal{R}}_\lambda(P_{B_*,\tilde{m}} f_\lambda^{B_*}) \\
&\quad + \widehat{\mathcal{R}}_\lambda(P_{B_*,\tilde{m}} f_\lambda^{B_*}) - \mathcal{R}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) + \mathcal{R}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) - \mathcal{R}(f_*) + \lambda \|P_{B_*,\tilde{m}} f_\lambda^{B_*}\|_{\mathcal{H}}^2 \\
&\leq \mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \widehat{\mathcal{R}}_\lambda(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) + \widehat{\mathcal{R}}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) - \mathcal{R}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) \\
&\quad + \mathcal{R}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) - \mathcal{R}(f_*) + \lambda \|P_{B_*,\tilde{m}} f_\lambda^{B_*}\|_{\mathcal{H}}^2 \\
&\leq \left\{ \mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \widehat{\mathcal{R}}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) \right\} + \left\{ \widehat{\mathcal{R}}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) - \mathcal{R}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) \right\} \\
&\quad + \left\{ \mathcal{R}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) - \mathcal{R}(f_*) + \lambda \|P_{B_*,\tilde{m}} f_\lambda^{B_*}\|_{\mathcal{H}}^2 \right\}.
\end{aligned} \tag{32}
$$

Here, the first inequality follows because $\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}$ minimizes $\widehat{\mathcal{R}}_\lambda$ over $\mathcal{H}_{B,\tilde{m}}$. The first two terms are called estimation errors, whose controls are identical to those in the proof of Theorem 1 since $\mathcal{H}_{B,\tilde{m}} \subset \mathcal{H}_B$. For the last term, we further decompose it as

$$
\begin{aligned}
&\mathcal{R}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) - \mathcal{R}(f_*) + \lambda \|P_{B_*,\tilde{m}} f_\lambda^{B_*}\|_{\mathcal{H}}^2 \\
&\leq \|P_{B_*,\tilde{m}} f_\lambda^{B_*} - f_*\|_{\rho_X}^2 + \lambda \|f_\lambda^{B_*}\|_{B^*}^2 \\
&\leq \|P_{B_*,\tilde{m}} f_\lambda^{B_*} - f_\lambda^{B_*}\|_{\rho_X}^2 + \|f_\lambda^{B_*} - f_*\|_{\rho_X}^2 + \lambda \|f_\lambda^{B_*}\|_{B^*}^2.
\end{aligned}
$$

Denote the first term as $\mathcal{C}(\lambda) := \|P_{B_*,\tilde{m}} f_\lambda^{B_*} - f_\lambda^{B_*}\|_{\rho_X}^2$, which is the so-called computational error as given in Rudi et al. (2015). And the left part is the approximation error $\mathcal{A}(\lambda)$. Let Σ_{B_*} be the covariance operator associated with the kernel k_{B_*}, see equation 9, by the relationship between L_2 norm and RKHS norm (10), we have

$$
\begin{aligned}
\|P_{B_*,\tilde{m}} f_\lambda^{B_*} - f_\lambda^{B_*}\|_{\rho_X}^2 &= \|\Sigma_{B_*}^{1/2}(I - P_{B_*,\tilde{m}}) f_\lambda^{B_*}\|_{B^*}^2 \\
&\leq \|(I - P_{B_*,\tilde{m}})\Sigma_{B_*}^{1/2}\|^2 \|f_\lambda^{B_*}\|_{B^*}^2.
\end{aligned}
$$

Using the estimate $\|f_\lambda^{B_*}\|_{\mathcal{H}_{B_*}}^2 \leq \mathcal{A}(\lambda)/\lambda$ and applying (Rudi et al., 2015, Lemma 6), if $\tilde{m} \geq 67 \log \frac{4\kappa^2}{\lambda\delta} \vee 5\mathcal{N}_\infty(\lambda) \log \frac{4\kappa^2}{\lambda\delta}$, then with probability at least $1-\delta$, $\|(I - P_{B_*,\tilde{m}})\Sigma_{B_*}^{1/2}\|^2 \leq 3\lambda$. Therefore,

$$
\mathcal{R}(P_{B_*,\tilde{m}} f_\lambda^{B_*}) - \mathcal{R}(f_*) + \lambda \|P_{B_*,\tilde{m}} f_\lambda^{B_*}\|_{\mathcal{H}}^2 \leq 3\lambda \frac{\mathcal{A}(\lambda)}{\lambda} + \mathcal{A}(\lambda) = 4\mathcal{A}(\lambda), \tag{33}
$$

which coincides with the order of the approximation error in the proof of Theorem 1.

By adding the terms $\mathcal{R}(f_*)$ and $\widehat{\mathcal{R}}(f_*)$ to the first two components of equation 32, the estimation errors can be controlled using the same argument as in Theorem 1. Specifically, applying Lemma 3 by taking $f_0 = \hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}$, $R = c_2\sqrt{\log(2/\delta)}\left(\sqrt{\mathcal{A}(\lambda_m)/\lambda_m} + 1\right)$, $s^* = d_*/r$, and together with Lemma 4, we obtain

$$
\begin{aligned}
\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \mathcal{R}(f_*) \leq{}& 2C_3\, D d_*\, c_2^2 \log^2\!\left(\tfrac{2}{\delta}\right)\left(\tfrac{\mathcal{A}(\lambda)}{\lambda} + 1\right)\left(\tfrac{1}{m}\right)^{\frac{1}{s^*+1}} \\
&+ 2\left(\tfrac{14\kappa^2 \log(2/\delta)}{3m\lambda} + 1\right)\mathcal{A}(\lambda) + \tfrac{84M^2 \log(2/\delta)}{m} + 8\mathcal{A}(\lambda).
\end{aligned}
$$

The remainder of the proof follows the same computations as in Theorem 1, yielding

$$
\mathcal{R}(\hat{f}_\lambda^{\hat{B}_{d_*,\tilde{m}}}) - \mathcal{R}(f_*) \leq C_2\, D d_*\, \log^2\!\left(\tfrac{2}{\delta}\right) m^{-\theta\zeta},
$$

where

$$
C_2 = 28\kappa^2 + 84M^2 + 10 + 2C_3 c_2^2. \tag{34}
$$

Note that the assumption for ALS Nyström ensures that $\|(I - P_{B_*,\tilde{m}})\Sigma_{B_*}^{1/2}\|^2 \leq 3\lambda$ (Rudi et al., 2015), and hence equation 33 follows. The proof proceeds exactly as in the case of plain Nyström and is therefore omitted.

□

B.2 ADAPTIVITY: PROOF OF THEOREM 3

The following gives the proof of Theorem 3. We recall the following concentration inequality (see for example Caponnetto & Yao (2010). Let $Z_1, \dots, Z_{m'}$ be a sequence of i.i.d. real random variables with mean μ, such that $|Z_i| \le a$ a.s. and $\mathbb{E}[|Z_i - \mu|^2] \le \sigma^2$. Then for all $\alpha, \varepsilon > 0$,

$$P\left(\left|\frac{1}{m'}\sum_{i=1}^{m'} Z_i - \mu\right| \ge \varepsilon + \alpha\sigma^2\right) \le 2e^{-\frac{6m'\alpha\varepsilon}{3+4\alpha a}}. \tag{35}$$

Proof of Theorem 3. Let

$$(\tilde{d}, \tilde{\lambda}) = \operatorname*{arg\,min}_{(d,\lambda)\in\Gamma} \mathbb{E}[(T_M \hat{f}_\lambda^{\hat{B}_d}(x') - y')^2].$$

Here, the expectation is taken with respect to the pair (x', y') according to ρ.

For any $(d, \lambda) \in \Gamma$, we apply equation 35 with the choice $Z_i = Z_i^{(d,\lambda)} := \left(T_M \hat{f}_\lambda^{\hat{B}_d}(x_i') - y_i'\right)^2 - (f_*(x_i') - y_i')^2$, with $i = 1, \dots, m'$. Note that $|Z_i^{(d,\lambda)}| \le 8M^2$ and $\mathbb{E}(Z_i^{(d,\lambda)}) = \mathcal{R}(T_M \hat{f}_\lambda^{\hat{B}_d}) - \mathcal{R}(f_*) > 0$, then $\mathbb{E}((Z_i^{(d,\lambda)})^2) \le 8M^2\mathbb{E}(Z_i^{(d,\lambda)})$. Therefore, by a union bound over Γ, for all $(d, \lambda) \in \Gamma$, with probability at least $1 - \delta$, there hold

$$\frac{1}{m'}\sum_{i=1}^{m'} Z_i^{(d,\lambda)} \le (1 + 8\alpha M^2)\mathbb{E}(Z_i^{(d,\lambda)}) + \epsilon'$$

and

$$\mathbb{E}(Z_i^{(d,\lambda)}) \le \frac{1}{1 - 8\alpha M^2}\frac{1}{m'}\sum_{i=1}^{m'} Z_i^{(d,\lambda)} + \frac{\epsilon'}{1 - 8\alpha M^2}, \quad \text{for} \quad \alpha < \frac{1}{8M^2},$$

where $\epsilon' = \frac{3+32\alpha M^2}{6m'\alpha}\log\frac{2DN}{\delta}$. Therefore,

$$\begin{aligned}
&\mathcal{R}(T_M \hat{f}_\lambda^{\hat{B}_d}) - \mathcal{R}(f_*) = \mathbb{E}(Z_i^{(d,\lambda)}) \\
&\le \frac{1}{1 - 8\alpha M^2}\frac{1}{m'}\sum_{i=1}^{m'} Z_i^{(d,\lambda)} + \frac{\epsilon'}{1 - 8\alpha M^2} \\
&\le \frac{1 + 8\alpha M^2}{1 - 8\alpha M^2}\mathbb{E}(Z_i^{(d,\lambda)}) + \frac{2\epsilon'}{1 - 8\alpha M^2}.
\end{aligned}$$

Then, since $(\hat{d}, \hat{\lambda}), (\tilde{d}, \tilde{\lambda}) \in \Gamma$ and $(\hat{d}, \hat{\lambda})$ is the minimizer of $\frac{1}{m'}\sum_{i=1}^{m'}\left(T_M \hat{f}_\lambda^{\hat{B}_d}(x_i') - y_i'\right)^2$,

$$\begin{aligned}
\mathcal{R}(T_M \hat{f}_{\hat{\lambda}}^{\hat{B}_{\hat{d}}}) - \mathcal{R}(f_*) &\le \frac{1}{1 - 8\alpha M^2}\frac{1}{m'}\sum_{i=1}^{m'} Z_i^{(\hat{d},\hat{\lambda})} + \frac{\epsilon'}{1 - 8\alpha M^2} \\
&\le \frac{1}{1 - 8\alpha M^2}\frac{1}{m'}\sum_{i=1}^{m'} Z_i^{(\tilde{d},\tilde{\lambda})} + \frac{\epsilon'}{1 - 8\alpha M^2} \\
&\le \frac{1 + 8\alpha M^2}{1 - 8\alpha M^2}\mathbb{E}(Z_i^{(\tilde{d},\tilde{\lambda})}) + \frac{2\epsilon'}{1 - 8\alpha M^2}.
\end{aligned}$$

With the choice of $\alpha = 1/(24M^2)$, we get that

$$\begin{aligned}
\mathcal{R}(T_M \hat{f}_{\hat{\lambda}}^{\hat{B}_{\hat{d}}}) - \mathcal{R}(f_*) &\le 2(\mathcal{R}(T_M \hat{f}_{\tilde{\lambda}}^{\hat{B}_{\tilde{d}}}) - \mathcal{R}(f_*)) + \frac{52M^2}{m'}\log\frac{2DN}{\delta} \\
&\le 2(\mathcal{R}(T_M \hat{f}_{\tilde{\lambda}}^{\hat{B}_{d_*}}) - \mathcal{R}(f_*)) + \frac{52M^2}{m'}\log\frac{2DN}{\delta} \\
&\le 2(\mathcal{R}(\hat{f}_{\tilde{\lambda}}^{\hat{B}_{d_*}}) - \mathcal{R}(f_*)) + \frac{52M^2}{m'}\log\frac{2DN}{\delta} \\
&\le 2U(\tilde{\lambda}) + \frac{52M^2}{m'}\log\frac{2DN}{\delta},
\end{aligned}$$

where $U(\lambda)$ is the right-hand side of equation 29. The second and third steps are obtained from the definition of $(\tilde{d}, \tilde{\lambda})$ and the fact that, for any function f,

$$\mathcal{R}(T_M f) - \mathcal{R}(f_*) \leq \mathcal{R}(f) - \mathcal{R}(f_*).$$

As in (Chirinos-Rodríguez et al., 2024, Lemma 2), by the definition of the grid Λ, there exists $q \in [1, Q]$ such that $\tilde{\lambda} = q\lambda_*$, where λ_* is the optimal parameter choice according to the bound in equation 29, namely

$$\lambda_* = \arg\min_{\lambda > 0} U(\lambda) = m^{-\zeta}.$$

Furthermore, one can easily show that

$$U(q\lambda) \leq q^\theta U(\lambda)$$

for a suitable θ, so that

$$\mathcal{R}(T_M \hat{f}_{\hat{\lambda}}^{\hat{B}_d}) - \mathcal{R}(f_*) \leq 2q^\theta U(\lambda_*) + \frac{52M^2}{m'} \log \frac{2ND}{\delta}.$$

□

C TECHNICAL DETAILS ON OPTIMIZATION FOR HKRR

In this section, we provide further insights into the optimization procedure introduced in Section 4.4.

We start by presenting the complete versions of Algorithms 1 and 2, which include additional implementation specifics:

- a **non-monotone Armijo backtracking strategy** to automatically tune the learning-rates s_α and s_B. This classical approach, featuring a contraction parameter $\rho \in (0, 1)$, a dilatation parameter $\delta \in (0, 1]$ and a decay rate parameter $c > 0$, allows avoiding fine-tuning while adapting to the local behavior of $\hat{\mathcal{L}}$ thanks to its non-monotonicity (Calatroni & Chambolle, 2019).
- a **projection step** in B to handle the constraint $B \in \mathcal{B}$. At each iteration, B^i is obtained through a projected gradient step involving the projection operator $\mathcal{P}_\mathcal{B}$. In practice, this step consists in thresholding the singular values of the matrix B^i.

Variable Projection (VarPro) The complete version of the method is specified in Algorithm 3 below.

Algorithm 3 VarPro

Require B^0, $s_{B,-1} > 0$, $s_{max} > 0$, $\rho \in (0, 1)$, $\delta \in (0, 1]$ and $c > 0$.
$\alpha^0 = \arg\min_{\alpha \in \mathbb{R}^{\tilde{m}}} \hat{\mathcal{L}}\left(B^0, \alpha\right)$
for i in $0, 1, \ldots$ **do**
  $s_{B,i} = \min\left\{\frac{s_{B,i-1}}{\rho\delta}, s_{max}\right\}$
  repeat
    $s_{B,i} = \rho s_{B,i}$
    $B^{i+1} = \mathcal{P}_\mathcal{B}\left(B^i - s_{B,i}\nabla_B \hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right)$
  until $\hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right) - \hat{\mathcal{L}}\left(B^i, \alpha^i\right) < -cs_{B,i}\left\|\nabla_B \hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right\|^2$
  $\alpha^{i+1} = \arg\min_{\alpha \in \mathbb{R}^{\tilde{m}}} \hat{\mathcal{L}}\left(B^{i+1}, \alpha\right)$
end for
return $\left(B^{i+1}, \alpha^{i+1}\right)$

This method involves a linesearch strategy for determining the sequence of learning-rates $(s_{B,i})_{i \in \mathbb{N}}$. The following lemma guarantees the well-posedness of this procedure under continuity conditions on the kernel and a boundedness assumption on the sequence $(\alpha_i)_{i \in \mathbb{N}}$.

Lemma 7. *Let $k(\cdot,\cdot)$ be a continuous, twice differentiable kernel with continuous second order derivatives, and $\left(\alpha^i\right)_{i\in\mathbb{N}}$ be a bounded sequence. Then, there exists $L_1 > 0$ such that,*

$$\forall B \in \mathcal{B},\ \forall i \in \mathbb{N}, \quad \left\|\nabla_B^2 \hat{\mathcal{L}}\left(B, \alpha^i\right)\right\|_2 \leq L_1, \tag{36}$$

implying that for any $i \in \mathbb{N}$, $B \mapsto \hat{\mathcal{L}}\left(B, \alpha^i\right)$ is L_1-smooth. It follows that the linesearch procedure used to generate $(s_{B,i})_{i\in\mathbb{N}}$ in Algorithm 3 and Algorithm 4 is well defined, and $s_{B,i} \geq \min\left\{\frac{2\rho(1-c)}{L_1}, s_{B,-1}\right\}$ for any $i \in \mathbb{N}$.

Proof of Lemma 7. Since we consider a dataset of dimension $m < +\infty$, we can ensure that there exists $C < +\infty$ such that for any $j \in \{1,\dots,m\}$, $\|x_j\| < C$. As a consequence, for any $B \in \mathcal{B}$,

$$\|Bx_j\| \leq \|B\|_\infty \|x_j\| \leq C.$$

From the above inequality and the continuity of k and its first and second order derivatives, we get that $k(Bx_{j_1}, Bx_{j_2})$, $\|k_p(Bx_{j_1}, Bx_{j_2})\|$ and $\|k_{pq}(Bx_{j_1}, Bx_{j_2})\|$, where $(p,q) \in \{1,2\}^2$, can be bounded independently from B and $(j_1, j_2) \in \{1,\dots,m\}^2$.

It is then straightforward to show that $\left\|\nabla_B^2 \hat{\mathcal{L}}\left(B, \alpha^i\right)\right\|_2$ can be bounded independently from B and i since it only depends on $(k(Bx_{j_1}, Bx_{j_2}))_{j_1,j_2=1}^m$, $(k_p(Bx_{j_1}, Bx_{j_2}))_{j_1,j_2=1}^m$, $(k_{pq}(Bx_{j_1}, Bx_{j_2}))_{j_1,j_2=1}^m$, $(y_j)_{j=1}^m$ and α^i which can be bounded independently from i by assumption.

The Lipschitz continuity of $B \mapsto \hat{\mathcal{L}}\left(B, \alpha^i\right)$ directly ensures that the Armijo backtracking procedure in Algorithm 4 is well defined. In particular, if B^i is not a critical point of $\Psi : B \mapsto \hat{\mathcal{L}}\left(B, \alpha^i\right) + i_{\mathcal{B}}(B)$, the Armijo condition is satisfied for any step size $s \leq \frac{2(1-c)}{L_1}$. Since the $s_{B,i}$ is updated by multiplying it by ρ until the condition is satisfied, we can deduce the desired lower bound. □

We can then exploit this lemma to prove the convergence of Algorithm 3, using the Kurdyka-Łojasiewicz property of $\hat{L}$.

Theorem 8. *Let $k(\cdot,\cdot)$ be an analytic kernel. Let $\left(B^i\right)_{i\in\mathbb{N}}$ and $\left(\alpha^i\right)_{i\in\mathbb{N}}$ be the sequences generated by Algorithm 3 and suppose that for any $i \in \mathbb{N}$, $\lambda_{min}\left(\hat{K}_{\tilde{m}\tilde{m}}^{B^i}\right) \geq \sigma > 0$ where λ_{min} denotes the smallest eigenvalue. Then, **the sequence** $\left(B^i, \alpha^i\right)_{i\in\mathbb{N}}$ **converges to a critical point** of $\Psi : B, \alpha \mapsto \mathcal{L}(B,\alpha) + i_{\mathcal{B}}(B)$ as i goes to infinity. In addition, the sequences $\left(B^i\right)_{i\in\mathbb{N}}$ and $\left(\alpha^i\right)_{i\in\mathbb{N}}$ have finite length, i.e.*

$$\sum_{i=0}^{+\infty} \|B^{i+1} - B^i\| < +\infty, \quad \sum_{i=0}^{+\infty} \|\alpha^{i+1} - \alpha^i\| < +\infty,$$

and there exists $C > 0$ such that after N iterations, $\left(B^N, \alpha^N\right)$ is a critical point of Ψ or

$$\min_{0\leq i\leq N} \left\|\nabla_B \hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right\|^2 \leq \frac{C}{N}. \tag{37}$$

Proof of Theorem 8. The proof of this theorem relies on (Attouch et al., 2013, Theorem 2.9) stating convergence towards critical points for algorithms minimizing functions having the Kurdyka-Łojasiewicz property under several assumptions. It requires to prove the following points:

(a) the function $\Psi : B, \alpha \mapsto \mathcal{L}(B,\alpha) + i_{\mathcal{B}}(B)$ has the Kurdyka-Łojasiewicz (KL) property,

(b) there exists $a > 0$ such that for each $i \in \mathbb{N}$,

$$\Psi\left(B^{i+1}, \alpha^{i+1}\right) + a\left(\left\|B^{i+1} - B^i\right\|^2 + \left\|\alpha^{i+1} - \alpha^i\right\|^2\right) \leq \Psi\left(B^i, \alpha^i\right), \tag{38}$$

(c) there exists $b > 0$ such that for each $i \in \mathbb{N}$, there is $g^{i+1} \in \partial\Psi\left(B^{i+1}, \alpha^{i+1}\right)$ satisfying,

$$\left\|g^{i+1}\right\|^2 \leq b\left(\left\|B^{i+1} - B^i\right\|^2 + \left\|\alpha^{i+1} - \alpha^i\right\|^2\right), \tag{39}$$

where $\partial\Psi$ denotes the convex subdifferential of Ψ which is defined for any $(B,\alpha) \in \mathbb{R}^{d\times D} \times \mathbb{R}^{\tilde{m}}$ as

$$\partial\Psi(B,\alpha) = \left\{s \in \mathbb{R}^{d\times D} \times \mathbb{R}^{\tilde{m}} \mid \forall\left(B',\alpha'\right) \in \mathbb{R}^{d\times D} \times \mathbb{R}^{\tilde{m}}, \Psi\left(B',\alpha'\right) \geq \Psi\left(B,\alpha\right) + \langle s, \left(B',\alpha'\right) - (B,\alpha)\rangle\right\}.$$

(d) the sequence $\left(B^i, \alpha^i\right)_{i\in\mathbb{N}}$ admits a converging subsequence.

Before proving each point above, we first show that the sequence $\left(\alpha^i\right)_{i\in\mathbb{N}}$ in Algorithm 3 is well defined and bounded. According to the assumptions of the theorem, there exists $\sigma > 0$ such that for any $i \in \mathbb{N}$, we have

$$\lambda_{min}\left(\hat{K}^{B^i}_{\tilde{m}\tilde{m}}\right) \geq \sigma.$$

This ensures that the matrix $\left(\frac{1}{m}(\hat{K}^{B^i}_{m\tilde{m}})^T\hat{K}^{B^i}_{m\tilde{m}} + \lambda\hat{K}^{B^i}_{\tilde{m}\tilde{m}}\right)$ is invertible at each iteration i and thus that α^i is well defined. In addition, we have that for any $i \in \mathbb{N}$,

$$\|\alpha^i\| \leq \left\|\left(\frac{1}{m}(\hat{K}^{B^i}_{m\tilde{m}})^T\hat{K}^{B^i}_{m\tilde{m}} + \lambda\hat{K}^{B^i}_{\tilde{m}\tilde{m}}\right)^{-1}\right\|_2 \left\|\hat{K}^{B^i}_{\tilde{m}\tilde{m}}\right\|_2 \|y\| \leq \frac{\left\|\hat{K}^{B^i}_{\tilde{m}\tilde{m}}\right\|_2 \|y\|}{\lambda\sigma},$$

and since k is continuous (because analytic), $B^i \in \mathcal{B}$ and the dataset is bounded, we can conclude that $\left\|\hat{K}^{B^i}_{\tilde{m}\tilde{m}}\right\|_2$, and consequently the sequence $\left(\alpha^i\right)_{i\in\mathbb{N}}$, are bounded. In addition, k is analytic and therefore continuous, twice differentiable with continuous second-order derivatives. Consequently, we can apply Lemma 7.
We now prove the four properties stated at the beginning of the proof:
1. The analyticity of k directly ensures that $\hat{\mathcal{L}}$ is analytic (and therefore has the KL property) in both variables. Moreover, the indicator function $i_{\mathcal{B}}$ where $\mathcal{B} = \{B \in \mathbb{R}^{d\times D}, \|B\|_\infty \leqslant 1\}$ is semi algebraic and also has the KL property. As a consequence, the first statement is directly satisfied.
2. First, notice that since $B^0 \in \mathcal{B}$ and the sequence $\left(B^i\right)_{i\in\mathbb{N}}$ is built via the step

$$B^{i+1} = \mathcal{P}_{\mathcal{B}}\left(B^i - s_{B,i}\nabla_B\hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right),$$

where $s_{B,i} > 0$, we have that $B^i \in \mathcal{B}$ for any $i \in \mathbb{N}$. It follows that for any $i \in \mathbb{N}$ and $\alpha \in \mathbb{R}^{\tilde{m}}$,

$$\Psi\left(B^i, \alpha\right) = \mathcal{L}\left(B^i, \alpha\right). \tag{40}$$

Lemma 7 guarantees that for each $i \in \mathbb{N}$, Algorithm 3 provides $s_{B,i}$ such that

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right) - \hat{\mathcal{L}}\left(B^i, \alpha^i\right) < -cs_{B,i}\left\|\nabla_B\hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right\|^2.$$

Note that due to the firm non-expansiveness of $\mathcal{P}_{\mathcal{B}}$, we have that for any $i \in \mathbb{N}$,

$$\left\|B^{i+1} - B^i\right\| = \left\|\mathcal{P}_{\mathcal{B}}\left(B^i - s_{B,i}\nabla_B\hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right) - \mathcal{P}_{\mathcal{B}}\left(B^i\right)\right\| \leq s_{B,i}\left\|\nabla_B\hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right\|.$$

Consequently, at any iteration $i \in \mathbb{N}$,

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right) - \hat{\mathcal{L}}\left(B^i, \alpha^i\right) < -\frac{c}{s_{B,i}}\left\|B^{i+1} - B^i\right\|^2 \leq -\frac{c}{s_{max}}\left\|B^{i+1} - B^i\right\|^2. \tag{41}$$

Since it is assumed that there exists some $\sigma > 0$ such that for any $i \in \mathbb{N}$ we have $\lambda_{min}\left(\hat{K}^{B^i}_{\tilde{m}\tilde{m}}\right) \geq \sigma$, we can prove that $\alpha \mapsto \hat{\mathcal{L}}\left(B^i, \alpha\right)$ is $\lambda\sigma$-strongly convex for any $i \in \mathbb{N}$. From the definition of $\left(\alpha^i\right)_{i\in\mathbb{N}}$ and this property, we get that

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i+1}\right) + \frac{\lambda\sigma}{2}\left\|\alpha^{i+1} - \alpha^i\right\|^2 \leq \hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right).$$

We can deduce that

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i+1}\right) - \hat{\mathcal{L}}\left(B^i, \alpha^i\right) < -\frac{c}{s_{max}}\left\|B^{i+1} - B^i\right\|^2 - \frac{\lambda\sigma}{2}\left\|\alpha^{i+1} - \alpha^i\right\|^2, \tag{42}$$

which leads to the desired inequality.
3. We aim at showing that for a well-chosen $b > 0$, for any $i \in \mathbb{N}$, there exists $g^{i+1} \in \partial\Psi\left(B^{i+1}, \alpha^{i+1}\right)$ (i.e. $g^{i+1} = \left(g_B^{i+1}, g_\alpha^{i+1}\right)$ where $g_B^{i+1} \in \partial_B\Psi\left(B^{i+1}, \alpha^{i+1}\right)$ and $g_\alpha^{i+1} \in \partial_\alpha\Psi\left(B^{i+1}, \alpha^{i+1}\right)$) such that

$$\left\|g^{i+1}\right\|^2 \leq b\left(\left\|B^{i+1} - B^i\right\|^2 + \left\|\alpha^{i+1} - \alpha^i\right\|^2\right).$$

Due to the structure of Ψ, it is then sufficient to show that for some $v^{i+1} \in \partial i_{\mathcal{B}}(B^{i+1})$, the choice $g^{i+1} = \left(v^{i+1} + \nabla_B \mathcal{L}(B^{i+1}, \alpha^{i+1}), \nabla_\alpha \mathcal{L}(B^{i+1}, \alpha^{i+1})\right)$ is valid for the above equation.
From Algorithm 4, the sequence $\left(B^i\right)_{i\in\mathbb{N}}$ is defined via a step which can be seen as a proximal gradient step on $B \mapsto \mathcal{L}(B, \alpha^i) + i_{\mathcal{B}}(B)$. As a consequence, we can write that

$$B^i - B^{i+1} - s_{B,i}\nabla_B \mathcal{L}(B^i, \alpha^i) \in \partial i_{\mathcal{B}}(B^{i+1}),$$

and we will choose $v^{i+1} = \frac{1}{s_{B,i}}\left(B^i - B^{i+1}\right) - \nabla_B \mathcal{L}(B^i, \alpha^i) \in \partial i_{\mathcal{B}}(B^{i+1})$ (due to the properties of $\partial i_{\mathcal{B}}$ which is the normal cone onto $\mathcal{B}$). It follows that

$$\begin{aligned}\left\|g^{i+1}\right\|^2 &= \left\|v^{i+1} + \nabla_B \mathcal{L}(B^{i+1}, \alpha^{i+1})\right\|^2 + \left\|\nabla_\alpha \mathcal{L}(B^{i+1}, \alpha^{i+1})\right\|^2 \\ &= \left\|\frac{1}{s_{B,i}}\left(B^i - B^{i+1}\right) + \nabla_B \mathcal{L}(B^{i+1}, \alpha^{i+1}) - \nabla_B \mathcal{L}(B^i, \alpha^i)\right\|^2 + \left\|\nabla_\alpha \mathcal{L}(B^{i+1}, \alpha^{i+1})\right\|^2. \end{aligned} \tag{43}$$

Elementary computations ensure that

$$\begin{aligned}\left\|\frac{1}{s_{B,i}}\left(B^i - B^{i+1}\right) + \nabla_B \mathcal{L}(B^{i+1}, \alpha^{i+1}) - \nabla_B \mathcal{L}(B^i, \alpha^i)\right\|^2 \leq\ & \frac{2}{s_{B,i}}\left\|B^{i+1} - B^i\right\|^2 \\ & + 2\left\|\nabla_B \mathcal{L}(B^{i+1}, \alpha^{i+1}) - \nabla_B \mathcal{L}(B^i, \alpha^i)\right\|^2.\end{aligned}$$

By using similar arguments to that in the proof of Lemma 7 namely boundedness of $\left(B^i\right)_{i\in\mathbb{N}}$ and $\left(\alpha^i\right)_{i\in\mathbb{N}}$, and continuity of the second order derivatives of k, we can show that $\hat{\mathcal{L}}$ is jointly Lipschitz smooth in (B, α) on a compact containing $\left(B^i, \alpha^i\right)_{i\in\mathbb{N}}$. We can deduce that there exists $L > 0$ such that for any (B_1, α_1) and (B_2, α_2):

$$\|\nabla_B \mathcal{L}(B_1, \alpha_1) - \nabla_B \mathcal{L}(B_2, \alpha_2)\|^2 + \|\nabla_\alpha \mathcal{L}(B_1, \alpha_1) - \nabla_\alpha \mathcal{L}(B_2, \alpha_2)\|^2 \leq L\left(\|B_1 - B_2\|^2 + \|\alpha_1 - \alpha_2\|^2\right).$$

We can then use the above inequality and the lower bound on $s_{B,i}$ from Lemma 7 to write

$$\begin{aligned}\left\|\frac{1}{s_{B,i}}\left(B^i - B^{i+1}\right) + \nabla_B \mathcal{L}(B^{i+1}, \alpha^{i+1}) - \nabla_B \mathcal{L}(B^i, \alpha^i)\right\|^2 \leq\ & 2\left(\max\left\{\frac{L_1}{2\rho(1-c)}, s_{B,-1}^{-1}\right\} + L\right)\left\|B^{i+1} - B^i\right\|^2 \\ & + 2L\left\|\alpha^{i+1} - \alpha^i\right\|^2.\end{aligned}$$

In addition, since α^{i+1} minimizes the function $\alpha \mapsto \hat{\mathcal{L}}\left(B^{i+1}, \alpha\right)$, we directly get that

$$\left\|\nabla_\alpha \mathcal{L}(B^{i+1}, \alpha^{i+1})\right\| = 0.$$

Therefore, we get that

$$\left\|g^{i+1}\right\|^2 \leq 2\left(\max\left\{\frac{L_1}{2\rho(1-c)}, s_{B,-1}^{-1}\right\} + L\right)\left\|B^{i+1} - B^i\right\|^2 + 2L\left\|\alpha^{i+1} - \alpha^i\right\|^2,$$

which implies equation 39.
4. This point is trivially satisfied as both $\left(B^i\right)_{i\in\mathbb{N}}$ and $\left(\alpha^i\right)_{i\in\mathbb{N}}$ are bounded.
We have proved the convergence of Algorithm 3 towards a critical point of Ψ. We now demonstrate that equation 37 holds after N iterations if $\left(B^N, \alpha^N\right)$ is not a critical point. From the definition of the method, for any $i \in \{0, \ldots, N\}$,

$$\begin{aligned}\left\|\nabla_B \hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right\|^2 &< \frac{1}{cs_{B,i}}\left(\hat{\mathcal{L}}\left(B^i, \alpha^i\right) - \hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right)\right) \\ &< c^{-1}\max\left\{\frac{L_1}{2\rho(1-c)}, s_{B,-1}^{-1}\right\}\left(\hat{\mathcal{L}}\left(B^i, \alpha^i\right) - \hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right)\right).\end{aligned}$$

By summing this inequality on $i \in \{0, \ldots, N\}$, we get that

$$\begin{aligned}\sum_{i=0}^{N}\left\|\nabla_B \hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right\|^2 &< c^{-1}\max\left\{\frac{L_1}{2\rho(1-c)}, s_{B,-1}^{-1}\right\}\sum_{i=0}^{N}\left(\hat{\mathcal{L}}\left(B^i, \alpha^i\right) - \hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right)\right) \\ &\leq c^{-1}\max\left\{\frac{L_1}{2\rho(1-c)}, s_{B,-1}^{-1}\right\}\hat{\mathcal{L}}\left(B^0, \alpha^0\right) := C.\end{aligned}$$

We can then deduce equation 37. □

Alternating Gradient Descent. The complete version of the scheme can be found below.

Algorithm 4 Alternating Gradient Descent

Require $B^0 \in \mathcal{B}$, $\alpha^0 \in \mathbb{R}^{\tilde{m}}$, $s_{B,-1} > 0$, $s_{\alpha,-1} > 0$, $s_{max} > 0$, $\rho \in (0,1)$, $\delta \in (0,1]$, $c > 0$ and $n_\alpha \in \mathbb{N}^*$.
for i in $0, 1, ...$ **do**
  $s_{B,i} = \min\left\{\frac{s_{B,i-1}}{\rho\delta}, s_{max}\right\}$
  repeat
    $s_{B,i} = \rho s_{B,i}$
    $B^{i+1} = \mathcal{P}_{\mathcal{B}}\left(B^i - s_{B,i}\nabla_B\hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right)$
  until $\hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right) - \hat{\mathcal{L}}\left(B^i, \alpha^i\right) < -cs_{B,i}\left\|\nabla_B\hat{\mathcal{L}}\left(B^i, \alpha^i\right)\right\|^2$
  $\alpha^{i,0} = \alpha^i$
  $s_{\alpha,i,-1} = s_{\alpha,i-1}$
  for j in $0, 1, ..., n_\alpha - 1$ **do**
    $s_{\alpha,i,j} = \min\left\{\frac{s_{\alpha,i,j-1}}{\rho\delta}, s_{max}\right\}$
    repeat
      $s_{\alpha,i,j} = \rho s_{\alpha,i,j}$
      $\alpha^{i,j+1} = \alpha^{i,j} - s_{\alpha,i,j}\nabla_\alpha\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i,j}\right)$
    until $\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i,j+1}\right) - \hat{\mathcal{L}}\left(B^i, \alpha^{i,j}\right) < -cs_{\alpha,i,j}\left\|\nabla_\alpha\hat{\mathcal{L}}\left(B^i, \alpha^{i,j}\right)\right\|^2$
  end for
  $s_{\alpha,i} = s_{\alpha,i,n_\alpha-1}$
  $\alpha^{i+1} = \alpha^{i,n_\alpha-1}$
end for
return $\left(B^{i+1}, \alpha^{i+1}\right)$

Similarly to Algorithm 3, Algorithm 4 leverages a backtracking strategy to set both $(s_{B,i})_{i\in\mathbb{N}}$ and $(s_{\alpha,i,j})_{(i,j)\in\mathbb{N}\times\{0,\dots,n_\alpha-1\}}$. In addition to Lemma 7, we introduce the following lemma guaranteeing that these sequences are well-defined.

Lemma 8. *Let $k(\cdot,\cdot)$ be a continuous kernel and $\left(B^i\right)_{i\in\mathbb{N}}$ such that $B^i \in \mathcal{B}$ for any $i \in \mathbb{N}$. Then, there exists $L_2 > 0$ such that,*

$$\forall\alpha \in \mathbb{R}^{\tilde{m}},\ \forall i \in \mathbb{N}, \quad \left\|\nabla^2_\alpha\hat{\mathcal{L}}\left(B^i, \alpha\right)\right\|_2 \leq L_2, \tag{44}$$

implying that for any $i \in \mathbb{N}$, $\alpha \mapsto \hat{\mathcal{L}}\left(B^i, \alpha\right)$ is L_2-smooth.
More precisely, $\left\|\nabla^2_\alpha\hat{\mathcal{L}}\left(B^i, \alpha\right)\right\|_2 = \left\|\frac{1}{m}(\hat{K}^{B^i}_{m\tilde{m}})^T\hat{K}^{B^i}_{m\tilde{m}} + \lambda\hat{K}^{B^i}_{\tilde{m}\tilde{m}}\right\|_2 \leq L_2$. It follows that the linesearch procedure used to generate $(s_{\alpha,i,j})_{i,j\in\mathbb{N}\times\{1,\dots,n_\alpha\}}$ in Algorithm 4 is well defined, and $s_{\alpha,i,j} \geq \min\left\{\frac{2\rho(1-c)}{L_2}, s_{\alpha,-1}\right\}$ for any $i \in \mathbb{N}$ and $j \in \{0,\dots,n_\alpha - 1\}$.

Proof of Lemma 8. Elementary computations show that for any $\alpha \in \mathbb{R}^{\tilde{m}}$ and $B \in \mathcal{B}$,

$$\nabla^2_\alpha\hat{\mathcal{L}}\left(B, \alpha\right) = \frac{1}{m}(\hat{K}^B_{m\tilde{m}})^T\hat{K}^B_{m\tilde{m}} + \lambda\hat{K}^B_{\tilde{m}\tilde{m}}$$

Notice that $\hat{K}^B_{m\tilde{m}}$ and $\hat{K}^B_{\tilde{m}\tilde{m}}$ are submatrices of $\hat{K}^B$ defined as $\left(\hat{K}^B\right)_{i,j} = k\left(Bx_i, Bx_j\right)$. Since $\hat{K}^B$ is a positive semi-definite matrix, we have that $\|\hat{K}^B_{m\tilde{m}}\|_2 \leq \|\hat{K}^B\|_2$ and $\|\hat{K}^B_{\tilde{m}\tilde{m}}\|_2 \leq \|\hat{K}^B\|_2$.
We can then use the continuity of k, the boundedness of $(x_i)_{i=1}^m$ and the inequality $\|B\|_\infty \leq 1$ (see more details in the proof of Lemma 7) to show that there exists $C < +\infty$ such that $\|\hat{K}^B\|_2 \leq C$. This leads to equation 44 with $L_2 = \frac{C^2}{m} + \lambda C$.
Since $B^i \in \mathcal{B}$ for any $i \in \mathbb{N}$, the Lipschitz continuity of $\alpha \mapsto \hat{\mathcal{L}}\left(B^i, \alpha\right)$ directly ensures that the Armijo backtracking procedure in Algorithm 4 is well defined. In particular, the Armijo condition is satisfied for any step size

$s \leq \frac{2(1-c)}{L_2}$. Since the $s_{\alpha,i,j}$ is updated by multiplying it by ρ until the condition is satisfied, we can deduce the desired lower bound. □

Remark 10 (Backtracking strategy on $(s_{\alpha,i,j})_{(i,j)\in\mathbb{N}\times\{0,\dots,n_\alpha-1\}}$)**.** *For any iteration $i \in \mathbb{N}$, the function $\alpha \mapsto \hat{\mathcal{L}}\left(B^i, \alpha\right)$ is L_i-Lipschitz where $L_i = \left\| \frac{1}{m}(\hat{K}^{B^i}_{m\tilde{m}})^T \hat{K}^{B^i}_{m\tilde{m}} + \lambda \hat{K}^{B^i}_{\tilde{m}\tilde{m}} \right\|_2$ can be computed directly. It is therefore possible to replace the linesearch procedure for setting $(s_{\alpha,i,j})_{(i,j)\in\mathbb{N}\times\{0,\dots,n_\alpha-1\}}$ by the rule $s_{\alpha,i,j} = 1/L_i$ in Algorithm 4 (the convergence guarantees would be the same). It is worth noticing that this strategy involves computing the largest eigenvalue of $\frac{1}{m}(\hat{K}^{B^i}_{m\tilde{m}})^T \hat{K}^{B^i}_{m\tilde{m}} + \lambda \hat{K}^{B^i}_{\tilde{m}\tilde{m}}$ which can be costly depending on the parameter $\tilde{m}$.*

By applying the same strategy as that of the proof of Theorem 8, we demonstrate similar convergence properties of Algorithm 4.

Theorem 9. *Let $k(\cdot,\cdot)$ be an analytic kernel. Let $\left(B^i\right)_{i\in\mathbb{N}}$ and $\left(\alpha^i\right)_{i\in\mathbb{N}}$ be the sequences generated by Algorithm 4 and suppose that $\left(\alpha^i\right)_{i\in\mathbb{N}}$ is bounded. Then,* ***the sequence*** *$\left(B^i, \alpha^i\right)_{i\in\mathbb{N}}$* ***converges to a critical point*** *of $\Psi : B, \alpha \mapsto \mathcal{L}(B,\alpha) + i_{\mathcal{B}}(B)$ as i goes to infinity. In addition, the sequences $\left(B^i\right)_{i\in\mathbb{N}}$ and $\left(\alpha^i\right)_{i\in\mathbb{N}}$ have finite length, i.e.*

$$\sum_{i=0}^{+\infty} \|B^{i+1} - B^i\| < +\infty, \quad \sum_{i=0}^{+\infty} \|\alpha^{i+1} - \alpha^i\| < +\infty,$$

and there exists $C > 0$ such that after N iterations, $\left(B^N, \alpha^N\right)$ is a critical point of Ψ or

$$\min_{0\leq i\leq N} \left\| \nabla_B \hat{\mathcal{L}}\left(B^i, \alpha^i\right) \right\|^2 \leq \frac{C}{N}. \tag{45}$$

Proof of Theorem 9. Similarly to the proof of Theorem 8, we adapt (Attouch et al., 2013, Theorem 2.9) to our framework to show the desired convergence results. Therefore, we need to show the 4 assertions enumerated in the aforementioned proof.

1. The Kurdyka-Łojasiewicz property of $\hat{\mathcal{L}}$ is already shown in the proof of Theorem 8.
2. As stated in the proof of Theorem 8, it is trivial that $B^i \in \mathcal{B}$ for any $i \in \mathbb{N}$, and consequently for any $\alpha \in \mathbb{R}^{\tilde{m}}$,

$$\Psi\left(B^i, \alpha\right) = \mathcal{L}\left(B^i, \alpha\right). \tag{46}$$

In addition, similar computations allow to show that

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right) - \hat{\mathcal{L}}\left(B^i, \alpha^i\right) < -\frac{c}{s_{max}} \left\|B^{i+1} - B^i\right\|^2. \tag{47}$$

Lemma 8 ensures that at each step $i \in \mathbb{N}$ and substep $j \in \{1, \dots, n_\alpha\}$,

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i,j+1}\right) - \hat{\mathcal{L}}\left(B^i, \alpha^{i,j}\right) < -cs_{\alpha,i,j} \left\| \nabla_\alpha \hat{\mathcal{L}}\left(B^i, \alpha^{i,j}\right) \right\|^2, \tag{48}$$

which directly implies that

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i,j+1}\right) - \hat{\mathcal{L}}\left(B^i, \alpha^{i,j}\right) < -\frac{c}{s_{\alpha,i,j}} \left\|\alpha^{i,j+1} - \alpha^{i,j}\right\|^2 \leq -\frac{c}{s_{max}} \left\|\alpha^{i,j+1} - \alpha^{i,j}\right\|^2.$$

Since $\alpha^i = \alpha^{i,0}$ and $\alpha^{i+1} = \alpha^{i,n_\alpha-1}$, we get that

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i+1}\right) - \hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right) < -\frac{c}{s_{max}} \sum_{j=0}^{n_\alpha-2} \left\|\alpha^{i,j+1} - \alpha^{i,j}\right\|^2,$$

and since for any sequence $(x^i)_{i\in\mathbb{N}}$, $\left\|\sum_{i=1}^n x^i\right\|^2 \leq n \sum_{i=1}^n \left\|x^i\right\|^2$,

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i+1}\right) - \hat{\mathcal{L}}\left(B^{i+1}, \alpha^i\right) < -\frac{c}{s_{max}(n_\alpha - 1)} \left\|\alpha^{i+1} - \alpha^i\right\|^2. \tag{49}$$

From equation 48 and equation 49, we can prove that

$$\hat{\mathcal{L}}\left(B^{i+1}, \alpha^{i+1}\right) - \hat{\mathcal{L}}\left(B^i, \alpha^i\right) < -\frac{c}{s_{max}} \left(\left\|B^{i+1} - B^i\right\|^2 - \frac{1}{n_\alpha - 1} \left\|\alpha^{i+1} + \alpha^i\right\|^2 \right),$$

which leads to the desired conclusion, taking $a = \frac{c}{s_{max}(n_\alpha - 1)}$ and using equation 46.

3. We aim at showing that for a well-chosen $b > 0$, for any $i \in \mathbb{N}$, there exists $g^{i+1} \in \partial\Psi\left(B^{i+1}, \alpha^{i+1}\right)$ (i.e. $g^{i+1} = \left(g_B^{i+1}, g_\alpha^{i+1}\right)$ where $g_B^{i+1} \in \partial_B\Psi\left(B^{i+1}, \alpha^{i+1}\right)$ and $g_\alpha^{i+1} \in \partial_\alpha\Psi\left(B^{i+1}, \alpha^{i+1}\right)$) such that

$$\left\|g^{i+1}\right\|^2 \leq b\left(\left\|B^{i+1} - B^i\right\|^2 + \left\|\alpha^{i+1} - \alpha^i\right\|^2\right).$$

By taking $g^{i+1} = \left(v^{i+1} + \nabla_B\mathcal{L}(B^{i+1}, \alpha^{i+1}), \nabla_\alpha\mathcal{L}(B^{i+1}, \alpha^{i+1})\right)$ with $v^{i+1} = \frac{1}{s_{B,i}}\left(B^i - B^{i+1}\right) - \nabla_B\mathcal{L}(B^i, \alpha^i)$, we can apply the same reasoning as that of the proof of Theorem 8 to demonstrate that there exists $L > 0$ such that:

$$\begin{aligned}\left\|g^{i+1}\right\|^2 \leq 2\left(\max\left\{\frac{L_1}{2\rho(1-c)}, s_{B,-1}^{-1}\right\} + L\right)\left\|B^{i+1} - B^i\right\|^2 + 2L\left\|\alpha^{i+1} - \alpha^i\right\|^2 \\ + \left\|\nabla_\alpha\mathcal{L}(B^{i+1}, \alpha^{i+1})\right\|^2.\end{aligned} \tag{50}$$

By rewriting the second term of the above inequality,

$$\left\|\nabla_\alpha\mathcal{L}(B^{i+1}, \alpha^{i+1})\right\|^2 \leq \left\|\nabla_\alpha\mathcal{L}(B^{i+1}, \alpha^{i+1}) - \nabla_\alpha\mathcal{L}(B^{i+1}, \alpha^{i,n_{\alpha-1}})\right\|^2 + \left\|\nabla_\alpha\mathcal{L}(B^{i+1}, \alpha^{i,n_{\alpha-1}})\right\|^2.$$

From the joint Lipschitz smoothness of $\hat{\mathcal{L}}$ and the definition of the sequence $\left(\alpha^i\right)_{i\in\mathbb{N}}$, we get that

$$\begin{aligned}\left\|\nabla_\alpha\mathcal{L}(B^{i+1}, \alpha^{i+1})\right\|^2 &\leq 2\left(L + \frac{1}{s_{\alpha,i,n_\alpha - 1}}\right)\left\|\alpha^{i+1} - \alpha^i\right\|^2 \\ &\leq 2\left(L + \max\left\{\frac{L_2}{2\rho(1-c)}, s_{\alpha,-1}^{-1}\right\}\right)\left\|\alpha^{i+1} - \alpha^i\right\|^2,\end{aligned}$$

where we use the lower bound on $s_{\alpha,i,j}$ from Lemma 8. Combining the above inequalities, we can conclude that

$$\begin{aligned}\left\|g^{i+1}\right\|^2 \leq 2\left(L + \max\left\{\frac{L_1}{2\rho(1-c)}, s_{B,-1}^{-1}\right\}\right)\left\|B^{i+1} - B^i\right\|^2 \\ + 2\left(2L + \max\left\{\frac{L_2}{2\rho(1-c)}, s_{\alpha,-1}^{-1}\right\}\right)\left\|\alpha^{i+1} - \alpha^i\right\|^2\end{aligned}$$

which implies the desired inequality for

$$b = \max\left\{2\left(L + \max\left\{\frac{L_1}{2\rho(1-c)}, s_{B,-1}^{-1}\right\}\right), 2\left(2L + \max\left\{\frac{L_2}{2\rho(1-c)}, s_{\alpha,-1}^{-1}\right\}\right)\right\}.$$

4. This point is trivially satisfied as both $\left(B^i\right)_{i\in\mathbb{N}}$ and $\left(\alpha^i\right)_{i\in\mathbb{N}}$ are bounded.

We demonstrated above that the method converges to a critical point. Inequality (45) can be obtained using the same computations as in the proof of Theorem 8. □

Remark 11 (On the boundedness of $\left(\alpha^i\right)_{i\in\mathbb{N}}$)**.** *Theorem 9 relies on a boundedness assumption on the sequence $\left(\alpha^i\right)_{i\in\mathbb{N}}$. This can be enforced in different ways:*

1. *by applying the same hypothesis as in Theorem 8: suppose that for any $i \in \mathbb{N}$, $\lambda_{min}\left(\hat{K}_{\tilde{m}\tilde{m}}^{B^i}\right) \geq \sigma > 0$. The boundedness of the sequence can be obtained directly as Algorithm 4 is a descent method and consequently the term $\lambda(\alpha^i)^T\hat{K}_{\tilde{m}\tilde{m}}^{B^i}\alpha^i$ can not grow indefinitely.*
2. *by directly adding a constraint on α to the problem. The convergence analysis for Algorithm 4 would remain the same.*

Remark 12 (On δ and s_{max})**.** *Algorithm 3 and Algorithm 4 involve a non-monotone backtracking procedure for defining the learning rates. The non-monotonicity, which occurs when $\delta < 1$, allows for more aggressive step sizes that better adapt to the local geometry of the objective function. For technical reasons, the proofs require to set a maximum learning-rate $s_{max} > 0$ which, in practice, is not necessary. Note that in the monotone case, i.e. $\delta = 1$, the sequences of learning-rates are directly bounded by the initial value.*

D SUPPLEMENTAL MATERIAL ON NUMERICAL EXPERIMENTS

D.1 EXPERIMENTS SETTING

Experimental setting. The experiments presented in Figure 1a, Figure 1b, Figure 1c and Appendix D.2 were performed in Python on a 2,4 GHz Intel Core i5 quad-core laptop with 8 Gb of RAM. The remaining experiments were performed on a server on #99-Ubuntu SMP with 2 x AMD EPYC 7301 16-Core Processor and 256 Gb of RAM.

Datasets. Two synthetic datasets are generated for the experiments presented in the paper. The **first dataset** is generated by sampling $x_i \sim \mathcal{U}([-1,1])$ and setting the output as follows:

$$y_i = \underbrace{\sum_{j=1}^{d} \sin\left(\left(1+\frac{j}{d}\right)\pi\left(Bx_i\right)_j\right)}_{:=z_i} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0,\sigma^2\right), \tag{51}$$

where σ^2 depends on the variance of $(z_i)_{i\in\{1,\dots,m\}}$ (i.e. $\sigma^2 = \text{Var}(z)/100$). The matrix B is sampled randomly in $\mathcal{B}$ (for any $(i,j) \in \{1,\dots,d\}\times\{1,\dots,D\}$, $B_{i,j} \sim \mathcal{U}([0,1])$).
The **second dataset** is generated taking $x_i \in \mathcal{U}([-10,10])$ and

$$y_i = \underbrace{\sum_{j=1}^{d} \sin\left(0.5\left(Bx_i\right)_j - j\right) + \frac{1}{2}\left(Bx_i\right)_{j+1}\cos\left(0.4\left(Bx_i\right)_{j+2} - j + 1\right)}_{:=z_i} + \varepsilon_i, \tag{52}$$

where $\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$ and $\sigma^2 = \text{Var}(z)/100$. We sample B randomly as done for the first dataset.

Methodology. The convergence graphs of Figure 1a were obtained by running VarPro and AGD on the first dataset introduced above, setting $D = 300$ and $d = 2$, and the true value of B manually to $B_{i,j} = 1$ for $(i,j) \in \{(1,1),(2,2)\}$, i.e. selecting only the first two components. We used $\tilde{m} = 5$ Nyström centers from the training set of size $m = 300$.

The experiments presented at the top of Figure 2 were performed using VarPro and AGD on 5 Nyström centers for $\lambda = 10^{-7}$ and adjusting the parameter γ to the initial matrix B^0 (sampled randomly in $\mathcal{B}$):

$$\gamma = \frac{1}{2\tilde{\mu}^2},$$

where $\tilde{\mu} = \text{median}\{\|B^0(x_i - x_j)\|, i \neq j\}$. We then solve HKRR without the Nyström approximation setting B as the approximation given by the method, and setting the parameters γ and λ with cross-validation (on a validation set of size $m = 600$). The resulting B and α are used to compute the R2 score on a test set of size $8m = 4800$. For the graph on the left, we consider the first dataset with $D = 50$ and $d_* = 3$, while the left one is obtained based on the second dataset with $D = 50$ and $d_* = 3$. The methods are stopped after 60 seconds of computations and run 5 times per set of parameters.

The bottom graph in Figure 2 was obtained by applying the same process, running AGD and VarPro on $\tilde{m} = 25$ Nyström centers and setting $\lambda = 10^{-8}$. For each dataset size m in $\{1000, 4000, 8000, 12000, 16000\}$, we performed each algorithm 10 times, stopping after 80 seconds.

D.2 A 2D EXAMPLE FOR ALTERNATING MINIMIZATION

Recall that we consider the function $f : x, y \mapsto \left(x - y^2\right)^2 + \cos(\pi y) + (1-y)^2 + 1$. Despite its simplicity, this function shares several similarities with $\hat{\mathcal{L}} : B, \alpha \mapsto \frac{1}{m}\left\|\hat{K}^B_{m\tilde{m}}\alpha - \mathbf{y}\right\|^2 + \lambda\alpha^T\hat{K}^B_{\tilde{m}\tilde{m}}\alpha$:

- it is strongly convex w.r.t. its first variable x and minimizing $x \mapsto f(x,y)$ can be done directly ($x^* = y^2$).
- it is non convex in its second variable, with potentially local minimizers.

Because of this structural similarity to the HKRR objective function, we use Variable Projection (VarPro) and Alternating Gradient Descent (AGD) to minimize f. Note that f has a unique global minimum in $(x^*, y^*) = (1, 1)$.

Figure 3 shows the behavior of both methods applied to f. For the chosen initialization point, it highlights the advantage of taking in account the geometry of f in both x and y since VarPro provides a local minimizer while AGD goes to a global one. This phenomenon occurs not only for cherry-picked initialization points as shown in Figure 4: the attraction basin of the global minimizer of f is significantly larger for AGD on this function.

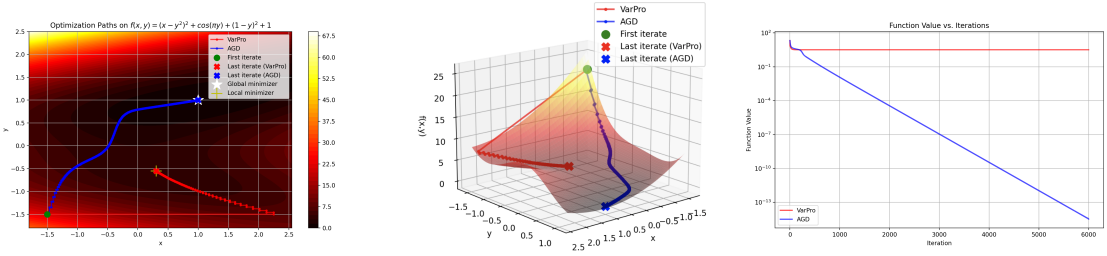


Figure 3: Left and center: Trajectories of the iterates of VarPro (in red) and AGD (in blue) for $f : x, y \mapsto \left(x - y^2\right)^2 + \cos(\pi y) + (1 - y)^2 + 1$ (taking $(x_0, y_0) = (-1.5, -1.5)$). Right: Value of the loss function w.r.t. the number of iterations.

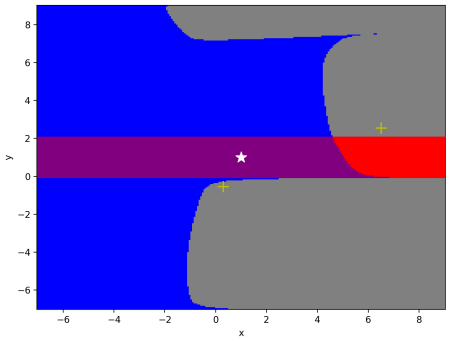


Figure 4: Convergence map of VarPro and AGD for minimizing $f : x, y \mapsto \left(x - y^2\right)^2 + \cos(\pi y) + (1 - y)^2 + 1$. Purple = both methods converge to the global minimum from the corresponding initialization point; Red = only VarPro converges to the global minimum; Blue = only AGD converges to the global minimum; Gray = no method converges to the global minimum. The white star is the global minimizer of f and the yellow '+' crosses are local minimizers.

We can observe a similar behavior on the function $f : x, y \mapsto (x - \sigma(y))^2 + \cos(\pi y) + (1 - y)^2 + 1$ where $\sigma : y \mapsto \frac{1}{1+e^{-y}}$ is the sigmoid function as illustrated in Figure 5 and 6.

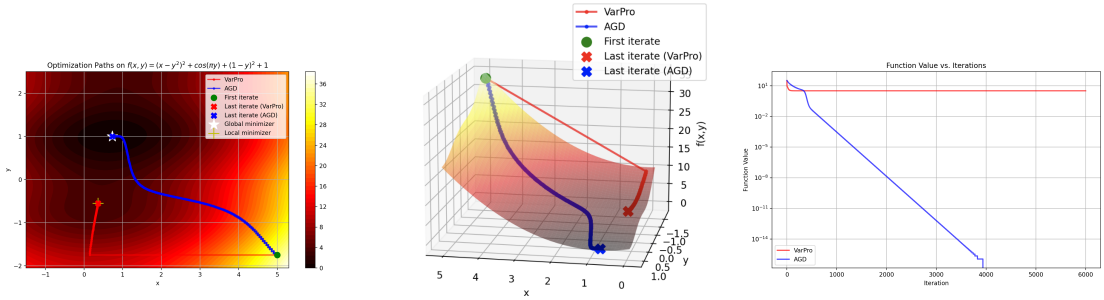


Figure 5: Left and center: Trajectories of the iterates of VarPro (in red) and AGD (in blue) for $f : x, y \mapsto (x - \sigma(y))^2 + \cos(\pi y) + (1 - y)^2 + 1$ (taking $(x_0, y_0) = (5, -1.75)$). Right: Value of the loss function w.r.t. the number of iterations.

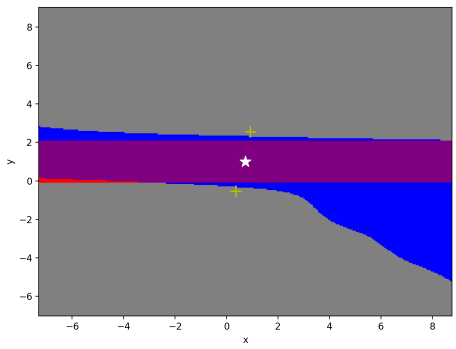


Figure 6: Convergence map of VarPro and AGD for minimizing $f : x, y \mapsto (x - \sigma(y))^2 + \cos(\pi y) + (1 - y)^2$. Purple = both methods converge to the global minimum from the corresponding initialization point; Red = only VarPro converges to the global minimum; Blue = only AGD converges to the global minimum; Gray = no method converges to the global minimum. The white star is the global minimizer of f and the yellow '+' crosses are local minimizers.

E THE USE OF LARGE LANGUAGE MODELS (LLMS)

We made minor use of LLMs to assist with typo detection and language polishing.