Beyond Least Squares: Uniform Approximation and the Hidden Cost of Misspecification

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of controlling worst-case errors in misspecified linear regression under the random design setting, where the regression function is estimated via (penalized) least-squares. This setting arises naturally in value function approximation for bandit algorithms and reinforcement learning. Our first main contribution is the observation that the amplification of the misspecification error when using least-squares is governed by the *Lebesgue constant*, a classical quantity from approximation theory that depends on the choice of the feature subspace and the covariate distribution. We also show that this dependence on the misspecification error is tight for least-squares regression: in general, no method minimizing the empirical squared loss can improve it substantially. As a second contribution, we propose a method that augments the original feature set with auxiliary features designed to reduce the error amplification. For this method we prove an oracle inequality that shows that the method successfully competes with an "oracle" that knows the best way of using the auxiliary features to reduce error amplification. As an illustration, when the domain is a real interval and the features are monomials, we prove that in the limit as $d \to \infty$, our method reduces the amplification factor to $\mathcal{O}(1)$. Note that without our method, least-squares with the monomials (and in fact polynomials) will suffers a worst-case error of order $\Omega(d)$ times the one of the best uniform linear approximator.

1 Introduction

Value function approximation plays a central role in modern reinforcement learning (RL) and contextual bandit algorithms [Sutton and Barto, 2018, Lattimore and Szepesvári, 2020]. In many such settings, policies are evaluated or selected based on value estimates obtained by regressing observed returns. In both cases, (penalized) linear regression—based on empirical squared loss—serves as a core subroutine due to its simplicity and favorable computational properties [Ernst et al., 2005, Antos et al., 2008]. A fundamental challenge arises, however, when the true value function or reward model lies outside the span of the chosen features—a situation referred to as *model misspecification*. Recent work by Du et al. [2020] highlighted that in this setting, the prediction error incurred by least-squares regression can be *amplified* by a factor as large as $\sqrt{d}$, even when the misspecification error itself is small and *the learner is allowed to choose the distribution of the features* (importance sampling). This amplification phenomenon has since drawn significant attention in the RL and bandits communities, due to its implications for the reliability of value estimation under function approximation [Lattimore et al., 2020, Dong and Yang, 2023, Amortila et al., 2023, Maran et al., 2024, Amortila et al., 2024].

In this paper, we study the problem of controlling such worst-case errors in *misspecified linear regression* under the *random design setting*, where inputs are drawn from an unknown distribution. Our first main result is a sharp characterization of how the *amplification of the misspecification error* depends on the interaction between the sampling distribution and the feature subspace. Specifically, we show that this amplification is governed by the *Lebesgue constant*—a classical quantity in approximation theory that captures how well the 2-norm projection underlying least-squares regression projects arbitrary functions onto the span of the features. This result is a significant improvement on previous results. While previous works pointed out that with the *best* covariate distribution the misspecification amplification factor cannot be larger than $\sqrt{d}$ regardless of the feature-map and at the same time, for some feature maps $\sqrt{d}$ is actually the best factor, our approach reveals that the amplification factor can range from as low as 1 for favorable features. In such scenario, one can obtain significantly tighter finite-sample guarantees than previously known, which universally assumed a worst-case $\sqrt{d}$ scaling. Moreover, we prove that this dependence is tight: no estimator based on least squares can substantially improve upon this bound in general.

Motivated by these insights, we propose a method for *reducing the misspecification error amplification* by augmenting the original feature set with extra features and then using a weighted ridge regression approach to regularize the corresponding projection operator. As an illustration of this idea, we show that when the domain is an interval and the base and extra features are monomials, our method reduces the amplification factor to 1 asymptotically as $d \to \infty$. In contrast, standard least squares remains susceptible to arbitrarily large worst-case errors for the same setting.

2 Problem Formulation

Our goal is to learn a linear predictor that enjoys *uniform* accuracy over the whole input space, even when the linear model is misspecified. We first detail the statistical setting, introduce the standing assumptions and define the performance criterion that will be used in the rest of the paper.

The learner receives a dataset $((x_t, y_t))_{t=1}^n$, where $x_t \in \mathcal{X}$ and $y_t \in \mathbb{R}$. Each x_t gives rise to a response $y_t = f(x_t) + \eta_t$ where $f : \mathcal{X} \to \mathbb{R}$ is an unknown function and η_t is a noise variable.

Assumption 1 (Sub–Gaussian Noise)**.** *The noise variables $(\eta_t)_{t=1}^n$ are independent, centered and σ–sub–Gaussian, meaning that, for every $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda\eta_1)] \le \exp(\sigma^2\lambda^2/2)$. The noise variables are independent of the inputs $(x_t)_{t=1}^n$.*

About $(x_t)_t^n$, we are going to make the following assumption.

Assumption 2 (Random design)**.** *Samples $(x_t)_{t=1}^n$ are drawn i.i.d. from a probability distribution μ over $\mathcal{X}$ (unknown to the learner).*

We are interested in the problem of *linear function approximation*, when the learner is given some feature map $\boldsymbol{\varphi}_d : \mathcal{X} \to \mathbb{R}^d$ and aims to approximate f using $f_\theta(\cdot) = \boldsymbol{\varphi}_d(\cdot)^\top\theta$ by selecting some $\theta \in \mathbb{R}^d$ based on the data available to it. In the rest of the paper, we use the short-hand $\varphi_i(x)$ for the i–th coordinate of $\boldsymbol{\varphi}_d(x)$ and index data points by $t = 1, \dots, n$. Differently from most of the literature about this setting, motivated by the applications mentioned earlier, the performance is going to be evaluated via the uniform, or maximum-norm, which for a function $g : \mathcal{X} \to \mathbb{R}$ is defined via $\|g\|_\infty = \sup_{x\in\mathcal{X}} |g(x)|$. We let $L^\infty(\mathcal{X})$ denote the set of functions with finite maximum norm. In what follows, we assume that both f and our features φ_i belong to this set. For $f \in L^\infty(\mathcal{X})$ and $\theta \in \mathbb{R}^d$ we let

$$\mathcal{E}_\infty(\theta, f) := \|f_\theta - f\|_\infty\,, \qquad \mathcal{E}_\infty(f) := \inf_{\theta\in\mathbb{R}^d} \mathcal{E}_\infty(\theta, f).$$

Thus, $\mathcal{E}_\infty(\theta, f)$ is the maximum error suffered when f is approximated using f_θ, while $\mathcal{E}_\infty(f)$ is the smallest possible value for this error; its value is *unknown* to the learner. When $\mathcal{E}_\infty(f) > 0$, we say that the problem of estimating f is *misspecified* and the error $\mathcal{E}_\infty(f)$ is known as the *misspecification error*. In the next section we will be interested in investigating how the error $\mathcal{E}_\infty(\hat\theta_n, f)$ behaves when $\hat\theta_n$ is given by *ordinary least-squares* (OLS) estimate:

$$\hat\theta_{n,\mathrm{OLS}} = \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d} \sum_{t=1}^n \left(y_t - f_\theta(x_t)\right)^2 .$$

3 Characterizing the behavior of OLS and Ridge Regression

Let $\mathcal{F}_d = \{f_\theta : \theta \in \mathbb{R}^d\}$ denote the subspace of $L^\infty(\mathcal{X})$ spanned by the basis functions underlying the feature-map $\boldsymbol{\varphi}_d$. As $n \to \infty$, $f_{\hat{\theta}_{n,\text{OLS}}}$ is known to converge to

$$\Pi_{d,\mu} f := \arg\min_{g \in \mathcal{F}_d} \|g - f\|_\mu^2 \tag{1}$$

with probability one (see Györfi et al. [2006]). Here, we define $\|\cdot\|_\mu^2$ to stand for the $L^2(\mu)$-norm: For $g : \mathcal{X} \to \mathbb{R}$ measurable, $\|g\|_\mu^2 = \int_{\mathcal{X}} g^2(x)\mu(dx)$. Since μ is a probability measure, we have $\|g\|_\mu^2 \le \|g\|_\infty^2$. The map in Eq. (1) is known to be *projection* onto $\mathcal{F}_d$: $\Pi_{d,\mu}$ is linear, idempotent and for all $f \in \mathcal{F}_d$, $\Pi_{d,\mu} f = f$ holds. Moreover, it is non-expansive in the $L^2(\mu)$-norm.

By continuity, the previous comment on the convergence of the OLS estimate implies that $\lim_{n\to\infty} \mathcal{E}_\infty(\hat{\theta}_{n,\text{OLS}}, f) = \|\Pi_{d,\mu} f - f\|_\infty$. The first question then is thus how large $\|\Pi_{d,\mu} f - f\|_\infty$ can be relative to $\mathcal{E}_\infty(f)$, or, in other words, by how much will the misspecification error $\mathcal{E}_\infty(f)$ be enlarged if one uses the linear projection of f to $\mathcal{F}_d$. The following definition will be useful:

Definition 1 (Lebesgue constant)**.** *Let $P : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ be a linear operator. Then, the L^∞−norm of P is called the* Lebesgue constant *associated with P:*

$$\Lambda(P) := (\|P\|_\infty =) \sup_{f \in L^\infty(\mathcal{X}) :\, \|f\|_\infty \le 1} \|Pf\|_\infty .$$

The following result holds (see Proposition 4.1 from Chapter 2 of DeVore and Lorentz [1993]):

Lemma 1 (Lebesgue's lemma)**.** *Let $P : L^\infty(\mathcal{X}) \to \mathcal{F}_d$ be a linear map such that P is an identity on $\mathcal{F}_d$. In particular, assume that for any $f \in \mathcal{F}_d$, $Pf = f$. Then, for any $f \in L^\infty(\mathcal{X})$,*

$$\|f - Pf\|_\infty \le (1 + \Lambda(P)) \inf_{g \in \mathcal{F}_d} \|f - g\|_\infty .$$

Since the Lebesgue constant of our projection operators will be frequently needed, to minimize clutter, we introduce the shorthand for them:

$$\Lambda_{d,\mu} := \Lambda(\Pi_{d,\mu}).$$

With the help of this notation, from Lemma 1 we thus have

$$\|f - \Pi_{d,\mu} f\|_\infty \le \Lambda_{d,\mu} \mathcal{E}_\infty(f) . \tag{2}$$

It is easy to see that $\Lambda_{d,\mu} \ge 1$: just take any $f \in \mathcal{F}_d$ such that $\|f\|_\infty = 1$ (such a function exist). Then $\Pi_{d,\mu} f = f$, which gives the result. Unfortunately, there is no upper limit on how large $\Lambda_{d,\mu}$ can be in general. What is more, the bound in Lemma 1 is essentially tight:

Theorem 3. *For any $\varepsilon > 0$ there exist $f \in L^\infty(\mathcal{X})$ such that*

$$\|f - \Pi_{d,\mu} f\|_\infty \ge (\Lambda_{d,\mu} - 1 - \varepsilon) \mathcal{E}_\infty(f) .$$

For the proof, see Appendix C.1. Because of the last result, we expect that any bound on $\mathcal{E}_\infty(\hat{\theta}_{n,\text{OLS}}, f)$ where $\hat{\theta}_n$ is estimated from data will involve $\Lambda_{d,\mu}\mathcal{E}_\infty(f)$. Our main result of this section is indeed of this form. To state the result, let $(\overline{\varphi}_i)_{1 \le i \le d}$ be the orthonormal basis in $L^2(\mu)$ of $\mathcal{F}_d$ given by the Gram-Schmidt procedure on the original features. We call $\overline{\boldsymbol{\varphi}}_d(x) = (\overline{\varphi}_1(x), \ldots, \overline{\varphi}_d(x))^\top$ and define $\overline{\varphi}_{d,2} = \sup_{x \in \mathcal{X}} \|\overline{\boldsymbol{\varphi}}_d(x)\|_2$.

Theorem 4. *Let $\mathcal{X}$ be finite. Let Assumptions 1 and 2 hold. Then, for any n positive integer and real $0 < \delta \le 1/3$ such that $n \ge 20\overline{\varphi}_{d,2}^2 \log(d/\delta)$, letting $\hat{\theta}_{n,OLS}$ be the parameter vector returned by OLS, with probability at least $1 - 3\delta$,*

$$\mathcal{E}_\infty(\hat{\theta}_{n,OLS}, f) \le (1 + \Lambda_{d,\mu})\mathcal{E}_\infty(f) + 3(\sigma + \Lambda_{d,\mu}\mathcal{E}_\infty(f))\overline{\varphi}_{d,2}\sqrt{\frac{\log(|\mathcal{X}|/\delta)}{n}} + \frac{\mathit{poly}(d, \overline{\varphi}_{d,2}, \Lambda_{d,\mu}\mathcal{E}_\infty(f))}{n} .$$

The first quantity in the bound is the same as in Eq. (2) and this is the quantity that accounts for the gap between $\Pi_{d,\mu} f$ and f. The other terms bound the finite sample error. We gave the result for $\mathcal{X}$ finite only for simplicity. Indeed, the result can be easily extended by means of a simple covering argument. For example, if $\mathcal{X} = [-1, 1]$ and the features are Lipschitz continuous with constant L_φ, we can achieve uniform error over $\mathcal{X}$ by making a $\varepsilon/L_\varphi-$covering of $\mathcal{X}$. In this way, the bound increases just by a factor $\propto \log(L_\varphi/\varepsilon)$. Another possibility, which sometimes can lead to tighter results, is to cover $\{\varphi(x) : x \in \mathcal{X}\} \subset \mathbb{R}^d$.

In addition to $\Lambda_{d,\mu}$, another constant that depends on the feature map is $\overline{\varphi}_{d,2}$. As it turns out, this value "hides" the dimension d. In particular, we show that regardless of the feature map, $\overline{\varphi}_{d,2} \geq \sqrt{d}$ (see Proposition 19 in the appendix).

The terms and their scaling with the relevant quantities are as expected. We have already discussed the first term and argued that it cannot be significantly improved for OLS. The second (and the lower order third) term accounts for the sampling errors. In particular, the effect of the noise is shown through the term involving σ. The next term, which also involves the Lebesgue constant and $\mathcal{E}_\infty(f)$ accounts for the random design sampling error.

Below we show that when an *a priori* upper bound ε on $\mathcal{E}_\infty(f)$ is available (as can be the case in certain numerical applications when the target function belongs to some known class of functions, such as a smoothness class), we can obtain an empirical bound that has the potential to significantly reduce the terms of the bound shown in the last result.

A uniform, semi-empirical bound Define $\mu_n = \frac{1}{n}\sum_{t=1}^n \delta_{x_t}$ to be the empirical measure underlying the inputs $(x_1, \ldots, x_n)$. We are interested in bounding the uniform error of the OLS estimate via empirical quantities. In particular, we will use the Lebesgue constant associated with the projection operator Π_{d,μ_n} corresponding to μ_n. Because of this, we can also remove the assumption that $(x_t)_t$ is sampled from μ; in fact, we will not need any assumptions concerning how $(x_t)_t$ are selected.

The operator Π_{d,μ_n} takes the form

$$\Pi_{d,\mu_n} f(\cdot) := \boldsymbol{\varphi}_d(\cdot)^\top (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{f} \qquad \mathbf{f} := [f(x_1), \ldots f(x_t), \ldots f(x_n)]^\top,$$

where Φ is the $n \times d$ matrix storing, as rows, the features corresponding to every $\boldsymbol{\varphi}_d(x_t)$. As before, we define $(\widehat{\varphi}_i)_{1\leq i\leq d}$ the orthonormal basis in $L^2(\mu_n)$ of $\mathcal{F}_d$. We let $\widehat{\boldsymbol{\varphi}}_d(x) = (\widehat{\varphi}_1(x), \ldots, \widehat{\varphi}_d(x))^\top$ and define $\widehat{\varphi}_{d,2} = \sup_{x\in\mathcal{X}} \|\widehat{\boldsymbol{\varphi}}_d(x)\|_2$.

Theorem 5. *Let Assumption 1 hold. Then, for any fixed $\delta > 0$, with probability at least $1 - \delta$,*

$$\mathcal{E}_\infty(\widehat{\theta}_{n,\mathrm{OLS}}) \leq (1 + \Lambda_{d,\mu_n})\mathcal{E}_\infty(f) + \frac{\sigma \widehat{\varphi}_{d,2}\sqrt{2\log(2\mathcal{X}/\delta)}}{\sqrt{n}}.$$

Compared to Theorem 4, we both removed Assumption 2 and the lower-order term $n^{-1}\mathrm{poly}(d, \overline{\varphi}_{d,2}, \Lambda_{d,\mu}\mathcal{E}_\infty(f))$. At the same time, the Lebesgue constant $\Lambda_{d,\mu}$ is replaced with Λ_{d,μ_n}, which may be smaller or larger than $\Lambda_{d,\mu}$. When $\mathcal{X}$ is finite then Λ_{d,μ_n} can be calculated in $O(n|\mathcal{X}|)$ time (it is just a matrix maximum norm).

If choosing the points $(x_1, \ldots, x_n)$ is an option, one may attempt to optimize the bound. Here, besides the term Λ_{d,μ_n}, $\widehat{\varphi}_{d,2}$ also hides μ_n. In experimental optimal design, the G-optimal design is defined as the one that minimizes $\widehat{\varphi}_{d,2}$ by choosing an appropriate distribution μ_n. Here, one of the main results is that for $n = \Omega(d)$ (or slightly larger), one can find μ_n such that $\widehat{\varphi}_{d,2} = O(\sqrt{d})$ and this is the best possible value Kiefer and Wolfowitz [1960].

Under the assumption that μ_n is an optimal design, we can compare our result with Proposition 5.1 from Lattimore et al. [2020] (see equation (2) and the corresponding bound in high probability). Rephrasing their proposition in our notation, we get, roughly that, if μ is an optimal design for $\boldsymbol{\varphi}_d$, then for $\sigma = 1$,

$$\mathcal{E}_\infty(\hat{\theta}_{n,\mathrm{OLS}}, f) \leq \mathcal{O}\left(\sqrt{d}\mathcal{E}_\infty(f) + \sqrt{\frac{d\log(|\mathcal{X}|/\delta)}{n}}\right).$$

This result is a particular case of our Theorem 4: indeed $\widehat{\varphi}_{d,2} = \sqrt{d}$ as we are using an optimal design, while it is not hard to see that $\Lambda_{d,\mu_n} \leq \widehat{\varphi}_{d,2}$ holds without any assumptions. In this bound,

Basis functions	μ	$\Lambda_{d,\mu}$	**Source**	**Note**
Polynomial	uniform on regular d-grid	$\Omega(2^d)$	[Quarteroni et al., 2010]	
Polynomials	uniform	$\Theta(d)$	DeVore and Lorentz [1993]	$\overline{\varphi}_{d,2} \approx d$
Fourier	uniform	$\mathcal{O}(\log(d))$	[Katznelson, 2004, p.59, Excercise 1]	
Continuous B-splines	uniform	$\mathcal{O}(1)$	Huang [2003]	
Wavelets	uniform	$\mathcal{O}(1)$	Chen and Christensen [2013]	

Table 1: Examples of Lebesgue constants. Domain is $\mathcal{X} = [-1, 1]$.

for large n, $\sqrt{d}\mathcal{E}_\infty(f)$ is the dominant term. Therefore, our tighter bound with the Lebesgue constant achieves a better result whenever the Lebesgue constant is significantly smaller than $\sqrt{d}$. As an example, of a feature-map with a small Lebesgue constant, consider any partitioning $(\mathcal{X}_i)_i$ of $\mathcal{X}$ and set φ_i to be the indicator of part $\mathcal{X}_i$, $i = 1, \ldots, d$. Then, the Lebesgue constant of $\boldsymbol{\varphi}_d$ is 1 regardless of the choice of μ. Note that with some extra work we can extract a more refined result from the proof of Proposition 5.1 when $\sqrt{d}$ on the right-hand-side above is essentially replaced by $\widehat{\varphi}_{d,2}$.

3.1 The Lebesgue constant: properties and particular cases

While as noted earlier $\Lambda_{d,\mu} \leq \overline{\varphi}_{d,2}$ always hold, since $\overline{\varphi}_{d,2}$ is never lower than $\sqrt{d}$, to get a better understanding of the Lebesgue constants associated with specific feature maps, we need to resort to feature-map dependent analysis. For many of the classical feature-maps, to under we have to resort of some feature map-dependent analysis. In the following, we enumerate few well-known/novel results in Table 1. As seen in the table, the range of values is quite large. Notably, polynomials with regular d-grids, show the worst-behavior (though this is somewhat unusual since here for every d one uses a different measure). Yet, polynomials with the uniform measure still exhibit quite big Lebesgue constants. Perhaps surprisingly, when switching to Fourier series, the Lebesgue constant decreases to $O(\log d)$. As such, if the convergence of a Fourier series to the target function is fast enough, there is little to no incentive to go beyond the L^2-projections, or least-squares. It is interesting to note that some researchers have empirically found the Fourier series as a good "general" basis to be used in reinforcement learning [Konidaris et al., 2011]. This raises the hypothesis that this is primarily due to the reasonable error extrapolation properties of Fourier series, as attested by its relatively slow-growing Lebesgue constant. Finally, we mention that perhaps unsurprisingly localized basis functions such as wavelets and B-splines have Lebesgue constants that are constant independently of the number of basis functions used. As such, when uniform approximation is important, it seems that there are good reasons to prefer these systems. We speculate that tile-coding, which is itself localized and which is quite popular in reinforcement learning also shares the good extrapolation properties of these localized systems of basis functions.

One weakness of the above results is that they are dependent on the choice of the sampling distribution μ. The following result shows that the Lebesgue constant changes gradually as one moves from one distribution to another, provided that the overlap between the two distributions is well-controlled:

Proposition 6. *Let μ, ν be two probability distributions on the discrete set $\mathcal{X}$ such that for all $x \in \mathcal{X}, C \geq \frac{\mu(x)}{\nu(x)} \geq c > 0$. Then, $\Lambda_{d,\mu} \leq \frac{C}{c}\Lambda_{d,\nu}$.*

4 Regularized estimators

The previous Theorem 3 shows that, whatever the feature map, if we use the OLS estimator, the error is forced to scale with the misspecification multiplied by the Lebesgue constant. This is not a matter of overfitting, as the bound holds for infinite data; still, the problem is related to the LS solution becoming "too big" for some choices of f. Therefore, an idea would be to enforce a regularization on the LS loss, to limit the magnitude of the estimated function. In the next theorem, we show that the standard Ridge Regression approach is ineffective, even when knowing the orthonormal basis $\overline{\boldsymbol{\varphi}}_d$.

Theorem 7. *Let $\widehat{\theta}_{n,\text{RIDGE}}$ the output of $\lambda-$ridge regression. For any feature map $\boldsymbol{\varphi}_d(\cdot) : \mathcal{X} \to \mathbb{R}^d$ there is $f \in L^\infty(\mathcal{X})$ such that, for infinite data $\mathcal{E}_\infty(\widehat{\theta}_{\infty,\text{RIDGE}}) = \Omega\left(\max\left\{(\Lambda_{d,\mu} - 2\lambda)\mathcal{E}_\infty(f), \frac{\lambda}{\lambda+1}\right\}\right)$.*

This result tells that the Ridge regularization is ineffective: if we take $\lambda \approx \Lambda_{d,\mu}/2$ the second term is close to one (we do not go to zero even for $\mathcal{E}_\infty(f) \to 0$), if we do not, we get the same lower bound of OLS. Crucially, this phenomenon persists even in the infinite data regime, indicating that it is not merely a sample size issue, but a geometric defect of the projection operator itself. Other regularization techniques like Cross-Validation and Early Stopping [Ghojogh and Crowley, 2019], that are designed for dealing with overfitting, cannot overcome this result, as they aim at minimizing the test error MSE, which is achieved by OLS for infinite data.

Let us dig deeper into the reason behind the failure of ridge regression. The proof of Theorem 7, builds on the fact that the corresponding operator $\Pi^{\text{Ridge}}_{d,\mu}$ can be written in the following form

$$\Pi^{\text{Ridge}}_{d,\mu} f(x) = \sum_{i=1}^{d} \alpha \overline{\varphi}_i(x) \int_{\mathcal{X}} \overline{\varphi}_i(z) f(z)\, d\mu(z) \qquad \alpha = \frac{1}{1+\lambda}. \tag{3}$$

In fact, this is *not* a projection operator. Indeed, this does to corresponds to the identity when applied over $\mathcal{F}_d$. F.e. taking $f(\cdot) = \overline{\varphi}_1(\cdot)$, we get $\overline{\varphi}_1(x) - \Pi^{\text{Ridge}}_{d,\mu}\overline{\varphi}_1(x) = \frac{\lambda}{1+\lambda}\overline{\varphi}_1(x)$. Indeed, in order to scale with $\mathcal{E}_\infty(f)$ we have to ensure that any function in $\mathcal{F}_d$ is kept fixed by the operator associated to our estimator for θ. Still, keeping $\alpha = 1$, that means $\lambda = 0$ means going back to OLS.

Extending the feature map Let us assume to add one more feature to $\boldsymbol{\varphi}_d(\cdot)$, which we now call $\boldsymbol{\varphi}_{d+1}(\cdot)$. Indeed, regardless of the nature of the feature that we add, $\mathcal{E}_\infty(f)$ can only decrease, as we are taking the infimum over a larger set. Intuition, together with the results from Section 3.1, would suggest that the corresponding Lebesgue constant, that passes from $\Lambda_{d,\mu} \to \Lambda_{d+1,\mu}$ could only get bigger. Surprisingly, this is not the case: there are examples of feature maps such that adding one feature may correspond to the Lebesgue constant getting *smaller*.

To formalize this idea, let us assume to expand the original feature map with $D-d$ different functions, for some integer $D > d$, so that the full feature map can be written as $\boldsymbol{\varphi}_D(\cdot) := [\boldsymbol{\varphi}_d(\cdot), \boldsymbol{\varphi}'_{D-d}(\cdot)]$. Even if the added features $\boldsymbol{\varphi}'_{D-d}(\cdot)$ can be arbitrary, we argue that in many real problems there is a way to select them in a reasonable way. For example, in all the examples mentioned in Section 3.1 (Fourier features, polynomials, splines...) the sequence considered can be easily extended up to infinity by enlarging the maximum degree. We call, as before, $\overline{\boldsymbol{\varphi}}_D(\cdot)$ the feature map that we obtain from the Gram-Schmidt procedure on the basis $\boldsymbol{\varphi}_D(\cdot)$ with measure $\mu(\cdot)$. We consider the operator associated to *weighted* ridge regression over this extended basis, which generalizes Eq. (3). This writes, for any sequence of D weights λ_i in $[0,\infty)$ as:

$$\Pi^{\text{Ridge}}_{\boldsymbol{\alpha},\mu} f(x) := \sum_{i=1}^{D} \alpha_i \overline{\varphi}_i(x) \int_{\mathcal{X}} \overline{\varphi}_i(z) f(z)\, d\mu(z). \qquad \alpha_i = \frac{1}{1+\lambda_i}. \tag{4}$$

In fact, it shall be proved that this operator is the minimizer of the weighted Ridge loss, when adding a penalization λ_i on each component $\overline{\varphi}_i(x)$. Not every value for the sequence $\boldsymbol{\alpha}$ is meaningful: in fact, the original components $i \le d$ must *not* be penalized. Moreover, as the ridge penalization of each component ranges in $[0,+\infty)$, the corresponding $\alpha_i \in (0,1]$. These two constrains are formalized in the following set:

$$\mathcal{A}^D_d := \{\boldsymbol{\alpha} \in [0,1]^D : \forall i \le d,\ \alpha_i = 1\}, \tag{5}$$

which we call the set of *attenuation parameters*.

4.1 Weighted ridge estimator and the Oracle Operator

Each of there operators $\Pi^{\text{Ridge}}_{\boldsymbol{\alpha},\mu}$, as long as $\boldsymbol{\alpha} \in \mathcal{A}^D_d$, maintains every element of $\mathcal{F}_d$ (the span of the original feature map $\boldsymbol{\varphi}_d(\cdot)$) as fixed point. We call $\Lambda_{\boldsymbol{\alpha},\mu}$ the Lebesgue constant of the corresponding operator. The following result holds, generalizing Lebesgue's Lemma.

Proposition 8. *Let $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ (equation (5)) and $\Pi^{\text{Ridge}}_{\boldsymbol{\alpha},\mu}$ be defined according to equation (4). Then,*

$$\|f(\cdot) - \Pi^{\text{Ridge}}_{\boldsymbol{\alpha},\mu} f(\cdot)\|_\infty \leq (1 + \Lambda_{\boldsymbol{\alpha},\mu})\mathcal{E}_\infty(f).$$

Crucial for the proof of Proposition 8 is the definition of $\mathcal{A}_d^D$, ensuring that the attenuation factor α_i is one for $i \leq d$, so that the projection on the original features is maintained exactly as it is. This ensures that the original features are *not* penalized by the Ridge regularization. Under our point of view, to minimize the expansion factor corresponds to using the value $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ which has the minimal Lebesgue constant. We call this value ORACLE:

$$\boldsymbol{\alpha}^{\text{Oracle}}_\mu := \arg\min_{\boldsymbol{\alpha} \in \mathcal{A}_d^D} \Lambda_{\boldsymbol{\alpha},\mu} \qquad \Lambda^{\text{Oracle}}_\mu := \min_{\boldsymbol{\alpha} \in \mathcal{A}_d^D} \Lambda_{\boldsymbol{\alpha},\mu}.$$

Unfortunately, $\boldsymbol{\alpha}^{\text{Oracle}}_\mu$ is unknown to the learner, as it depends on the unknown distribution μ. Our questions for the rest of this section are the following:

- **Q1** Can we design a finite sample estimator that, for fixed $\boldsymbol{\alpha} \in \mathcal{A}_d^D$, asymptotically, scales as $\Lambda_{\boldsymbol{\alpha},\mu}$?
- **Q2** Can we design a finite sample estimator that, asymptotically, scales as $\Lambda^{\text{Oracle}}_\mu$?

Q1 To answer both the previous questions, we start by generalizing Theorem 5 to the case of regularization with the desired parameter $\boldsymbol{\alpha}$. Indeed, even being μ unknown, we can define an empirical counterpart of the operator defined in Eq. (4) by means of the empirical measure $\mu_n(\cdot)$. In fact, recalling that $\widehat{\boldsymbol{\varphi}}_D(\cdot)$ is the feature map obtained by orthogonalizing $\boldsymbol{\varphi}_D(\cdot)$ w.r.t. $\mu_n(\cdot)$, we have

$$\Pi^{\text{Ridge}}_{\boldsymbol{\alpha},\mu_n} f(\cdot) := \sum_{i=1}^D \alpha_i \widehat{\varphi}_i(x) \frac{1}{n} \sum_{t=1}^n \widehat{\varphi}_i(x_t) f(x_t). \tag{6}$$

This operator 1) takes as argument only the evaluations of $f(\cdot)$ at x_t and 2) Outputs a linear combination of the features $\widehat{\varphi}_i(x)$. Thanks to the first point, we can estimate $\Pi^{\text{Ridge}}_{\boldsymbol{\alpha},\mu_n} f(\cdot)$ by our noisy samples by replacing $f(x_t)$ with y_t. Thanks to the second one, there exists an estimator $\widehat{\theta}$ such that the result of the operator is written as $\boldsymbol{\varphi}_D(\cdot)^\top \widehat{\theta}$. Calling R_n the triangular matrix such that $\boldsymbol{\varphi}_D(\cdot)^\top = \widehat{\boldsymbol{\varphi}}_D(\cdot)^\top R_n$ (the matrix corresponding to Grham Scmhidt procedure), we call

$$\widehat{\theta}_{n,\boldsymbol{\alpha}} := R_n^{-1} I_{\boldsymbol{\alpha}} \frac{1}{n} \sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_d(x_t) y_t, \tag{7}$$

where $I_\alpha = \text{diag}(\boldsymbol{\alpha})$ takes into account the regularization, we can prove the following result.

Theorem 9. *Let assumption 1 hold. Then, for any $\delta > 0$, with probability $1 - \delta$,*

$$\mathcal{E}_\infty(\widehat{\theta}_{n,\boldsymbol{\alpha}}) \leq (1 + \Lambda_{\boldsymbol{\alpha},\mu_n})\mathcal{E}_\infty(f) + \frac{\sigma \widehat{\varphi}_{2,D} \sqrt{2\log(2\mathcal{X}/\delta)}}{\sqrt{n}}.$$

Therefore, our estimator $\widehat{\theta}_{n,\boldsymbol{\alpha}}$ is able to scale with $\Lambda_{\boldsymbol{\alpha},\mu_n}$. To answer question **Q1** completely, we just need to show that, for large enough n, we can replace the Lebesgue constant of μ_n with the one of μ. This is done in the following proposition.

Proposition 10. *Under assumption 2 we have, with probability $1 - \delta$ for every $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ at the same time, $|\widehat{\varphi}_{D,2} - \overline{\varphi}_{D,2}| \leq \widetilde{\mathcal{O}}(\overline{\varphi}_{D,2}^2 \sqrt{\log(1/\delta)/n})$, and $|\Lambda_{\boldsymbol{\alpha},\mu_n} - \Lambda_{\boldsymbol{\alpha},\mu}| \leq \widetilde{\mathcal{O}}\left(\frac{\sqrt{d}\overline{\varphi}_{D,2}^2\sqrt{\log(1/\delta)}}{\sqrt{n}} + \frac{\sqrt{d}\overline{\varphi}_{D,2}^3 \log(1/\delta)}{n}\right)$.*[1]

[1]The statement of this theorem is slightly different from the one in the main paper of the submission, as we have made the orders of magnitude more precise.

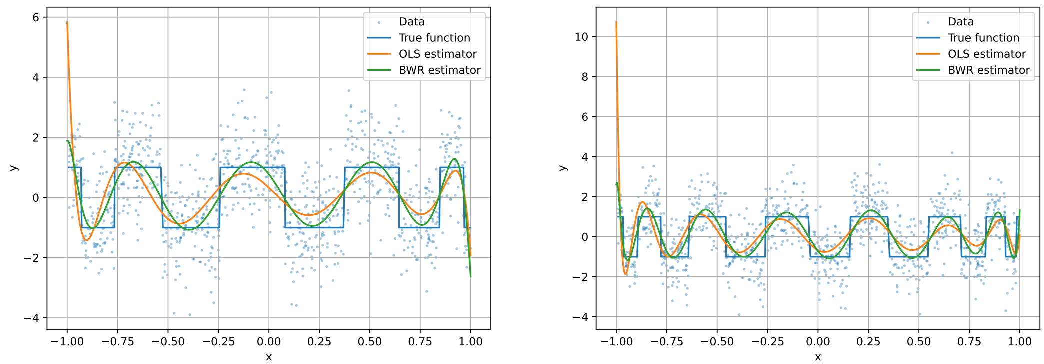

Figure 1: Comparison between the OLS estimator and the BWR estimator using polynomial features on $[-1, 1]$, with a basis of length $d = 10$ (left) and $d = 15$ (right). Even if the true function is bounded, OLS suffers from large oscillations near the boundaries due to the high Lebesgue constant. In contrast, BWR achieves a much more uniform approximation error across the domain by effectively controlling the amplification effect.

Q2 For this more challenging result, we have to optimize the value of $\boldsymbol{\alpha}$ in order to converge to the one of the oracle, even not knowing the true distribution $\mu(\cdot)$. Our strategy, also for this point, is to work with we can compute: the operator $\Pi^{\text{Ridge}}_{\boldsymbol{\alpha},\mu_n}$ (equation (6)) and its Lebesgue constant $\Lambda_{\boldsymbol{\alpha},\mu_n}$. One observation is key for this goal: the Lebesgue constant is convex in $\boldsymbol{\alpha}$.

Proposition 11. *The function $J : \mathcal{A}_d^D \to (0, +\infty)$ given by $J(\boldsymbol{\alpha}) := \Lambda_{\boldsymbol{\alpha},\mu_n}$ is convex in $\boldsymbol{\alpha}$.*

This result allows us to provably arrive to one minimizer of the Lebesgue constant in a finite number of iteration. The idea is what follows: we start from any $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ and update it iteratively with the sub-Gradient method until convergence. This algorithm is well-known Boyd et al. [2003], but, for completeness, we include it; see Algorithm 1 in appendix D.4.

Theorem 12. *Fix $\epsilon > 0$. Algorithm 1, after a number of iterations $I = \widetilde{\mathcal{O}}(\epsilon^{-2}\widehat{\varphi}^2_{2,D}(D-d))$ outputs $\boldsymbol{\alpha}^{(I)} \in \mathcal{A}_d^D$ such that $J(\boldsymbol{\alpha}^{(I)}) \le \inf_{\boldsymbol{\alpha}\in\mathcal{A}_d^D} J(\boldsymbol{\alpha}) + \epsilon$.*

By definition of $J(\cdot)$, the former result entails that $\boldsymbol{\alpha}^{(I)}$ is close to be a minimizer of $\Lambda_{\boldsymbol{\alpha},\mu_n}$. To finally answer **Q2**, we define the following estimator, which corresponds to equation (7) for $\boldsymbol{\alpha}^{(I)}$:

$$\widehat{\theta}_{n,\text{BWR}} := R_n^{-1} I_{\boldsymbol{\alpha}^{(I)}} \frac{1}{n} \sum_{t=1}^{n} \widehat{\boldsymbol{\varphi}}_D(x_t) y_t. \tag{8}$$

The estimator is called BWR, which stands for "Best Weighted Regularizer". We close this section with its performance guarantee.

Theorem 13. *Let Assumptions 1 and 2 hold and fix $\delta > 0$. Then, with probability $1 - \delta$,*

$$\mathcal{E}_\infty(\widehat{\theta}_{n,BWR}) \le (1 + \Lambda_\mu^{Oracle})\mathcal{E}_\infty(f) + \widetilde{\mathcal{O}}\left(\frac{\overline{\varphi}_{2,D}\sqrt{D\log(|\mathcal{X}|/\delta)}}{\sqrt{n}} + \frac{\overline{\varphi}^2_{2,D}\log(|\mathcal{X}|/\delta)}{n}\right).$$

This oracle inequality answer also **Q2** in a positive way: our estimator is asymptotically able to compete with the Oracle Lebesgue constant. We close this paper with a case study of wide interest where this property allows $\widehat{\theta}_{n,\text{BWR}}$ to get a much better result than the one of OLS.

5 Case study: polynomial basis

The method introduced in the previous section aims at reducing the amplification of the misspecification error by controlling the Lebesgue constant. While it applies broadly, its impact is best illustrated in settings where standard estimators suffer from poor uniform behavior. One such setting is the classical case where the feature map φ_d consists of the first d monomials $\{1, x, x^2, \dots, x^{d-1}\}$ over a compact interval. We consider the scenario where $\mathcal{X} = [-1, 1]$ and the data-generating distribution μ is the uniform distribution on this interval. Even in this most favorable case, the Lebesgue constant associated with the polynomial basis grows linearly with the degree: $\Lambda_{d,\mu} \approx d$. This leads to a worst-case uniform error for the OLS estimator that scales as $\mathcal{O}(d \cdot \mathcal{E}_\infty(f))$, which can be arbitrarily large even when the misspecification bias $\mathcal{E}_\infty(f)$ is small.

In contrast, the BWR estimator described at the end in Section 4 augments the feature space with additional monomials and optimizes an attenuation vector to minimize the empirical Lebesgue constant. The result is a projection operator that preserves the behavior of the original features while drastically reducing the amplification of the uniform error. Theoretically, this allows reducing the amplification factor from $\mathcal{O}(d)$ down to $\mathcal{O}(1)$, as the following theorem shows.

Theorem 14. *Let $\mu(\cdot) = \mathcal{U}([-1, 1])$. There is a constant C independent of d such that, for $D = 2d$ and $\boldsymbol{\varphi}_d(x) = [1, \dots x^{d-1}]$, $\boldsymbol{\varphi}_D(x) = [1, \dots, x^{2d-1}]$, we have $\Lambda_\mu^{Oracle} \leq C$.*

This improvement is evident in the numerical simulation shown in Fig. 1, where we compare the OLS estimator to the BWR estimator on synthetic data. While OLS exhibits large oscillations near the boundary of the interval—a manifestation of the classical Runge phenomenon—BWR remains stable across the domain and achieves significantly smaller uniform error. Despite both estimators using the same base features, the control of the Lebesgue constant yields a qualitative and quantitative advantage for BWR.

The above simulations visually demonstrate how the amplification factor is exacerbated by increasing d. We complement them with an asymptotic result that shows how heavy this factor is, even for a function such that $\mathcal{E}_\infty(f) \overset{d}{\to} 0$. In fact, the exist a bounded function that can be uniformly approximated with polynomial features but such that the OLS estimator diverges with uniform error roughly of order $\Omega(d)$.

Proposition 15. *Fix $\gamma > 0$. There is a function $f : [-1, 1] \to \mathbb{R}$ such that, $\mathcal{E}_\infty(f) \overset{d}{\to} 0$ and under assumptions 1 and 2 for $\mu = \mathcal{U}([-1, 1])$, with probability one,*

$$\lim_{d\to\infty} \lim_{n\to\infty} \|f(\cdot) - \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,BWR}\|_\infty = 0 \qquad \lim_{n\to\infty} \|f(\cdot) - \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,OLS}\|_\infty \gtrsim d^{1-\gamma}.$$

6 Related works

The problem we deal in this paper, while motivated by Online Learning applications has roots in several different fields. Not just Mathematical Analysis and Fourier Series, but also Econometrics and Online Learning. Here, we give a short overview of the main papers, leaving an extended version for the appendix A.

In mathematical analysis the problem of projecting onto a linear subspace of $L^\infty(\mathcal{X})$ in a way that minimizes the uniform error have always been of great interest. Several results about orthogonal polynomials Szegő [1939] or Fourier Series Katznelson [2004] approximation have this goal. More recently, Kobos and Lewicki [2024] proposed an approach for general feature maps. Passing to the case when an unknown function is estimated via noisy samples, there is a line of research (Newey [1997], Belloni et al. [2015], and Li and Liao [2020]) that studies the properties of *pointwise* estimators based on LS. The latter can be naturally adapted to achieve a uniform convergence guarantee. A similar problem was recently studied, in a totally different context, by Online Learning papers (see Du et al. [2020], Lattimore et al. [2020], Maran et al. [2024], Dong and Yang [2023], Amortila et al. [2024]) under the name of *misspecified linear function approximation*.

The specific technique that we use in section 4 is inspired by an old method for regularizing Fourier series [de la Vallée Poussin, 1918, De La Vallée Poussin et al., 1919]. The technique he invented is still studied today in numerical mathematics [Németh, 2016, Themistoclakis and Van Barel, 2017, Occorsio and Themistoclakis, 2025].

7 Conclusion

We investigated the problem of uniform error control in misspecified linear regression under the random design setting. Our key insight is that the amplification of $\mathcal{E}_\infty(f)$ by least-squares methods is governed by the Lebesgue constant, a concept from approximation theory. We showed that this amplification is tight and intrinsic to the projection geometry, thereby exposing a fundamental limitation of ordinary and ridge least-squares methods, even in the infinite data regime.

To overcome this limitation, we introduced a novel regularization framework based on weighted ridge regression over extended feature sets, which preserves the desirable properties of the base features while attenuating the contribution of auxiliary ones. We proved that this approach allows us to, asymptotically for $n \to \infty$, compete with the best possible (oracle) projection in terms of uniform error, and we proposed an efficient algorithm for learning such weights from data. In the polynomial basis case, we demonstrated a dramatic improvement: from $\Omega(d)$ amplification with OLS to the optimal $\mathcal{O}(1)$ with our method.

References

Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford university press, 2000.

Philip Amortila, Nan Jiang, and Csaba Szepesvári. The optimal approximation factors in misspecified off-policy value function estimation. In *ICML*, pages 768–790, 2023.

Philip Amortila, Tongyi Cao, and Akshay Krishnamurthy. Mitigating covariate shift in misspecified regression with applications to reinforcement learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 130–160. PMLR, 2024.

Andras Antos, Csaba Szepesvári, and Rémi Munos. Fitted Q-iteration in continuous action-space MDPs. In *Advances in neural information processing systems*, pages 9–16, 2008.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2): 345–366, 2015.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

Stephen Boyd, Lin Xiao, and Almir Mutapcic. Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2004(01), 2003.

Xiaohong Chen and Timothy Christensen. Optimal uniform convergence rates for sieve nonparametric instrumental variables regression. *arXiv preprint arXiv:1311.0412*, 2013.

E.W. Cheney. *Introduction to approximation theory*. McGraw-Hill, 1966.

Ch de la Vallée Poussin. Sur la meilleure approximation des fonctions d'une variable réelle par des expressions d'ordre donné. *CR Acad. Sci. Paris*, 166:799–802, 1918.

Ch J De La Vallée Poussin et al. *Leçons sur l'approximation des fonctions d'une variable réelle*. Paris, 1919.

Ronald A DeVore and George G Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, January 1993.

Jialin Dong and Lin Yang. Does sparsity help in learning misspecified linear bandits? In *International Conference on Machine Learning*, pages 8317–8333. PMLR, 2023.

Zlatko Drmač, Matjaž Omladič, and Krešimir Veselić. On the perturbation of the cholesky factorization. *SIAM Journal on Matrix Analysis and Applications*, 15(4):1319–1332, 1994.

Simon S Du, Sham M Kakade, Ruosong Wang, and Lin F Yang. Is a good representation sufficient for sample efficient reinforcement learning? In *International Conference on Learning Representations*, 2020.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Benyamin Ghojogh and Mark Crowley. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*, 2019.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14:569–600, 2014.

Jianhua Z Huang. Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31 (5):1600–1635, 2003.

Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Tomasz Kobos and Grzegorz Lewicki. On the dimension of the set of minimal projections. *Journal of Mathematical Analysis and Applications*, 529(2):127250, 2024.

Andrei Kolmogoroff. Über die beste annäherung von funktionen einer gegebenen funktionenklasse. *Annals of Mathematics*, 37(1):107–110, 1936.

George Konidaris, Sarah Osentoski, and Philip Thomas. Value function approximation in reinforcement learning using the Fourier basis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 25, pages 380–385, 2011.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Tor Lattimore, Csaba Szepesvári, and Gellért Weisz. Learning with good feature representations in bandits and in RL with a generative model. In *International conference on machine learning*, pages 5662–5670. PMLR, 2020.

Jia Li and Zhipeng Liao. Uniform nonparametric inference for time series. *Journal of Econometrics*, 219(1):38–51, 2020.

G.G. Lorentz. *Approximation of functions*. Athena. New York, Holt, Rinehart and Winston, 1966.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restelli. Local linearity: the key for no-regret reinforcement learning in continuous MDPs. *arXiv preprint arXiv:2410.24071*, 2024.

Zsolt Németh. *De la Vallée Poussin Type Approximation Methods*. PhD thesis, Eötvös Loránd University, Hungary, 2016.

Whitney K Newey. Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168, 1997.

Donatella Occorsio and Woula Themistoclakis. De la Vallée Poussin filtered polynomial approximation on the half–line. *Applied Numerical Mathematics*, 207:569–584, 2025.

Allan Pinkus. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37. Springer Science & Business Media, 2010.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Gábor Szegő. *Orthogonal polynomials*, volume 23. American Mathematical Society, 1939.

Woula Themistoclakis and Marc Van Barel. Generalized de la Vallée Poussin approximations on $[-1, 1]$. *Numerical Algorithms*, 75:1–31, 2017.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: —

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limits of the paper and the future research directions in order to address them.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the statements are provided with proofs in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include the code in the supplementary material (very simple, just one very short Jupyter notebook)

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include the code in the supplementary material (very simple, just one very short Jupyter notebook)

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We include the code in the supplementary material (very simple, just one very short Jupyter notebook)

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: -

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The simulation is straightforward and its computational time is negligible.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper is coherent with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: —

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: —

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: —

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: —

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: —

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: —

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: —

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Related Works

Classical approximation theory The idea of approximating a class of functions with a family of vector spaces in a uniform sense has always been an important topic in mathematical analysis. On the more general level, this theory takes the name of *Kolmogorov's n-width* (Kolmogoroff [1936]; see Lorentz [1966] and Pinkus [2012] for a more modern formalization). The idea, central to this paper, of finding a *linear* operator that well approximates the *non-linear* L^∞ projection operator has also been the main topic of multiple line of research. In particular, many result about orthogonal polynomials Szegő [1939] or Fourier Series Katznelson [2004] approximation have this goal. More recently, Kobos and Lewicki [2024] studied the problem for general feature map, investigating the class of linear operators that achieve the lower bound.

Asymptotic pointwise and uniform convergence of LS series in the econometric literature In the econometric literature, the series least squares (LS) estimators have been analyzed primarily through an asymptotic lens: with the sample size $n \to +\infty$ and the basis dimension $d \to +\infty$, one studies asymptotic Gaussianity of the estimator of the function in each single point. Newey [1997] provided seminal results for this literature, which were then improved by Belloni et al. [2015], the first to use the Lebesgue constant in this field, and by Li and Liao [2020], who generalize the result to time series data. All these contributions, however, remain *asymptotic*: they provide limiting distributions or rates without explicit high–probability bounds, and—crucially—they do not propose algorithmic modifications capable of *reducing* the amplification factor induced by the Lebesgue constant.

Uniform bounds for linear regression in the context of Online Learning As anticipated in the introduction, the problem of getting L^∞ bounds for regression over a domain naturally arises in the context of Online Learning with linear function approximation; bandits and RL in particular. Du et al. [2020] established the first $\sqrt{d}$ amplification lower bound in some specific cases, which was then refined by Lattimore et al. [2020], who also derives the corresponding an upper bound of $\sqrt{d}$, using an optimal design argument. In fact, it can be proved that the factor $\sqrt{d}$ is precisely the maximal Lebesgue constant of any feature map for μ that is the optimal design. These lower bound hold for a worst-case feature map, but *allowing the learner to choose the data distribution*. Following these works, many papers tried to understand how this amplification factor could be reduced. Maran et al. [2024] shows how to remove it in case of a locally linear feature map; Dong and Yang [2023] improves the $\sqrt{d}$ amplification in case of sparsity. Perhaps, the most similar paper to our one is Amortila et al. [2024], which proposes a method to mitigate the effect of misspecification w.r.t. the least-squares fitting. Still, the latter focuses on a different objective, i.e. the error under covariate shift (measuring the MSE under a distribution $\nu \neq \mu$), and scales with the density ratio $\nu(\cdot)/\mu(\cdot)$. Generalizing to the uniform error would mean to take $\nu(\cdot)$ as a Dirac's delta, which would make this bound vacuous.

De la Valleè Poussin approach The to reduce the Lebesgue constant by adding auxiliary features is rooted in a concept that dates back in the history of mathematics to Baron de la Vallée Poussin [de la Vallée Poussin, 1918, De La Vallée Poussin et al., 1919]. The technique he invented is still studied today in numerical mathematics [Németh, 2016, Themistoclakis and Van Barel, 2017, Occorsio and Themistoclakis, 2025].

Finite-sample bounds for ridge regression Hsu et al. [2014] gives finite-sample bounds for ridge regression under random design. The results, when translated into our setting, bound the error between $f_{\hat{\theta}_n}$ and $\bar{f}$ where $\bar{f} := g \circ \varphi$ and the bound is expressed in terms of $\bar{f} - \Pi_{\mu,d} f$. Here for $u \in \mathbb{R}^d$, $g(u) = \int f(x)\mu(dx|u)$ where $\mu(dx|u)$ is the disintegration of μ with respect to the push-forward of μ under φ. In particular, for $S \subset \mathcal{X}$, $u \in \mathbb{R}^d$, $\mu(S|u) = \int \mathbb{I}(x \in S, \varphi(x) = u)\mu(dx)$. In the special case when φ is injective, $\bar{f} = f$. Just like in the result that can be extracted from the work of Lattimore et al. [2020], the bounds in this work depend on $\overline{\varphi}_{d,2}$ (or $\widehat{\varphi}_{d,2}$) and scale similarly.

B General-interest results

We start from the usual Bernstein's inequality Boucheron et al. [2003], here written for variables that are bounded in $[-B, B]$ and in the "high probability" form.

Theorem 16. *Let $\{X_t\}_{t=1}^n$ be a sequence of zero-mean random variable bounded in $[-B, B]$. Let $\sigma^2 := \sum_{t=1}^n Var(X_t)$. Then, with probability at least $1 - \delta$*

$$\left|\sum_{t=1}^{n} X_t\right| \leq \sqrt{2\sigma^2 \log(2/\delta)} + \frac{2B}{3}\log(2/\delta).$$

Lemma 2. *Let $\overline{\boldsymbol{\varphi}}_d$ be an orthonormal feature map w.r.t. ρ.*

$$\mathbb{E}_{x\sim\rho}\left[\overline{\boldsymbol{\varphi}}_d(x)\overline{\boldsymbol{\varphi}}_d(x)^\top\right] = I_d,$$

where I_d is the $d-$dimensional identity matrix.

Proof. In this proof, let us denote with $\boldsymbol{e}_i$, for $i = 1, \ldots d$, the standard basis of $\mathbb{R}^d$. By definition of outer product between two vectors we get what follows.

$$\begin{aligned}
\mathbb{E}_{x\sim\rho}\left[\overline{\boldsymbol{\varphi}}_d(x)\overline{\boldsymbol{\varphi}}_d(x)^\top\right] &= \mathbb{E}_{x\sim\rho}\left[\sum_{i=1}^{d}\sum_{j=1}^{d}\overline{\varphi}_i(x)\overline{\varphi}_j(x)\boldsymbol{e}_i\boldsymbol{e}_j^\top\right] \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\mathbb{E}_{x\sim\rho}\left[\overline{\varphi}_i(x)\overline{\varphi}_j(x)\right]\boldsymbol{e}_i\boldsymbol{e}_j^\top \\
&= \sum_{i=1}^{d}\sum_{j=1}^{d}\delta_{ij}\boldsymbol{e}_i\boldsymbol{e}_j^\top = I_d.
\end{aligned}$$

This completes the proof. □

Lemma 3. *Let $\{v_t\}_{t=1}^k$ be a sequence of independent $d-$dimensional random vectors such that*

$$\mathbb{E}[v_t v_t^\top] = \sigma I_d \qquad \|v_t\|_2^2 \leq B.$$

Let $V := \sum_{t=1}^k v_t v_t^\top$. Then,

1. *W.p. at least $1-\delta$*

$$\lambda_{\min}(V) \geq \left(1 - \sqrt{\frac{5B\log(d/\delta)}{k\sigma^2}}\right)k\sigma^2,$$

if $\left(1 - \sqrt{\frac{5B\log(d/\delta)}{k\sigma^2}}\right) \leq 1/2$.

2. *W.p. at least $1-\delta$*

$$\lambda_{\max}(V) \leq \left(1 + \sqrt{\frac{2B\log(d/\delta)}{k\sigma^2}}\right)k\sigma^2,$$

if $\left(1 + \sqrt{\frac{2B\log(d/\delta)}{k\sigma^2}}\right) \leq 1$.

Proof. Note that, as $\lambda_{\max}(v_t v_t^\top) = \|v_t\|_s^2 \leq B$, we can then apply Theorem 5.1.1 from Tropp et al. [2015] taking

$$\mu_{\min} = \mu_{\max} = k\sigma^2 \qquad L = B,$$

which ensures that

$$\forall\varepsilon \in (0,1),\ \mathbb{P}\left(\lambda_{\min}(V) \leq (1-\varepsilon)k\sigma^2\right) \leq d\left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{k\sigma^2/B},$$

while

$$\forall \varepsilon > 0,\ \mathbb{P}\left(\lambda_{\max}(V) \geq (1+\varepsilon)k\sigma^2\right) \leq d \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{k\sigma^2/B}.$$

The thesis is going to follow by just simplifying the previous expressions. We recall from elementary Taylor expansions that

$$\varepsilon < 0.5 \implies -\varepsilon - 4\varepsilon^2/5 \leq \log(1-\varepsilon) \leq -\varepsilon - \frac{\varepsilon^2}{2}.$$

and

$$\varepsilon < 1 \implies \varepsilon - \frac{\varepsilon^2}{2} \leq \log(1+\varepsilon) \leq \varepsilon - \frac{\varepsilon^2}{4}.$$

Therefore, we have, for $\varepsilon < 0.5$

$$\begin{aligned}\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}} &= \exp(-\varepsilon - (1-\varepsilon)\log(1-\varepsilon)) \\ &\leq \exp\left(-\varepsilon - (1-\varepsilon)(-\varepsilon - 4\varepsilon^2/5)\right) \\ &= \exp\left(-\varepsilon + \varepsilon - \varepsilon^2/5 + \mathcal{O}(\varepsilon^3))\right) \approx e^{-\varepsilon^2/5}.\end{aligned}$$

On the other side, for $\varepsilon \leq 1$,

$$\begin{aligned}\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}} &= \exp(\varepsilon - (1+\varepsilon)\log(1+\varepsilon)) \\ &\leq \exp(\varepsilon - (1+\varepsilon)(\varepsilon - \varepsilon^2/2)) \\ &= \exp(-\varepsilon^2/2 + \mathcal{O}(\varepsilon^3)) \approx e^{-\varepsilon^2/2}.\end{aligned}$$

This tells us that

$$\forall \varepsilon \in (0, 1/2),\ \mathbb{P}\left(\lambda_{\min}(V) \leq (1-\varepsilon)k\sigma^2\right) \leq d e^{-k\sigma^2\varepsilon^2/(5B)},$$

and

$$\forall \varepsilon \in (0, 1),\ \mathbb{P}\left(\lambda_{\max}(V) \geq (1+\varepsilon)k\sigma^2\right) \leq d e^{-k\sigma^2\varepsilon^2/(2B)}.$$

We can reformulate the previous results in the high-probability notation. Indeed, taking $\delta = d e^{-k\sigma^2\varepsilon^2/(5B)}$, we get

$$\varepsilon = \sqrt{\frac{5B\log(d/\delta)}{k\sigma^2}},$$

which entails that

$$\sqrt{\frac{5B\log(d/\delta)}{k\sigma^2}} \leq 1/2 \implies \mathbb{P}\left(\lambda_{\min}(V) \leq \left(1 - \sqrt{\frac{5B\log(d/\delta)}{k\sigma^2}}\right) k\sigma^2\right) \leq \delta.$$

Doing the same for the other result, we get

$$\sqrt{\frac{2B\log(d/\delta)}{k\sigma^2}} \leq 1 \implies \mathbb{P}\left(\lambda_{\max}(V) \geq \left(1 + \sqrt{\frac{2B\log(d/\delta)}{k\sigma^2}}\right) k\sigma^2\right) \leq \delta,$$

which completes the proof.

□

Proposition 17. *The Lebesgue constant satisfies* $\Lambda_{d,\mu} = \sup_{x\in\mathcal{X}} \int_{\mathcal{X}} \left|\sum_{i=1}^{d} \overline{\varphi}_i(z)\overline{\varphi}_i(x)\right| \, d\mu(z)$.

Proof. See Cheney [1966], chapter 4. □

C Proofs from section 3

C.1 Lower bound for LS

Recall that $\mathcal{F} = L^2(\mu) \cap L^\infty(\mathcal{X})$. Let $\Pi_\infty f = \arg\min_{g\in\mathcal{F}_d} \|f-g\|_\infty$ with ties broken arbitrarily. Note that Theorem 1.1 of Chapter 3 in the book of DeVore and Lorentz [1993] guarantees that at least one minimizer exists. (As discussed there, uniqueness may or may not hold.)

Lemma 4. *We have*

$$\sup_{f\in\mathcal{F}} \frac{\|\Pi_{d,\mu} f - f\|_\infty}{\mathcal{E}_\infty(f)} \geq \Lambda_{d,\mu} - 1\,.$$

Proof. By definition of Lebesgue constant, for every $\varepsilon > 0$ there is a function g such that

$$\|\Pi_{d,\mu} g\|_\infty \geq (\Lambda_{d,\mu} - \varepsilon)\|g\|_\infty.$$

Take $f = \Pi_\infty g - g$. We will use twice that for any $h \in \mathcal{F}$, $\|h|\_\infty = \|0-h\|_\infty \geq \inf_{u\in\mathcal{F}_d} \|u - h\|_\infty = \|\Pi_\infty h - h\|_\infty$. Now,

$$\begin{aligned}
\|\Pi_{d,\mu} f - f\|_\infty &= \|\Pi_{d,\mu}(\Pi_\infty g - g) - \Pi_\infty g + g\|_\infty \\
&= \|\Pi_\infty g - \Pi_{d,\mu} g - \Pi_\infty g + g\|_\infty \\
&= \|-\Pi_{d,\mu} g + g\|_\infty \\
&\geq \|\Pi_{d,\mu} g\|_\infty - \|g\|_\infty \\
&\geq (\Lambda_{d,\mu} - 1 - \varepsilon)\|g\|_\infty \\
&\geq (\Lambda_{d,\mu} - 1 - \varepsilon)\|\Pi_\infty g - g\|_\infty \\
&= (\Lambda_{d,\mu} - 1 - \varepsilon)\|f\|_\infty \\
&\geq (\Lambda_{d,\mu} - 1 - \varepsilon)\|\Pi_\infty f - f\|_\infty.
\end{aligned}$$

The result follows by letting $\varepsilon \to 0$. □

Theorem 3. *For any $\varepsilon > 0$ there exist $f \in L^\infty(\mathcal{X})$ such that*

$$\|f - \Pi_{d,\mu} f\|_\infty \geq (\Lambda_{d,\mu} - 1 - \varepsilon)\mathcal{E}_\infty(f)\,.$$

Proof. The result is immediate from Lemma 4. □

C.2 Towards the proof of theorem 4

Lemma 5. *Fix $\delta > 0$, and $n \geq 20\overline{\varphi}_{d,2}^2 \log(d/\delta)$. Let*

$$V_n = \sum_{t=1}^{n} \overline{\boldsymbol{\varphi}}_d(x_t)\overline{\boldsymbol{\varphi}}_d(x_t)^\top.$$

Then, $\lambda_{\min}(V_n) \geq n/2$.

Proof. The matrices we are summing correspond to $\overline{\boldsymbol{\varphi}}_d(x_t)\overline{\boldsymbol{\varphi}}_d(x_t)^\top$ each one being semi-positive definite with the biggest eigenvalue bounded by $\overline{\varphi}_{d,2}^2$ almost surely (indeed, $v^\top\overline{\boldsymbol{\varphi}}_d(x_t)\overline{\boldsymbol{\varphi}}_d(x_t)^\top v$ is maximized for v parallel to $\overline{\boldsymbol{\varphi}}_d(x_t)$ and produces $\|\overline{\boldsymbol{\varphi}}_d(x_t)\|_2^2$). Moreover, as we have seen in lemma 2,

$$\mathbb{E}\left[\sum_{t=1}^{n}\overline{\boldsymbol{\varphi}}_d(x_t)\overline{\boldsymbol{\varphi}}_d(x_t)^\top\right] = \sum_{t=1}^{n}\mathbb{E}\left[\overline{\boldsymbol{\varphi}}_d(x_t)\overline{\boldsymbol{\varphi}}_d(x_t)^\top\right] = nI_d.$$

These two ingredients allow us to apply lemma 3 part one, which ensures that with probability at least $1-\delta$

$$\lambda_{\min}(V_n) \geq \left(1-\sqrt{\frac{5\overline{\varphi}_{d,2}^2\log(d/\delta)}{n}}\right)n,$$

if $\left(1-\sqrt{\frac{5\overline{\varphi}_{d,2}^2\log(d/\delta)}{n}}\right) \leq 1/2$. Therefore, taking $n \geq 20\overline{\varphi}_{d,2}^2\log(d/\delta)$, we get $\lambda_{\min}(V_n) \geq n/2$, which completes the proof.

□

Lemma 6. *Let $\zeta_{d,\mu}(\cdot) := f(\cdot) - \Pi_{d,\mu}f(\cdot)$. With probability at least $1-\delta$,*

$$\left|\overline{\boldsymbol{\varphi}}_d(z)^\top V_n^{-1}\sum_{t=1}^{n}\overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)\right| \leq \frac{2\Lambda_{d,\mu}\mathcal{E}_\infty(f)\overline{\varphi}_{d,2}}{\sqrt{n}}\sqrt{\log(1/\delta)},$$

plus a lower-order term depending on n^{-1} which takes the form of $\widetilde{\mathcal{O}}\left(n^{-1}d^{1/2}\overline{\varphi}_{d,2}^2\Lambda_{d,\mu}\mathcal{E}_\infty(f) + n^{-3/2}d\overline{\varphi}_{d,2}^3\Lambda_{d,\mu}\mathcal{E}_\infty(f)\right)$.

Proof. We start rearranging the equation as follows

$$\begin{aligned}
\left|\overline{\boldsymbol{\varphi}}_d(z)^\top V_n^{-1}\sum_{t=1}^{n}\overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)\right| &= \left|\overline{\boldsymbol{\varphi}}_d(z)^\top\left(\frac{1}{n}V_n\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)\right| \\
&= \left|\overline{\boldsymbol{\varphi}}_d(z)^\top\left(I_d+\Delta_n\right)\frac{1}{n}\sum_{t=1}^{n}\overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)\right| \\
&\leq \left|\overline{\boldsymbol{\varphi}}_d(z)^\top\frac{1}{n}\sum_{t=1}^{n}\overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)\right| \\
&\quad + \left|\overline{\boldsymbol{\varphi}}_d(z)^\top\Delta_n\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\varphi}_d(x_t)\zeta_{d,\mu}(x_t)\right|.
\end{aligned}$$

For $\Delta_n := (V_n/n)^{-1} - I_d$. To bound both parts, we start by giving a result for $\frac{1}{n}\sum_{t=1}^{n}v^\top\overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)$ that holds for one fixed $v\in\mathbb{R}^d$. Indeed,

1. Every random variable $v^\top\overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)$ is bounded by $\|v\|_2\overline{\varphi}_{d,2}\Lambda_{d,\mu}\mathcal{E}_\infty(f)$ a.s.
2. The variance of the same random variable is

$$\begin{aligned}
\mathbb{E}_{x\sim\rho}[(v^\top\overline{\boldsymbol{\varphi}}_d(x)\zeta_{d,\mu}(x))^2] &= \mathbb{E}_{x\sim\rho}[\zeta_{d,\mu}(x)^2v^\top\overline{\boldsymbol{\varphi}}_d(x)^\top\overline{\boldsymbol{\varphi}}_d(x)v] \\
&\leq (\Lambda_{d,\mu}\mathcal{E}_\infty(f))^2v^\top\mathbb{E}_{x\sim\rho}[\overline{\boldsymbol{\varphi}}_d(x)^\top\overline{\boldsymbol{\varphi}}_d(x)]v \\
&= (\Lambda_{d,\mu}\mathcal{E}_\infty(f))^2v^\top I_dv \\
&= (\Lambda_{d,\mu}\mathcal{E}_\infty(f))^2\|v\|_2^2,
\end{aligned}$$

   the main step following from lemma 2.

So by Bernstein's inequality (theorem 16),

$$\frac{1}{n}\sum_{t=1}^{n}v^\top\overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) \leq \frac{2\Lambda_{d,\mu}\mathcal{E}_\infty(f)\|v\|_2}{\sqrt{n}}\sqrt{\log(1/\delta)} + \frac{2\|v\|_2\overline{\varphi}_{d,2}\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{3n}\log(1/\delta). \quad (9)$$

We can use the previous equation to bound both parts. For the first, we just take $v = \overline{\boldsymbol{\varphi}}_d(z)$, which respects $\|v\|_2 \le \overline{\varphi}_{d,2}$, in equation 9 and get

$$\left| \overline{\boldsymbol{\varphi}}_d(z)^\top \frac{1}{n}\sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) \right| \le \frac{2\Lambda_{d,\mu}\mathcal{E}_\infty(f)\overline{\varphi}_{d,2}}{\sqrt{n}}\sqrt{\log(1/\delta)} + \frac{2\overline{\varphi}_{d,2}\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{3n}\log(1/\delta).$$

Let us now focus on the second part. Indeed,

$$\left| \overline{\boldsymbol{\varphi}}_d(z)^\top \Delta_n \frac{1}{n}\sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) \right| \le \overline{\varphi}_{d,2}\|\Delta_n\|_2 \left\| \frac{1}{n}\sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) \right\|_2$$

Now, using lemma 3 as done in the proof of lemma 5, we have

$$\|\Delta_n\|_2 \le \overline{\varphi}_{d,2}\sqrt{\frac{5\log(d/\delta)}{n}},$$

while for the last part we can write

$$\begin{aligned}\left\| \frac{1}{n}\sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) \right\|_2 &= \sup_{\|v\|_2=1} \frac{1}{n}\sum_{t=1}^n v^\top \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) \\ &\le \sup_{\|v\|_2 \in B_d^{1/n}} \frac{1}{n}\sum_{t=1}^n v^\top \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) + \frac{\overline{\varphi}_{d,2}\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{n},\end{aligned}$$

where $B_d^{1/n}$ is a $1/n$ covering of the set of vectors such that $\|v\|_2 = 1$. It is well-known that we can choose $B_d^{1/n}$ so that $|B_d^{1/n}| \approx n^{-d}$, so that, making a union bound together with equation 9, we get

$$\left\| \frac{1}{n}\sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) \right\|_2 \le \frac{2\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{\sqrt{n}}\sqrt{d\log(1/\delta)} + \frac{2\overline{\varphi}_{d,2}\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{3n}\log(1/\delta) + \frac{\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{n}.$$

As a consequence,

$$\begin{aligned}&\left| \overline{\boldsymbol{\varphi}}_d(z)^\top \Delta_n \frac{1}{n}\sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t) \right| \\ &\quad \le \overline{\varphi}_{d,2}^2\sqrt{\frac{5\log(d/\delta)}{n}}\left(\frac{2\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{\sqrt{n}}\sqrt{d\log(1/\delta)} + \frac{2\overline{\varphi}_{d,2}d\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{3n}\log(1/\delta) + \frac{\Lambda_{d,\mu}\mathcal{E}_\infty(f)}{n}\right) \\ &\quad \le \widetilde{\mathcal{O}}\left(n^{-1}d^{1/2}\overline{\varphi}_{d,2}^2\Lambda_{d,\mu}\mathcal{E}_\infty(f) + n^{-3/2}d\overline{\varphi}_{d,2}^3\Lambda_{d,\mu}\mathcal{E}_\infty(f)\right).\end{aligned}$$

This completes the proof. □

C.3 Proof of theorem 4

Theorem 4. *Let $\mathcal{X}$ be finite. Let Assumptions 1 and 2 hold. Then, for any n positive integer and real $0 < \delta \le 1/3$ such that $n \ge 20\overline{\varphi}_{d,2}^2\log(d/\delta)$, letting $\hat{\theta}_{n,OLS}$ be the parameter vector returned by OLS, with probability at least $1 - 3\delta$,*

$$\begin{aligned}\mathcal{E}_\infty(\hat{\theta}_{n,OLS}, f) \le (1 + \Lambda_{d,\mu})\mathcal{E}_\infty(f) + 3(\sigma + \Lambda_{d,\mu}\mathcal{E}_\infty(f))\overline{\varphi}_{d,2}\sqrt{\frac{\log(|\mathcal{X}|/\delta)}{n}} \\ + \frac{poly(d, \overline{\varphi}_{d,2}, \Lambda_{d,\mu}\mathcal{E}_\infty(f))}{n}.\end{aligned}$$

Proof. In this proof, as before, we call Let $\zeta_{d,\mu}(\cdot) := f(\cdot) - \Pi_{d,\mu} f(\cdot)$.

Trough this proof we call $\widehat{\theta}_n$ the OLS estimator and $\widehat{f}_n(\cdot)$ the corresponding estimated function. We start making the following decomposition:

$$\begin{aligned} |\overline{\boldsymbol{\varphi}}_d(x)^\top \widehat{\theta}_n - f(x)| &\le |\overline{\boldsymbol{\varphi}}_d(x)^\top \widehat{\theta}_n - \Pi_{d,\mu} f(x)| + \|\Pi_{d,\mu} f - f\|_\infty \\ &\le |\overline{\boldsymbol{\varphi}}_d(x)^\top \widehat{\theta}_n - \Pi_{d,\mu} f(x)| + (1+\Lambda_{d,\mu})\mathcal{E}_\infty(f). \end{aligned}$$

To bound the first part, we let $\theta_\star$ be such that $\Pi_{d,\mu} f(\cdot) = \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta_\star$. By Assumption 1, the samples take the form $y_t = \overline{\boldsymbol{\varphi}}_d(x_t)^\top \theta_\star + \zeta_{d,\mu}(x_t) + \eta_t$, where $\{\eta_t\}_{t=1}^n$ is a family of independent $\sigma-$subgaussian random variables. By definition, letting $V_n = \sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t) \overline{\boldsymbol{\varphi}}_d(x_t)^\top$, the LS solution takes the form $\overline{\boldsymbol{\varphi}}_d(x_t)^\top \widehat{\theta}_n$, where

$$\begin{aligned} \widehat{\theta}_n &= {V_n}^{-1} \sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t) y_t \\ &= {V_n}^{-1} \sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)(\overline{\boldsymbol{\varphi}}_d(x_t)^\top \theta_\star + \eta_t + \zeta_{d,\mu}(x_t)) \\ &= \theta_\star + {V_n}^{-1} \sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)(\eta_t + \zeta_{d,\mu}(x_t)). \end{aligned}$$

Therefore, we have

$$|\overline{\boldsymbol{\varphi}}_d(x)^\top \widehat{\theta}_n - \Pi_{d,\mu} f(x)| \le \underbrace{\left|\overline{\boldsymbol{\varphi}}_d(x)^\top {V_n}^{-1} \sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\eta_t\right|}_{(I)} + \underbrace{\left|\overline{\boldsymbol{\varphi}}_d(x)^\top {V_n}^{-1} \sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)\right|}_{(II)}.$$

We are going to bound the two terms separately. First, let $E := \{\lambda_{\min}(V_n) \ge n/2\}$. From lemma 5, under the assumptions of this theorem, we have $\mathbb{P}(E) \ge 1 - \delta$.

(I) Since η_t are independent and $\sigma-$subgaussian, Lemma 5.4 and Theorem 5.3 from Lattimore and Szepesvári [2020] ensure that, with probability at least $1 - 2\delta$

$$\begin{aligned} \left|\overline{\boldsymbol{\varphi}}_d(x)^\top {V_n}^{-1} \sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\eta_t\right| &\le \sqrt{2\sigma^2 \sum_{t=1}^n \left(\overline{\boldsymbol{\varphi}}_d(x)^\top {V_n}^{-1} \overline{\boldsymbol{\varphi}}_d(x_t)\right)^2 \log(1/\delta)} \\ &= \sqrt{2\sigma^2 \|\overline{\boldsymbol{\varphi}}_d(x)\|_{{V_n}^{-1}}^2 \log(1/\delta)} \\ &= \sqrt{2\log(1/\delta)}\sigma \|\overline{\boldsymbol{\varphi}}_d(x)\|_{{V_n}^{-1}}. \end{aligned}$$

Moreover, if event E holds,

$$\|\overline{\boldsymbol{\varphi}}_d(x)\|_{{V_n}^{-1}} \le \frac{2\|\overline{\boldsymbol{\varphi}}_d(x)\|_2}{\sqrt{n}} \le \frac{2\overline{\varphi}_{d,2}}{\sqrt{n}},$$

so that the full term is bounded by $\sqrt{8\log(1/\delta)}\sigma\overline{\varphi}_{d,2} n^{-1/2}$.

(II) This term is bounded by lemma 6 which, with probability at least $1 - \delta$ gives

$$\left|\overline{\boldsymbol{\varphi}}_d(z)^\top V_n^{-1} \sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu}(x_t)\right| \le \frac{2\Lambda_{d,\mu}\mathcal{E}_\infty(f)\overline{\varphi}_{d,2}}{\sqrt{n}} \sqrt{\log(1/\delta)},$$

plus lower-order terms of the form $\frac{\mathrm{poly}(d, \overline{\varphi}_{d,2}, \Lambda_{d,\mu}\mathcal{E}_\infty(f))}{n}$.

Note that, thanks to lemma 5, event E holds with probability $1-\delta$ under the assumptions of this theorem. Moreover, imposing that both events in (I) and (II) verify, we get, with probability at least $1-3\delta$,

$$
\begin{aligned}
|\overline{\boldsymbol{\varphi}}_d(x)^\top \widehat{\theta}_n - f(x)|_\infty &\leq (1+\Lambda_{d,\mu})\mathcal{E}_\infty(f) + |\overline{\boldsymbol{\varphi}}_d(x)^\top \widehat{\theta}_n - \Pi_{d,\mu} f(x)| \\
&\leq (1+\Lambda_{d,\mu})\mathcal{E}_\infty(f) + \frac{3(\sigma + \Lambda_{d,\mu}\mathcal{E}_\infty(f))\overline{\varphi}_{d,2}}{\sqrt{n}}\sqrt{\log(1/\delta)}
\end{aligned}
$$

plus lower-order terms of the form $\frac{\text{poly}(d,\overline{\varphi}_{d,2},\Lambda_{d,\mu}\mathcal{E}_\infty(f))}{n}$. This completes the proof. □

C.4 Bound scaling with the empirical lebesgue constant

Theorem 5. *Let Assumption 1 hold. Then, for any fixed $\delta > 0$, with probability at least $1-\delta$,*

$$
\mathcal{E}_\infty(\widehat{\theta}_{n,\text{OLS}}) \leq (1+\Lambda_{d,\mu_n})\mathcal{E}_\infty(f) + \frac{\sigma\widehat{\varphi}_{d,2}\sqrt{2\log(2\mathcal{X}/\delta)}}{\sqrt{n}}.
$$

Proof. Let $\widehat{\theta}_n$ the estimator corresponding to $\widehat{\theta}_{n,\text{OLS}}$ in the parameterization of $\widehat{\boldsymbol{\varphi}}_d(\cdot)$, so that

$$
\widehat{\boldsymbol{\varphi}}_d(\cdot)^\top \widehat{\theta}_n = \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,\text{OLS}} =: \widehat{f}_n(\cdot).
$$

The following decomposition holds:

$$
\begin{aligned}
\|f(\cdot) - \widehat{f}_n(\cdot)\|_\infty &\leq \|f(\cdot) - \Pi_{d,\mu_n} f(\cdot)\|_\infty + \|\Pi_{d,\mu_n} f(\cdot) - \widehat{f}_n(\cdot)\|_\infty \\
&\leq (1+\widehat{\Lambda}_{d,\mu})\mathcal{E}_\infty(f) + \|\Pi_{d,\mu_n} f(\cdot) - \widehat{f}_n(\cdot)\|_\infty.
\end{aligned}
$$

Now, we focus on the second term. As done in the previous proof of theorem 4, we let $\theta_\star$ be such that $\Pi_{d,\mu_n} f(\cdot) = \widehat{\boldsymbol{\varphi}}_d(\cdot)^\top \theta_\star$ and $\zeta_{d,\mu_n}(\cdot) := f(\cdot) - \widehat{\boldsymbol{\varphi}}_d(\cdot)^\top \theta_\star$. In this way, our samples take the form $y_t = \widehat{\boldsymbol{\varphi}}_d(x_t)^\top \theta_\star + \zeta_{d,\mu_n}(x_t) + \eta_t$.

For any fixed $x \in \mathcal{X}$ we have

$$
\begin{aligned}
\widehat{f}_n(x) &= \widehat{\boldsymbol{\varphi}}_d(x)^\top \widehat{\theta}_n \\
&= \widehat{\boldsymbol{\varphi}}_d(x)^\top \frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_d(x_t) y_t \\
&= \widehat{\boldsymbol{\varphi}}_d(x)^\top \frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_d(x_t)(\widehat{\boldsymbol{\varphi}}_d(x_t)^\top \theta_\star + \zeta_{d,\mu_n}(x_t) + \eta_t) \\
&= \widehat{\boldsymbol{\varphi}}_d(x)^\top \theta_\star + \underbrace{\widehat{\boldsymbol{\varphi}}_d(x)^\top \frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu_n}(x_t)}_{(I)} + \underbrace{\widehat{\boldsymbol{\varphi}}_d(x)^\top \frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_d(x_t)\eta_t}_{(II)}.
\end{aligned}
$$

Here, the last passage is due to the fact that, being $\widehat{\boldsymbol{\varphi}}_d(\cdot)$ orthogonal w.r.t. $\mu_n(\cdot)$, it follows $\frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_d(x_t)\widehat{\boldsymbol{\varphi}}_d(x_t)^\top = I_d$. Now, we analyze the two terms (I) and (II) separately.

$$
\begin{aligned}
(I) &= \widehat{\boldsymbol{\varphi}}_d(x)^\top \frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_d(x_t)\zeta_{d,\mu_n}(x_t) \\
&= \widehat{\boldsymbol{\varphi}}_d(x)^\top \int_{\mathcal{X}} \widehat{\boldsymbol{\varphi}}_d(z)\zeta_{d,\mu_n}(z)\, d\mu_n(z) = \widehat{\boldsymbol{\varphi}}_d(x)^\top \mathbf{0} = 0.
\end{aligned}
$$

In fact, by definition of orthogonal projection, $\zeta_{d,\mu_n}(\cdot)$ is orthogonal in $L^2(\mu_n)$ to the span of $\widehat{\boldsymbol{\varphi}}_d(\cdot)$, so to each of its components in particular.

Let us look at the second term. Since η_t are independent and $\sigma-$subgaussian, Lemma 5.4 and Theorem 5.3 from Lattimore and Szepesvári [2020] ensure that, with probability at least $1-2\delta$

$$\begin{aligned}\left|\widehat{\varphi}_d(x)^\top n^{-1}\sum_{t=1}^n \widehat{\varphi}_d(x_t)\eta_t\right| &\le \sqrt{2\sigma^2 n^{-1}\sum_{t=1}^n \left(\widehat{\varphi}_d(x)^\top \widehat{\varphi}_d(x_t)\right)^2 \log(1/\delta)} \\ &= \sqrt{2\sigma^2 n^{-1}\|\widehat{\varphi}_d(x)\|_2^2 \log(1/\delta)} \\ &= \sqrt{2\log(1/\delta)}\sigma n^{-1/2}\|\widehat{\varphi}_d(x)\|_2.\end{aligned}$$

Where the second passage comes once again from the fact that $\frac{1}{n}\sum_{t=1}^n \widehat{\varphi}_d(x_t)\widehat{\varphi}_d(x_t)^\top = I_d$. This proves that (II) is bounded by $\sqrt{2\log(1/\delta)}\sigma n^{-1/2}\widehat{\varphi}_{2,d}$. Making a union bound over $x \in \mathcal{X}$, this entails w.p. $1-\delta$,

$$\sup_{x\in\mathcal{X}}\left|\widehat{\varphi}_d(x)^\top n^{-1}\sum_{t=1}^n \widehat{\varphi}_d(x_t)\eta_t\right| \le \sqrt{2\log(|\mathcal{X}|/\delta)}\sigma n^{-1/2}\widehat{\varphi}_{2,d}.$$

We have proved that

$$\begin{aligned}\mathcal{E}_\infty(\widehat{\theta}_{n,\mathrm{OLS}}) &= \|f(\cdot) - \widehat{f}_n(\cdot)\|_\infty \\ &\le (1+\widehat{\Lambda}_{d,\mu})\mathcal{E}_\infty(f) + \|\Pi_{d,\mu_n} f(\cdot) - \widehat{f}_n(\cdot)\|_\infty \\ &\le (1+\widehat{\Lambda}_{d,\mu})\mathcal{E}_\infty(f) + \sqrt{2\log(|\mathcal{X}|/\delta)}\sigma n^{-1/2}\widehat{\varphi}_{2,d}.\end{aligned}$$

□

C.5 Proofs from section 3.1

Proposition 18. *The Lebesgue constant is bounded by $\Lambda_{d,\mu} \le \overline{\varphi}_{d,2}$.*

Proof. Let $f \in L^\infty(\mathcal{X})$ with $\|f\|_\infty = 1$. We have, for any $x \in \mathcal{X}$,

$$\begin{aligned}|\Pi_{d,\mu} f(x)| &= \left|\sum_{i=1}^d \langle f, \overline{\varphi}_i\rangle \overline{\varphi}_i(x)\right| \\ &\le \sqrt{\sum_{i=1}^d \langle f, \overline{\varphi}_i\rangle^2 \sum_{i=1}^d \overline{\varphi}_i(x)^2} \\ &\le \sqrt{\|f\|_\mu^2 \|\overline{\varphi}_i(x)\|_2^2} \\ &\le \|f\|_\infty \sqrt{\|\overline{\varphi}_i(x)\|_2^2} \le \overline{\varphi}_{d,2},\end{aligned}$$

the last passage coming from the fact that as ρ is a probability measure, $\|f\|_\mu \le \|f\|_\infty$. The thesis follows taking the supremum on f, x. □

Proposition 19. *Let $\boldsymbol{\varphi}_d : \mathcal{X} \to \mathbb{R}^d$ be any feature map, and ρ a probability measure. Then,*

$$\overline{\varphi}_2 \ge \sqrt{d}.$$

Proof. The key for this result is to note that, being ρ a probability measure, $\overline{\varphi}_{d,2}^2 \geq \mathbb{E}_{x\sim\rho}\left[\|\overline{\boldsymbol{\varphi}}_d(x)\|_2^2\right]$ (the supremum of a function upper bounds its integral on any probability measure). Then,

$$\begin{aligned}\overline{\varphi}_{d,2} &\geq \sqrt{\mathbb{E}_{x\sim\rho}\left[\|\overline{\boldsymbol{\varphi}}_d(x)\|_2^2\right]}\\ &= \sqrt{\mathbb{E}_{x\sim\rho}\left[\overline{\boldsymbol{\varphi}}_d(x)^\top\overline{\boldsymbol{\varphi}}_d(x)\right]}\\ &= \sqrt{\mathbb{E}_{x\sim\rho}\left[\mathrm{Tr}(\overline{\boldsymbol{\varphi}}_d(x)^\top\overline{\boldsymbol{\varphi}}_d(x))\right]}\\ &= \sqrt{\mathbb{E}_{x\sim\rho}\left[\mathrm{Tr}(\overline{\boldsymbol{\varphi}}_d(x)\overline{\boldsymbol{\varphi}}_d(x)^\top)\right]}\\ &= \sqrt{\mathrm{Tr}(\mathbb{E}_{x\sim\rho}\left[\overline{\boldsymbol{\varphi}}_d(x)\overline{\boldsymbol{\varphi}}_d(x)^\top\right])}\\ &\overset{*}{=} \sqrt{\mathrm{Tr}(I_d)} = \sqrt{d}.\end{aligned}$$

Where the passage $(*)$ comes from lemma 2. ☐

Proposition 20. *Let $\mathcal{X} = [k]$ and $\varphi_i(j) = X_{ij}$, with all the X_{ij} being independent bounded zero-mean unit variance random variables. Then, if $d = \mathcal{O}(\sqrt{k})$, the feature map $\boldsymbol{\varphi}_d$, satisfies*

$$\Lambda_{d,\mu} = \mathcal{O}(\sqrt{d\log(k/\delta)})$$

with probability at least $1-\delta$. Moreover, $\mathbb{E}[\Lambda_{d,\mu}] \geq \Omega(\sqrt{d})$.

Proof. By convenience, we call $\Phi \in \mathbb{R}^{k\times d}$ the matrix having, as columns, the features of $\boldsymbol{\varphi}_d$. Precisely, the $i-$th column of Φ corresponds to φ_i. It is well-known that, in a finite dimensional space the orthogonal projection operator writes as

$$\Pi_{d,\mu} := \Phi(\Phi^\top\Phi)^{-1}\Phi^\top.$$

We call $\Phi_{m\cdot}$ the $m-$th row of Φ which, by assumption, is a random vector of independent entries bounded in $[-B, B]$ and with variance one. We have

$$\Phi^\top\Phi = \sum_{m=1}^{k}\Phi_{m\cdot}\Phi_{m\cdot}^\top, \qquad \mathbb{E}[\Phi_{m\cdot}\Phi_{m\cdot}^\top] = \sigma^2 I_d, \qquad \lambda_d(\Phi_{m\cdot}\Phi_{m\cdot}^\top) \leq dB^2.$$

At this point, we can apply lemma 3, that ensures with probability $1-2\delta$, for k sufficiently large,

$$\left(1-\sqrt{\frac{5dB^2\log(d/\delta)}{k\sigma^2}}\right)k\sigma^2 \leq \lambda_{\min}(\Phi^\top\Phi) \leq \lambda_{\max}(\Phi^\top\Phi) \leq \left(1+\sqrt{\frac{2dB^2\log(d/\delta)}{k\sigma^2}}\right)k\sigma^2.$$

Now, we can fix $\sigma = 1$ as in the assumption and rewrite the projection operator in the following form

$$\Pi_{d,\mu} := k^{-1}\Phi(k^{-1}\Phi^\top\Phi)^{-1}\Phi^\top = k^{-1}\Phi\Phi^\top + k^{-1}\Phi\Delta\Phi^\top,$$

where Δ has all the eigenvalues of magnitude less than $\sqrt{\frac{5dB^2\log(d/\delta)}{k\sigma^2}}$, by the previous result.

We now bound the infinity norm of the two terms separately. First,

$$\begin{aligned}\|k^{-1}\Phi\Phi^\top\|_\infty &\overset{*}{=} \frac{1}{k}\max_{m=1,\ldots k}\|(\Phi\Phi^\top)_{m\cdot}\|_1\\ &= \max_{m=1,\ldots k}\frac{1}{k}\sum_{n=1}^{k}\left|\sum_{i=1}^{d}\Phi_{mi}\Phi_{ni}\right|,\end{aligned}$$

where $*$ holds since the infinity norm of a matrix corresponds to the maximum $1-$norm between its rows. Now, note that, as the rows are independent, each variable $\sum_{i=1}^{d} \Phi_{mi}\Phi_{ni}$, for $m \neq n$ is a sum of i.i.d. random variables such that

- $\Phi_{mi}\Phi_{ni}$ is bounded in $[-B^2, B^2]$ almost surely.
- The variance is

$$\mathbb{E}[(\Phi_{mi}\Phi_{ni})^2] = \mathbb{E}[\Phi_{mi}^2\Phi_{ni}^2] = \mathbb{E}[\Phi_{mi}^2]\mathbb{E}[\Phi_{ni}^2] = 1.$$

Therefore, Bernstein's inequality 16 ensures that, w.p. $1-\delta$

$$\left|\sum_{i=1}^{d} \Phi_{mi}\Phi_{ni}\right| \leq \sqrt{2d\log(2/\delta)} + \frac{2B^2}{3}\log(2/\delta).$$

Making a union bound over the $k^2 - k$ pairs $m \neq n$, we get, still with probability at least $1-\delta$,

$$\forall n \neq m \qquad \left|\sum_{i=1}^{d} \Phi_{mi}\Phi_{ni}\right| \leq \sqrt{4d\log(2k/\delta)} + \frac{4B^2}{3}\log(2k/\delta). \tag{10}$$

At this point, we simply have, with probability $1-\delta$,

$$\begin{aligned}
\|k^{-1}\Phi\Phi^\top\|_\infty &= \max_{m=1,\dots k} \frac{1}{k}\sum_{n=1}^{k}\left|\sum_{i=1}^{d}\Phi_{mi}\Phi_{ni}\right| \\
&\leq \frac{dB^2}{k} + \max_{m=1,\dots k}\frac{1}{k}\sum_{n=1,n\neq m}^{k}\left|\sum_{i=1}^{d}\Phi_{mi}\Phi_{ni}\right| \\
&\overset{10}{\leq} \frac{dB^2}{k} + \max_{m=1,\dots k}\frac{1}{k}\sum_{n=1,n\neq m}^{k}\left(\sqrt{4d\log(2k/\delta)} + \frac{4B^2}{3}\log(2k/\delta)\right) \\
&= \sqrt{4d\log(2k/\delta)} + \frac{4B^2}{3}\log(2k/\delta) + \frac{dB^2}{k}.
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
\|k^{-1}\Phi\Delta\Phi^\top\|_\infty &\leq k^{-1}\max_{m=1,\dots k}\sum_{n=1}^{k}\left|\langle\Phi_{m\cdot}, (\Delta\Phi^\top)_{\cdot n}\rangle\right| \\
&\leq k^{-1}\max_{m=1,\dots k}\sum_{n=1}^{k}\|\Phi_{m\cdot}\|_2\|(\Delta\Phi^\top)_{\cdot n}\|_2 \\
&\overset{*}{\leq} k^{-1}\max_{m=1,\dots k}\sum_{n=1}^{k}\frac{dB^2}{\sqrt{k}} \\
&\leq \frac{dB^2}{\sqrt{k}},
\end{aligned}$$

where $*$ comes from the bound on the eigenvalues of Δ. Putting everything together, we have proved that

$$\|\Pi_{d,\mu}\|_\infty \leq \sqrt{4d\log(2k/\delta)} + \frac{4B^2}{3}\log(2k/\delta) + \frac{dB^2}{k} + \frac{dB^2}{\sqrt{k}} = \sqrt{4d\log(2k/\delta)} + \mathcal{O}(d/\sqrt{k}).$$

To show that we cannot go much lower than this quantity, note that, even ignoring the contribution of Δ we have

$$\|\Pi_{d,\mu}\|_\infty \approx \|k^{-1}\Phi\Phi^\top\|_\infty = \max_{m=1,\dots k} \frac{1}{k}\sum_{n=1}^{k}\left|\sum_{i=1}^{d}\Phi_{mi}\Phi_{ni}\right|.$$

Therefore,

$$\begin{aligned}\mathbb{E}[\|\Pi_{d,\mu}\|_\infty] &\approx \mathbb{E}\left[\max_{m=1,\dots k} \frac{1}{k}\sum_{n=1}^{k}\left|\sum_{i=1}^{d}\Phi_{mi}\Phi_{ni}\right|\right]\\ &\geq \max_{m=1,\dots k} \frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\left[\left|\sum_{i=1}^{d}\Phi_{mi}\Phi_{ni}\right|\right]\\ &\geq \max_m \frac{1}{k}\sum_{n=1,n\neq m}^{k}\Omega(\sqrt{d}) = \Omega(\sqrt{d}).\end{aligned}$$

The last passage comes from the fact that, for $n \neq m$, we have the expected value of the modulus a sum of d independent random variables, which grows as $\sqrt{d}$. □

Proposition 21. *Let μ, ν be two probability distributions on the discrete set $\mathcal{X}$ such that for all $x \in \mathcal{X}, C \geq \frac{\mu(x)}{\nu(x)} \geq c > 0$. Then, $\Lambda_{d,\mu} \leq \frac{C}{c}\Lambda_{d,\nu}$.*

Proof. The following identity holds for the Lebesgue constant

$$\begin{aligned}\Lambda_{d,\mu} &= \sup_{x\in\mathcal{X}} \int_{\mathcal{X}} \overline{\boldsymbol{\varphi}}_d(x)^\top \overline{\boldsymbol{\varphi}}_d(z)\, d\mu(z)\\ &= \sup_{x\in\mathcal{X}} \int_{\mathcal{X}} \boldsymbol{\varphi}_d(x)^\top R(\mu)^{-1}R(\mu)^{-\top}\, d\mu(z)\\ &= \sup_{x\in\mathcal{X}} \int_{\mathcal{X}} |\boldsymbol{\varphi}_d(x)^\top G(\mu)^{-1}\boldsymbol{\varphi}_d(z)|\, d\mu(z),\end{aligned}$$

where $G(\mu) = \int_{\mathcal{X}} \boldsymbol{\varphi}_d(x)\boldsymbol{\varphi}_d(x)^\top\, d\mu(x)$ and $R(\mu)$ is its Cholesky factor, such that $R(\mu)^\top R(\mu) = G(\mu)$; here, the second passage comes from the fact that the Cholesky factor of a matrix corresponds to the R factor in the QR factorization, which is the one giving Graham-Schmidt orthogonalization Quarteroni et al. [2010]. In fact, letting $\overline{\boldsymbol{\varphi}}_d(x)$ be the basis orthonomalized w.r.t. μ, we have

$$\overline{\boldsymbol{\varphi}}_d(x)^\top \overline{\boldsymbol{\varphi}}_d(z)^\top = \boldsymbol{\varphi}_d(x)^\top G(\mu)^{-1}\boldsymbol{\varphi}_d(z).$$

Note that, by absolute continuity, we have, for any $x \in \mathcal{X}$

$$\begin{aligned}\int_{\mathcal{X}} |\boldsymbol{\varphi}_d(x)^\top G(\mu)^{-1}\boldsymbol{\varphi}_d(z)|\, d\mu(z) &\leq C\int_{\mathcal{X}} |\boldsymbol{\varphi}_d(x)^\top G(\mu)^{-1}\boldsymbol{\varphi}_d(z)|\, d\nu(z)\\ &\leq C\int_{\mathcal{X}} \left|\boldsymbol{\varphi}_d(x)^\top \left(\int_{\mathcal{X}} \boldsymbol{\varphi}_d(z')\boldsymbol{\varphi}_d(z')^\top\, d\mu(z')\right)^{-1}\boldsymbol{\varphi}_d(z)\right|\, d\nu(z)\\ &\leq C\int_{\mathcal{X}} \left|\boldsymbol{\varphi}_d(x)^\top c^{-1} \left(\int_{\mathcal{X}} \boldsymbol{\varphi}_d(z')\boldsymbol{\varphi}_d(z')^\top\, d\nu(z')\right)^{-1}\boldsymbol{\varphi}_d(z)\right|\, d\nu(z)\\ &= \frac{C}{c}\int_{\mathcal{X}} \left|\boldsymbol{\varphi}_d(x)^\top \left(\int_{\mathcal{X}} \boldsymbol{\varphi}_d(z')\boldsymbol{\varphi}_d(z')^\top\, d\nu(z')\right)^{-1}\boldsymbol{\varphi}_d(z)\right|\, d\nu(z)\\ &= \frac{C}{c}\int_{\mathcal{X}} \left|\boldsymbol{\varphi}_d(x)^\top G(\nu)^{-1}\boldsymbol{\varphi}_d(z)\right|\, d\nu(z).\end{aligned}$$

Passing to the supremum, we get the thesis. □

D Proofs from section 4

D.1 Lower bound for standard ridge regression

Lemma 7. *Let $\Pi^{\lambda}_{d,\mu}$ be the operator defined in this way:*

$$\Pi^{\lambda}_{d,\mu} f := \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta_\lambda \qquad \theta_\lambda = \arg\min_\theta \|f(\cdot) - \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta\|^2_{L^2} + \lambda \|\theta\|^2_2. \tag{11}$$

Then, we have

$$\Pi^{\lambda}_{d,\mu} f = \frac{\Pi_{d,\mu} f}{1+\lambda}.$$

Proof. We start from the definition of θ_λ:

$$\begin{aligned}
\theta_\lambda &= \arg\min_\theta \|f(\cdot) - \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta\|^2_{L^2} + \lambda \|\theta\|^2_2 \\
&= \arg\min_\theta \|\Pi_{d,\mu} f(\cdot) + \zeta_{d,\mu}(\cdot) - \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta\|^2_{L^2} + \lambda \|\theta\|^2_2 \\
&= \arg\min_\theta \|\zeta_{d,\mu}\|^2_{L^2} + \|\Pi_{d,\mu} f(\cdot) - \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta\|^2_{L^2} + \lambda \|\theta\|^2_2,
\end{aligned}$$

where the last passage comes from Parseval's theorem, as $\zeta_{d,\mu}$ is orthogonal in L^2 to the span of $\boldsymbol{\varphi}_d$, while $\Pi_{d,\mu} f(\cdot), \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta$ belongs to this vector space. We then write the operator $\Pi_{d,\mu} f$ explicitly:

$$\begin{aligned}
\theta_\lambda &= \arg\min_\theta \|\Pi_{d,\mu} f(\cdot) - \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta\|^2_{L^2} + \lambda \|\theta\|^2_2 \\
&= \arg\min_\theta \left\| \sum_{i=1}^d \langle f, \overline{\varphi}_i \rangle_{L^2} \overline{\varphi}_i(\cdot) - \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta \right\|^2_{L^2} + \lambda \|\theta\|^2_2 \\
&= \arg\min_\theta \sum_{i=1}^d (\langle f, \overline{\varphi}_i \rangle_{L^2} - \theta_i)^2 + \lambda \theta_i^2.
\end{aligned}$$

The last passage holds from Parseval's theorem since $\overline{\varphi}_i$ are orthonormal in L^2. Note that, as the θ_i in the last minimization problem are disentangled, we can find as explicit solution

$$\theta_{\lambda,i} = \frac{\langle f, \overline{\varphi}_i \rangle_{L^2}}{1+\lambda}, \qquad \Pi^{\lambda}_{d,\mu} f = \frac{\Pi_{d,\mu} f}{1+\lambda}.$$

This completes the proof. □

Lemma 8. *Let $\Pi^{\lambda}_{d,\mu}$ be defined according to equation 11. For every feature map $\boldsymbol{\varphi}_d$ we have*

$$\sup_{f \in L^\infty(\mathcal{X})} \frac{\|\Pi^{\lambda}_{d,\mu} f - f\|_\infty}{\|\Pi_\infty f - f\|_\infty} \geq \left(\frac{\Lambda_{d,\mu} - 1 - 2\lambda}{1+\lambda} \right).$$

Proof. By definition of Lebesgue constant, for every $\varepsilon > 0$ there is a function g such that

$$\|\Pi_{d,\mu} g\|_\infty = (\Lambda_{d,\mu} - \varepsilon)\|g\|_\infty.$$

Take $f = \Pi_\infty g - g$. We have, by lemma 7,

$$
\begin{aligned}
\|\Pi^{\lambda}_{d,\mu} f - f\|_\infty &= \left\| \frac{\Pi_{d,\mu} f}{1+\lambda} - f \right\|_\infty \\
&= \|(1+\lambda)^{-1}\Pi_{d,\mu}(P^d_\infty g - g) - \Pi_\infty g + g\|_\infty \\
&= \|(1+\lambda)^{-1}\Pi_\infty g - (1+\lambda)^{-1}\Pi_{d,\mu} g - P^d_\infty g + g\|_\infty \\
&= \left\| -(1+\lambda)^{-1}\Pi_{d,\mu} g - \frac{\lambda}{1+\lambda} P^d_\infty g + g \right\|_\infty .
\end{aligned}
$$

At this point, note that

$$
\|\Pi_\infty g\|_\infty \leq 2\|g\|_\infty,
$$

as follows from

$$
\begin{aligned}
\|\Pi_\infty g\|_\infty &\leq \|g - \Pi_\infty g\|_\infty + \|g\|_\infty \\
&\leq \|g - 0\|_\infty + \|g\|_\infty = 2\|g\|_\infty .
\end{aligned}
$$

Using this property, we have

$$
\begin{aligned}
\|\Pi^{\lambda}_{d,\mu} f - f\|_\infty &\geq \left\| -(1+\lambda)^{-1}\Pi_{d,\mu} g - \frac{\lambda}{1+\lambda}\Pi_\infty g + g \right\|_\infty \\
&\geq \| - (1+\lambda)^{-1}\Pi_{d,\mu} g\|_\infty - \frac{1+2\lambda}{1+\lambda}\|g\|_\infty .
\end{aligned}
$$

At this point, using the definition of g,

$$
\begin{aligned}
\| - (1+\lambda)^{-1}\Pi_{d,\mu} g\|_\infty - \frac{1+2\lambda}{1+\lambda}\|g\|_\infty &\geq \left(\frac{\Lambda_{d,\mu}}{1+\lambda} - \varepsilon - \frac{1+2\lambda}{1+\lambda} \right) \|\Pi_\infty g - g\|_\infty \\
&= \left(\frac{\Lambda_{d,\mu}}{1+\lambda} - \varepsilon - \frac{1+2\lambda}{1+\lambda} \right) \|f\|_\infty \\
&\geq \left(\frac{\Lambda_{d,\mu}}{1+\lambda} - \varepsilon - \frac{1+2\lambda}{1+\lambda} \right) \|\Pi_\infty f - f\|_\infty .
\end{aligned}
$$

The thesis follows letting $\varepsilon \to 0$. □

Theorem 7. *Let $\widehat{\theta}_{n,\textsc{Ridge}}$ the output of $\lambda-$ridge regression. For any feature map $\boldsymbol{\varphi}_d(\cdot) : \mathcal{X} \to \mathbb{R}^d$ there is $f \in L^\infty(\mathcal{X})$ such that, for infinite data $\mathcal{E}_\infty(\widehat{\theta}_{\infty,\textsc{Ridge}}) = \Omega\left(\max\left\{(\Lambda_{d,\mu} - 2\lambda)\mathcal{E}_\infty(f), \frac{\lambda}{\lambda+1}\right\}\right)$.*

Proof. Let $\widehat{f}_n$ be the output of $\lambda-$ridge regression, that is the function $\overline{\boldsymbol{\varphi}}_d(\cdot)^\top \widehat{\theta}_n$, where

$$
\widehat{\theta}_n := \arg\min_{\theta \in \mathbb{R}^d} \sum_{t=1}^{n} (\overline{\boldsymbol{\varphi}}_d(x_t)^\top \theta - y_t)^2 + \lambda n \|\theta\|_2^2 \qquad x_t \overset{i.i.d.}{\sim} \mu.
$$

By the uniform law of large numbers, in the limit, the minimizer $\widehat{f}_n$ converges to $\Pi^{\lambda}_{d,\mu} f$, the regularized projection operator is defined as follows

$$
\Pi^{\lambda}_{d,\mu} f\,(\cdot) := \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta_\lambda \qquad \theta_\lambda = \arg\min_{\theta} \|f(\cdot) - \overline{\boldsymbol{\varphi}}_d(\cdot)^\top \theta\|^2_{L^2} + \lambda \|\theta\|_2^2 .
$$

We start showing the $\frac{\lambda}{\lambda+1}$ lower bound. Taking any function in the span of $\boldsymbol{\varphi}_d(\cdot)$ with $\|f\|_\infty = 1$ we have, by lemma 7,

$$
\|f - \Pi^{\lambda}_{d,\mu} f\|_\infty = \|f - (1+\lambda)^{-1} f\|_\infty = \frac{\lambda}{\lambda+1} .
$$

To show the other part, use lemma 8 to define a function f such that

$$\|\Pi_{d,\mu}^{\lambda} f - f\|_\infty \geq \left(\frac{\Lambda_{d,\mu} - 2 - 2\lambda}{1+\lambda}\right) \|\Pi_\infty f - f\|_\infty.$$

Replacing $\|\Pi_\infty f - f\|_\infty = \mathcal{E}_\infty(f)$ completes the proof. □

D.2 Proofs from section 4.1

Proposition 22. *Let $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ (equation (5)) and $\Pi_{\boldsymbol{\alpha},\mu}^{\textit{Ridge}}$ be defined according to equation (4). Then,*

$$\|f(\cdot) - \Pi_{\boldsymbol{\alpha},\mu}^{\textit{Ridge}} f(\cdot)\|_\infty \leq (1 + \Lambda_{\boldsymbol{\alpha},\mu})\mathcal{E}_\infty(f).$$

Proof. We have

$$\begin{aligned}
\|f - P_{\boldsymbol{h}}^2 f\|_\infty &= \|\Pi_{d,\infty} f + \xi_d - P_{\boldsymbol{h}}^2[\Pi_{d,\infty} f + \xi_d]\|_\infty \\
&\stackrel{*}{=} \|\xi_d - P_{\boldsymbol{h}}^2[\xi_d]\|_\infty \\
&\leq \|\xi_d\|_\infty + \|P_{\boldsymbol{h}}^2[\xi]\|_\infty \\
&= \|\xi_d\|_\infty + \Lambda_{\boldsymbol{h}}\|\xi_d\|_\infty.
\end{aligned}$$

Here, the key passage (*) holds since $P_{\boldsymbol{h}}^2$ by definition **??** is the identity over everything in the span of the first d features, so $\Pi_{d,\infty} f$ in particular. □

Theorem 9. *Let assumption 1 hold. Then, for any $\delta > 0$, with probability $1 - \delta$,*

$$\mathcal{E}_\infty(\widehat{\theta}_{n,\boldsymbol{\alpha}}) \leq (1 + \Lambda_{\boldsymbol{\alpha},\mu_n})\mathcal{E}_\infty(f) + \frac{\sigma \widehat{\varphi}_{2,D}\sqrt{2\log(2\mathcal{X}/\delta)}}{\sqrt{n}}.$$

Proof. Let $\widehat{\theta}_n$ the estimator corresponding to $P_{\boldsymbol{\alpha}}^2$ in the parameterization of $\widehat{\boldsymbol{\varphi}}_D(\cdot)$, so that

$$\widehat{\boldsymbol{\varphi}}_d(\cdot)^\top \widehat{\theta}_n = \widehat{P}_{\boldsymbol{\alpha}}^2 \mathbf{f} =: \widehat{f}_n(\cdot).$$

The following decomposition holds:

$$\begin{aligned}
\|f(\cdot) - \widehat{f}_n(\cdot)\|_\infty &\leq \|f(\cdot) - \widehat{P}_{\boldsymbol{\alpha}}^2 f(\cdot)\|_\infty + \|\widehat{P}_{\boldsymbol{\alpha}}^2 f(\cdot) - \widehat{f}_n(\cdot)\|_\infty \\
&\leq (1 + \widehat{\Lambda}_{\boldsymbol{\alpha}})\mathcal{E}_\infty(f) + \|\widehat{P}_{\boldsymbol{\alpha}}^2 f(\cdot) - \widehat{f}_n(\cdot)\|_\infty.
\end{aligned}$$

where we have applied proposition 8 for $\mu_n(\cdot)$. Let us focus on the second term. As in the proof of the previous theorems, we call $\theta_\star$ the vector corresponding to the orthogonal projection over $\widehat{\boldsymbol{\varphi}}_D(\cdot)$ so that we have, for every $x \in \mathcal{X}$

$$\begin{aligned}
\widehat{f}_n(x) &= \widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{t=1}^n y_t \widehat{\boldsymbol{\varphi}}_D(x_t) \\
&= \widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{t=1}^n (\widehat{\boldsymbol{\varphi}}_D(x_t)^\top \theta_\star + \zeta_D(x_t) + \eta_t)\widehat{\boldsymbol{\varphi}}_D(x_t) \\
&= \widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_D(x_t)\widehat{\boldsymbol{\varphi}}_D(x_t)^\top \theta_\star \\
&\quad + \widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{t=1}^n \zeta_D(x_t)\widehat{\boldsymbol{\varphi}}_D(x_t) \\
&\quad + \widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{t=1}^n \eta_t \widehat{\boldsymbol{\varphi}}_D(x_t).
\end{aligned}$$

By orthogonality, the first term corresponds to

$$\widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \underbrace{\frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_D(x_t)\widehat{\boldsymbol{\varphi}}_D(x_t)^\top}_{I_D} \theta_\star = \widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}}\theta_\star = P^2_{\boldsymbol{\alpha}}(x).$$

The second term is

$$\widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{t=1}^n \zeta_D(x_t)\widehat{\boldsymbol{\varphi}}_D(x_t) = \widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \underbrace{\int_{\mathcal{X}} \zeta_D(z)\widehat{\boldsymbol{\varphi}}_D(z)\, d\mu_n(z)}_{\mathbf{0}\text{ vector}} = 0,$$

by definition of orthogonal projection. The third term is

$$\widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \frac{1}{n}\sum_{t=1}^n \eta_t \widehat{\boldsymbol{\varphi}}_D(x_t),$$

which can be bounded as the corresponding terms in theorems 4 and 5: as η_t are independent and $\sigma-$subgaussian, Lemma 5.4 and Theorem 5.3 from Lattimore and Szepesvári [2020] ensure that, with probability at least $1-2\delta$

$$\begin{aligned}\left|\widehat{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} n^{-1}\sum_{t=1}^n \widehat{\boldsymbol{\varphi}}_d(x_t)\eta_t\right| &\le \sqrt{2\sigma^2 n^{-1}\sum_{t=1}^n \left(\widehat{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}\widehat{\boldsymbol{\varphi}}_d(x_t)\right)^2 \log(1/\delta)} \\ &= \sqrt{2\sigma^2 n^{-1}\|\widehat{\boldsymbol{\varphi}}_d(x)\|_2^2 \log(1/\delta)} \\ &= \sqrt{2\log(1/\delta)}\sigma n^{-1/2}\|\widehat{\boldsymbol{\varphi}}_d(x)\|_2.\end{aligned}$$

Where the only difference w.r.t. the other proofs is the presence of $I_{\boldsymbol{\alpha}}$, which is erased after the first step since, being $\boldsymbol{\alpha} \in \mathcal{A}_d^D$, its norm is ≤ 1. This proves that the last term is bounded by $\sqrt{2\log(1/\delta)}\sigma n^{-1/2}\widehat{\varphi}_{2,D}$. Making a union bound over $\mathcal{X}$ gives, w.p. $1-\delta$,

$$\sup_{x\in\mathcal{X}} |\widehat{P}^2_{\boldsymbol{\alpha}} f(x) - \widehat{f}_n(x)| \le \sqrt{2\log(1/\delta)}\sigma n^{-1/2}\widehat{\varphi}_{2,D}.$$

Putting everything together, we have proved that

$$\mathcal{E}_\infty(\widehat{\theta}_{n,\boldsymbol{\alpha}}) \le \|f(\cdot) - \widehat{f}_n(\cdot)\|_\infty \le (1+\widehat{\Lambda}_{\boldsymbol{\alpha}})\mathcal{E}_\infty(f) + \frac{\sigma\widehat{\varphi}_{2,D}\sqrt{2\log(2\mathcal{X}/\delta)}}{\sqrt{n}}.$$

□

Proposition 23. *Under assumption 2 we have, with probability* $1-\delta$ *for every* $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ *at the same time,* $|\widehat{\varphi}_{D,2} - \overline{\varphi}_{D,2}| \le \widetilde{\mathcal{O}}(\overline{\varphi}^2_{D,2}\sqrt{\log(1/\delta)/n})$, *and* $|\Lambda_{\boldsymbol{\alpha},\mu_n} - \Lambda_{\boldsymbol{\alpha},\mu}| \le \widetilde{\mathcal{O}}\left(\frac{\sqrt{d}\overline{\varphi}^2_{D,2}\sqrt{\log(1/\delta)}}{\sqrt{n}} + \frac{\sqrt{d}\overline{\varphi}^3_{D,2}\log(1/\delta)}{n}\right)$.[2]

We prove this theorem for a generic $d \in \mathbb{N}$. The result follows for $d = D$.

We define $V_n := \frac{1}{n}\sum_{t=1}^n \overline{\boldsymbol{\varphi}}_d(x_t)\overline{\boldsymbol{\varphi}}_d(x_t)^\top$. Let $\widehat{\boldsymbol{\varphi}}_d(\cdot)$ the basis obtained from $\boldsymbol{\varphi}_d$ by Gram-Schmidt orthogonalization w.r.t. μ_n, the empirical distribution of the $\{x_t\}_t$. As in the main paper, we let $R_n = \mathrm{Chol}(V_n)$ and, since the Cholesky factor corresponds to the matrix given by Graham Schmidt orthogonalization (proposition 3.4 in Quarteroni et al. [2010]),

$$\overline{\boldsymbol{\varphi}}_d(x_t) = R_n^\top \widehat{\boldsymbol{\varphi}}_d(x_t) \qquad \widehat{\boldsymbol{\varphi}}_d(x_t) = R_n^{-\top}\overline{\boldsymbol{\varphi}}_d(x_t). \tag{12}$$

[2] The statement of this theorem is slightly different from the one in the main paper of the submission, as we have made the orders of magnitude more precise.

so that, under this convenient normalization, we can pass from $\overline{\varphi}_d(x_t)$ to $\widehat{\varphi}_d(x_t)$ trough a matrix that is exactly the Cholesky factor of V_n. In this setting, Theorem 2.1. in Drmač et al. [1994], which provides a stability result for the Cholesky decomposition which, combined with our theorem gives

$$1 - \mathcal{O}\left(\overline{\varphi}_{d,2}\sqrt{\log(1/\delta)/n}\log(d)\right) \leq \lambda_{\min}(R_n) \leq \lambda_{\max}(R_n) \leq 1 + \mathcal{O}\left(\overline{\varphi}_{d,2}\sqrt{\log(1/\delta)/n}\log(d)\right) \tag{13}$$

We can now proceed with the proof.

Proof. **Bounding norm difference**

We have to measure

$$\sup_{x\in\mathcal{X}} \|\widehat{\varphi}_d(x) - \overline{\varphi}_d(x)\|_2.$$

As we said, the relation between the two is $\overline{\varphi}_d(x) = R_n^\top \widehat{\varphi}_d(x)$ which we can also wite as $R_n^{-\top}\overline{\varphi}_d(x) = \widehat{\varphi}_d(x)$, so that

$$\sup_{x\in\mathcal{X}} \|\widehat{\varphi}_d(x) - \overline{\varphi}_d(x)\|_2 = \sup_{x\in\mathcal{X}} \|(I_d - R_n^{-\top})\overline{\varphi}_d(x)\|_2.$$

At this point, equation (13) ensures that $\|I_d - R_n^{-\top}\|_{2\to 2} = \mathcal{O}\left(\overline{\varphi}_{d,2}\sqrt{\log(1/\delta)/n}\log(d)\right)$, so we get

$$\sup_{x\in\mathcal{X}} \|\widehat{\varphi}_d(x) - \overline{\varphi}_d(x)\|_2 \leq \mathcal{O}\left(\overline{\varphi}_{d,2}^2\sqrt{\log(1/\delta)/n}\log(d)\right). \tag{14}$$

A simple yet useful consequence of this result is

$$\begin{aligned}
|\overline{\varphi}_{d,2} - \widehat{\varphi}_{d,2}| &= \sup_{x\in\mathcal{X}} |\|\widehat{\varphi}_d(x)\|_2 - \sup_{x\in\mathcal{X}} \|\overline{\varphi}_d(x)\|_2| && (15)\\
&\leq \sup_{x\in\mathcal{X}} |\|\widehat{\varphi}_d(x)\|_2 - \|\overline{\varphi}_d(x)\|_2| && (16)\\
&\leq \sup_{x\in\mathcal{X}} \|\widehat{\varphi}_d(x) - \overline{\varphi}_d(x)\|_2 && (17)\\
&= \mathcal{O}\left(\overline{\varphi}_{d,2}^2\sqrt{\log(1/\delta)/n}\log(d)\right) && (18)
\end{aligned}$$

Lebesgue constants difference

Let us bound the distance between the estimated and the true Lebesgue constant, for any $\boldsymbol{\alpha} \in \mathcal{A}_d^D$,

$$
\begin{aligned}
|\Lambda_{\boldsymbol{\alpha},\mu_n} - \Lambda_{\boldsymbol{\alpha},\mu}| &= \left| \sup_{x\in\mathcal{X}} \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t) \right| - \sup_{x\in\mathcal{X}} \int_{\mathcal{X}} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(z) \right| d\mu(z) \right| \\
&\leq \sup_{x\in\mathcal{X}} \left| \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t) \right| - \int_{\mathcal{X}} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(z) \right| d\mu(z) \right| \\
&\leq \sup_{x\in\mathcal{X}} \left| \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t) \right| - \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(x_t) \right| \right| \\
&\quad + \sup_{x\in\mathcal{X}} \left| \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(x_t) \right| - \int_{\mathcal{X}} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(z) \right| d\mu(z) \right| \\
&= \sup_{x\in\mathcal{X}} \left| \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t) \right| - \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(x_t) \right| \right| \\
&\quad + \sup_{x\in\mathcal{X}} \left| \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(x_t) \right| - \int_{\mathcal{X}} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(z) \right| d\mu(z) \right|.
\end{aligned}
$$

In the following, we call

$$
\text{First term} := \sup_{\boldsymbol{\alpha}\in\mathcal{A}_d^D, x\in\mathcal{X}} \left| \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t) \right| - \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(x_t) \right| \right|
$$

and

$$
\text{Second term} := \sup_{\boldsymbol{\alpha}\in\mathcal{A}_d^D, x\in\mathcal{X}} \left| \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(x_t) \right| - \int_{\mathcal{X}} \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(z) \right| d\mu(z) \right|.
$$

Bound the first term.

Fix $\boldsymbol{\alpha} \in \mathcal{A}_d^D$,

$$
\begin{aligned}
\text{First part} &= \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t) \right| - \left| \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(x_t) \right| \\
&\leq \frac{1}{n}\sum_{t=1}^{n} \left| \sum_{i=1}^{d} \alpha_i \widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t) - \sum_{i=1}^{d} \alpha_i \overline{\varphi}_i(x)\overline{\varphi}_i(x_t) \right| \\
&= \frac{1}{n}\sum_{t=1}^{n} \left| \widehat{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \widehat{\boldsymbol{\varphi}}_d(x_t) - \overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t) \right|.
\end{aligned}
$$

Where, $I_{\boldsymbol{\alpha}} = \text{diag}(\boldsymbol{\alpha})$. At this point, we can replace the result of equation 12: getting

$$
\begin{aligned}
\text{First part} &\le \frac{1}{n}\sum_{t=1}^{n}\left|\widehat{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}\widehat{\boldsymbol{\varphi}}_d(x_t) - \overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}\overline{\boldsymbol{\varphi}}_d(x_t)\right| \\
&= \frac{1}{n}\sum_{t=1}^{n}\left|\widehat{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}\widehat{\boldsymbol{\varphi}}_d(x_t) - \widehat{\boldsymbol{\varphi}}_d(x)^\top R_n I_{\boldsymbol{\alpha}} R_n^\top \widehat{\boldsymbol{\varphi}}_d(x_t)\right| \\
&\le \frac{1}{n}\sum_{t=1}^{n}\left|\widehat{\boldsymbol{\varphi}}_d(x)^\top (I_{\boldsymbol{\alpha}} - R_n I_{\boldsymbol{\alpha}} R_n^\top)\widehat{\boldsymbol{\varphi}}_d(x_t)\right| \\
&\le \frac{1}{n}\sum_{t=1}^{n}\|\widehat{\boldsymbol{\varphi}}_d(x)\|_2 \|I_{\boldsymbol{\alpha}} - R_n I_{\boldsymbol{\alpha}} R_n^\top\|_2 \|\widehat{\boldsymbol{\varphi}}_d(x_t)\|_2.
\end{aligned}
$$

This formulation allows us to apply equation 13: As $I_{\boldsymbol{\alpha}}$ is diagonal matrix with elements in $[0,1]$, we have

$$
\|I_{\boldsymbol{\alpha}} - R_n I_{\boldsymbol{\alpha}} R_n^\top\|_2 = \mathcal{O}\left(\overline{\varphi}_{d,2}\sqrt{\log(1/\delta)/n}\log(d)\right).
$$

This gives the following

$$
\begin{aligned}
\text{First part} &\le \mathcal{O}\left(\frac{1}{n}\sum_{t=1}^{n}\|\widehat{\boldsymbol{\varphi}}_d(x)\|_2\|H - R_n H R_n^\top\|_2\|\widehat{\boldsymbol{\varphi}}_d(x_t)\|_2\right) \\
&\le \mathcal{O}\left(\frac{\overline{\varphi}_{d,2}\sqrt{\log(1/\delta)}\log(d)}{\sqrt{n}}\frac{1}{n}\sum_{t=1}^{n}\|\widehat{\boldsymbol{\varphi}}_d(x)\|_2\|\widehat{\boldsymbol{\varphi}}_d(x_t)\|_2\right) \\
&\le \mathcal{O}\left(\frac{\overline{\varphi}_{d,2}\widehat{\varphi}_{d,2}\sqrt{\log(1/\delta)}\log(d)}{\sqrt{n}}\frac{\sum_{t=1}^{n}\|\widehat{\boldsymbol{\varphi}}_d(x_t)\|_2}{n}\right) \\
&\le \mathcal{O}\left(\frac{\overline{\varphi}_{d,2}\widehat{\varphi}_{d,2}\sqrt{\log(1/\delta)}\log(d)}{\sqrt{n}}\frac{\sqrt{n\sum_{t=1}^{n}\|\widehat{\boldsymbol{\varphi}}_d(x_t)\|_2^2}}{n}\right) \\
&= \mathcal{O}\left(\frac{\overline{\varphi}_{d,2}\widehat{\varphi}_{d,2}\sqrt{\log(1/\delta)}\log(d)}{\sqrt{n}}\frac{\sqrt{n^2 d}}{n}\right) \\
&= \mathcal{O}\left(\frac{\sqrt{d}\,\overline{\varphi}_{d,2}\widehat{\varphi}_{d,2}\sqrt{\log(1/\delta)}\log(d)}{\sqrt{n}}\right).
\end{aligned}
$$

Here, the first equality is due to the fact that, being $\widehat{\boldsymbol{\varphi}}_d$ orthonormal w.r.t. μ_n, we have $\sum_{t=1}^{n}\|\widehat{\boldsymbol{\varphi}}_d(x_t)\|_2^2 = nd$. This holds uniformly for every $\boldsymbol{\alpha}$, as we have only used the fact that $\|I_{\boldsymbol{\alpha}}\|_2 \le 1$.

Bounding the second term.

The second term corresponds to

$$
\text{Second term} = \sup_{x\in\mathcal{X}}\left|\frac{1}{n}\sum_{t=1}^{n}\left|\sum_{i=1}^{d}\alpha_i\overline{\varphi}_i(x)\overline{\varphi}_i(x_t)\right| - \int_{\mathcal{X}}\left|\sum_{i=1}^{d}\alpha_i\overline{\varphi}_i(x)\overline{\varphi}_i(z)\right|d\mu(z)\right|.
$$

First, we fix $x \in \mathcal{X}$ and $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ and use the scalar product to write it as

$$
\left|\frac{1}{n}\sum_{t=1}^{n}\left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}\overline{\boldsymbol{\varphi}}_d(x_t)\right| - \int_{\mathcal{X}}\left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}\overline{\boldsymbol{\varphi}}_d(z)\right|d\mu(z)\right|. \tag{19}
$$

Note that by definition

$$\mathbb{E}[|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)|] = \int_{\mathcal{X}} \left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(z)\right| d\mu(z).$$

Moreover,

$$\begin{aligned}
\text{Var}(|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)|) &\leq \mathbb{E}\left[|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)|^2\right] \\
&= \mathbb{E}\left[\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t) \overline{\boldsymbol{\varphi}}_d(x_t)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x)\right] \\
&= \overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \underbrace{\mathbb{E}\left[\overline{\boldsymbol{\varphi}}_d(x_t) \overline{\boldsymbol{\varphi}}_d(x_t)^\top\right]}_{=I_d} I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x) \\
&= \overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}^2 \overline{\boldsymbol{\varphi}}_d(x) \\
&\leq \overline{\varphi}_2^2,
\end{aligned}$$

where the last step comes from the fact that $I_{\boldsymbol{\alpha}}^2 \preceq I_d$. For the same reason, we also have $|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)| \leq \overline{\varphi}_2^2$ almost surely. These three results allow us to apply Bernstein's inequality 16 for

- $X_t = |\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)| - \mathbb{E}[|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)|]$.
- $\sigma^2 = \sum_{t=1}^n \text{Var}(|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)|) \leq n\overline{\varphi}_2^2$.
- $B = \overline{\varphi}_2^2$.

This gives, with probability at least $1 - \delta$,

$$\left|\sum_{t=1}^n X_t\right| \leq \sqrt{2n\overline{\varphi}_2^2 \log(2/\delta)} + \frac{2\overline{\varphi}_2^2}{3} \log(2/\delta).$$

So, we can bound equation 19, which corresponds to $\frac{1}{n}\left|\sum_{t=1}^n X_t\right|$, as follows.

$$\left|\frac{1}{n}\sum_{t=1}^n \left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)\right| - \int_{\mathcal{X}} \left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(z)\right| d\mu(z)\right| \leq \sqrt{\frac{2\overline{\varphi}_2^2 \log(2/\delta)}{n}} + \frac{2\overline{\varphi}_2^2}{3n} \log(2/\delta).$$

The former holds for any fixed $\boldsymbol{\alpha} \in \mathcal{A}_d^D$. To have a uniform bound, let

$$\mathcal{A}' = \varepsilon - \text{Cover of } \mathcal{A}_d^D \qquad \varepsilon = (n\overline{\varphi}_{d,2})^{-1},$$

so that $\log|\mathcal{A}'| \leq d\log(n\overline{\varphi}_{d,2})$. Making a union bound gives, $\forall \boldsymbol{\alpha} \in \mathcal{A}'$

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{t=1}^n \left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)\right| - \int_{\mathcal{X}} \left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(z)\right| d\mu(z)\right| \\
&\qquad \leq \sqrt{\frac{2d\overline{\varphi}_{d,2}^2 \log(2n\overline{\varphi}_{d,2}/\delta)}{n}} + \frac{2d\overline{\varphi}_2^2}{3n} \log(2n\overline{\varphi}_{d,2}/\delta).
\end{aligned}$$

To pass to the general case, note that for every $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ there is $\boldsymbol{\alpha}' \in \mathcal{A}'$ such that $\left|\frac{1}{n}\sum_{t=1}^n \left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(x_t)\right| - \int_{\mathcal{X}} \left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}} \overline{\boldsymbol{\varphi}}_d(z)\right| d\mu(z)\right|$ changes no more than $2\overline{\varphi}_{d,2}$ between the two, by definition of $\varepsilon-$cover. Therefore, we have, with probability at least $1-\delta$ over all $\boldsymbol{\alpha} \in \mathcal{A}_d^D$ at the same time

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{t=1}^{n}\left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}\overline{\boldsymbol{\varphi}}_d(x_t)\right| - \int_{\mathcal{X}}\left|\overline{\boldsymbol{\varphi}}_d(x)^\top I_{\boldsymbol{\alpha}}\overline{\boldsymbol{\varphi}}_d(z)\right| d\mu(z)\right| \\
&\qquad \leq \sqrt{\frac{2d\overline{\varphi}_{d,2}^2\log(2n\overline{\varphi}_{d,2}/\delta)}{n}} + \frac{2d\overline{\varphi}_2^2}{3n}\log(2n\overline{\varphi}_{d,2}/\delta) + 2\overline{\varphi}_{d,2}.
\end{aligned}
$$

This means,

$$
\text{Second term} \leq \widetilde{\mathcal{O}}\left(\sqrt{\frac{d\overline{\varphi}_{d,2}^2\log(1/\delta)}{n}} + \frac{d\overline{\varphi}_2^2}{n}\log(1/\delta)\right).
$$

Putting the two results toghether. By the two bounds that we got for the two terms, it follows with probability at least $1-\delta$

$$
\sup_{\boldsymbol{\alpha}\in\mathcal{A}_d^D}|\Lambda_{\boldsymbol{\alpha},\mu_n} - \Lambda_{\boldsymbol{\alpha},\mu}| \leq \widetilde{\mathcal{O}}\left(\frac{\sqrt{d}\overline{\varphi}_{d,2}\widehat{\varphi}_{d,2}\sqrt{\log(1/\delta)}}{\sqrt{n}} + \sqrt{\frac{d\overline{\varphi}_{d,2}^2\log(1/\delta)}{n}} + \frac{d\overline{\varphi}_2^2}{n}\log(1/\delta)\right).
$$

To end the proof, note that, using equation 18, the difference between $\overline{\varphi}_{d,2}$ and $\widehat{\varphi}_{d,2}$ is of order $\overline{\varphi}_{d,2}^2\sqrt{\log(1/\delta)/n}$, so that

$$
\begin{aligned}
\frac{\sqrt{d}\overline{\varphi}_{d,2}\widehat{\varphi}_{d,2}\sqrt{\log(1/\delta)}}{\sqrt{n}} &\leq \frac{\sqrt{d}\overline{\varphi}_{d,2}(\overline{\varphi}_{d,2}+\overline{\varphi}_{d,2}^2\sqrt{\log(1/\delta)/n})\sqrt{\log(1/\delta)}}{\sqrt{n}} \\
&= \frac{\sqrt{d}\overline{\varphi}_{d,2}^2\sqrt{\log(1/\delta)}}{\sqrt{n}} + \frac{\sqrt{d}\overline{\varphi}_{d,2}^3\log(1/\delta)}{n}.
\end{aligned}
$$

Finally, note that, as $\sqrt{d}\leq\overline{\varphi}_{d,2}$, the term $\frac{\sqrt{d}\overline{\varphi}_{d,2}^3\log(1/\delta)}{n}$ dominates over $\frac{d\overline{\varphi}_2^2}{n}\log(1/\delta)$ that we had before. □

D.3 Proofs about gradient method

Proposition 24. *The function $J:\mathcal{A}_d^D\to(0,+\infty)$ given by $J(\boldsymbol{\alpha}) := \Lambda_{\boldsymbol{\alpha},\mu_n}$ is convex in $\boldsymbol{\alpha}$.*

Proof. By definition,

$$
J(\boldsymbol{\alpha}) = \|M(\boldsymbol{\alpha})\|_\infty,
$$

where $M(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^d \alpha_i\widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t)$. Therefore, in particular

$$
J(\boldsymbol{\alpha}) = \sup_{x\in\mathcal{X},\mathbf{f}\in\{-1,1\}^n}\left|\frac{1}{n}\sum_{i=1}^d \alpha_i\widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t)\mathbf{f}\right|.
$$

This function is convex, being the supremum of a family of linear functions $\frac{1}{n}\sum_{i=1}^d \alpha_i\widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t)$ in $\boldsymbol{\alpha}$. □

Theorem 12. *Fix $\epsilon>0$. Algorithm 1, after a number of iterations $I=\widetilde{\mathcal{O}}(\epsilon^{-2}\widehat{\varphi}_{2,D}^2(D-d))$ outputs $\boldsymbol{\alpha}^{(I)}\in\mathcal{A}_d^D$ such that $J(\boldsymbol{\alpha}^{(I)})\leq\inf_{\boldsymbol{\alpha}\in\mathcal{A}_d^D}J(\boldsymbol{\alpha})+\epsilon$.*

Proof. The first step of this proof consists in finding an upper bound for any sub-gradient of $\boldsymbol{\alpha}$. As we said,

$$J(\boldsymbol{\alpha}) = \sup_{x\in\mathcal{X},\mathbf{f}\in\{-1,1\}^n} \left| \frac{1}{n}\sum_{i=1}^{d} \alpha_i \widehat{\varphi}_i(x)\widehat{\varphi}_i(x_t)\mathbf{f} \right| = \sup_{x\in\mathcal{X},\mathbf{f}\in\{-1,1\}^n} \left| \frac{1}{n}\widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top \mathbf{f} \right|,$$

where $I_{\boldsymbol{\alpha}} = \mathrm{diag}(\boldsymbol{\alpha})$ is a $D \times D$ diagonal matrix and $\widehat{\Phi}$ is the $n \times d$ matrix having, as rows, $\widehat{\boldsymbol{\varphi}}_D(x_t)$ for each $t = 1, \ldots n$. At this point note that, by duality

$$J(\boldsymbol{\alpha}) = \sup_{x\in\mathcal{X},\mathbf{f}\in\{-1,1\}^n} \left| \frac{1}{n}\widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top \mathbf{f} \right| = \sup_{x\in\mathcal{X}} \frac{1}{n}\sum_{t=1}^{n} |\{\widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top\}_t|,$$

where $\{\}_t$ denotes the $t-$th component of $\widehat{\boldsymbol{\varphi}}_D(x)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top$, which is a $1 \times n$ row vector. Now, assuming[3] that the supremum is obtained by just one value $x_* \in \mathcal{X}$, we can compute the gradient as

$$\begin{aligned}
\nabla J(\boldsymbol{\alpha}) &= \nabla \frac{1}{n}\sum_{t=1}^{n} |\{\widehat{\boldsymbol{\varphi}}_D(x_*)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top\}_t| \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathrm{sign}(\{\widehat{\boldsymbol{\varphi}}_D(x_*)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top\}_t) \nabla \{\widehat{\boldsymbol{\varphi}}_D(x_*)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top\}_t \\
&= \frac{1}{n}\sum_{t=1}^{n} \mathrm{sign}(\{\widehat{\boldsymbol{\varphi}}_D(x_*)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top\}_t) \widehat{\boldsymbol{\varphi}}_D(x_*)^\top \odot \{\widehat{\Phi}\}_t^\top.
\end{aligned}$$

In the last line, we have used the Hadamard product $\odot$, that is defined, for two vectors of length D like $\widehat{\boldsymbol{\varphi}}_D(x_*)^\top$ and $\{\widehat{\Phi}\}_t^\top$, as the component-wise product, generating another vector of length D.

Now, we are going to bound the two-norm of this gradient:

$$\begin{aligned}
\|\nabla J(\boldsymbol{\alpha})\|_2^2 &= \sum_{i=1}^{D} \left\{ \frac{1}{n}\sum_{t=1}^{n} \mathrm{sign}(\{\widehat{\boldsymbol{\varphi}}_D(x_*)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top\}_t) \widehat{\boldsymbol{\varphi}}_D(x_*)^\top \odot \{\widehat{\Phi}\}_t^\top \right\}_i^2 \\
&\leq \sum_{i=1}^{D} \frac{1}{n}\sum_{t=1}^{n} \left\{ \mathrm{sign}(\{\widehat{\boldsymbol{\varphi}}_D(x_*)^\top I_{\boldsymbol{\alpha}} \widehat{\Phi}^\top\}_t) \widehat{\boldsymbol{\varphi}}_D(x_*)^\top \odot \{\widehat{\Phi}\}_t^\top \right\}_i^2 \\
&\leq \sum_{i=1}^{D} \frac{1}{n}\sum_{t=1}^{n} \left\{ \widehat{\boldsymbol{\varphi}}_D(x_*)^\top \odot \{\widehat{\Phi}\}_t^\top \right\}_i^2 \\
&\leq \sum_{i=1}^{D} \frac{1}{n}\sum_{t=1}^{n} \widehat{\varphi}_i(x_*)^2 \widehat{\varphi}_i(x_t)^2 \\
&\leq \sum_{i=1}^{D} \widehat{\varphi}_i(x_*)^2 \underbrace{\frac{1}{n}\sum_{t=1}^{n} \widehat{\varphi}_i(x_t)^2}_{=1} = \widehat{\varphi}_{D,2}^2,
\end{aligned}$$

where the last passage holds since the features $\widehat{\varphi}_i(\cdot)$ are orthonormal w.r.t. $\mu_n(\cdot)$. Under these assumption, namely

1. J is convex
2. Each sub-gradient has norm bounded by $G := \widehat{\varphi}_{2,D}$
3. The diameter of the optimization space $\mathcal{H}_d^D$ is $R := \sqrt{D-d}$

[3] if there are ties, the argument applied to each of them still holds bounding the norm of the sub-gradient

equation (3) on Boyd et al. [2003] guarantees that running the subgradient method for I iterations with step size

$$\gamma_\ell = \frac{R}{G\sqrt{\ell+1}}$$

(corresponding to line 7), achieves suboptimality ϵ_I bounded by

$$\epsilon_I \leq \frac{R^2 + G^2 \sum_{\ell=1}^{I} \gamma_\ell^2}{2\sum_{\ell=1}^{I} \gamma_\ell} \leq \frac{R^2 + R^2(\log(I)+1)}{(R/G)\sqrt{I}} \leq \frac{2RG\log(I)}{\sqrt{I}} = \frac{2\widehat{\varphi}_{2,D}\sqrt{D-d}\log(I)}{\sqrt{I}}.$$

Therefore, a number of iterations $I = 4\epsilon^{-2}\widehat{\varphi}_{2,D}^2(D-d)\log^3(4\widehat{\varphi}_{2,D}^2(D-d))$ allows to ensure $\epsilon_I \leq \epsilon$. In this way, we have

$$\widehat{\Lambda}_{\boldsymbol{\alpha}^*} - \inf_{\boldsymbol{\alpha}\in\mathcal{A}_d^D} \widehat{\Lambda}_{\boldsymbol{\alpha}} = J(\boldsymbol{\alpha}^{(I)}) - \inf_{\boldsymbol{\alpha}\in\mathcal{A}_d^D} J(\boldsymbol{\alpha}) \leq \epsilon_I \leq \epsilon,$$

which completes the proof. □

Theorem 13. *Let Assumptions 1 and 2 hold and fix $\delta > 0$. Then, with probability $1-\delta$,*

$$\mathcal{E}_\infty(\widehat{\theta}_{n,BWR}) \leq (1+\Lambda_\mu^{Oracle})\mathcal{E}_\infty(f) + \widetilde{\mathcal{O}}\left(\frac{\overline{\varphi}_{2,D}\sqrt{D\log(|\mathcal{X}|/\delta)}}{\sqrt{n}} + \frac{\overline{\varphi}_{2,D}^2\log(|\mathcal{X}|/\delta)}{n}\right).$$

Proof. By theorem 9 and the definition of $\widehat{\theta}_{n,\mathrm{BWR}}$,

$$\mathcal{E}_\infty(\widehat{\theta}_{n,\mathrm{BWR}}) \leq (1+\Lambda_{\boldsymbol{\alpha}^{(I)},\mu_n})\mathcal{E}_\infty(f) + \frac{\sigma\widehat{\varphi}_{2,D}\sqrt{2\log(2\mathcal{X}/\delta)}}{\sqrt{n}}. \tag{20}$$

By theorem 12, for fixed ϵ, we have $\Lambda_{\boldsymbol{\alpha}^{(I)},\mu_n} \leq \min_{\boldsymbol{\alpha}\in\mathcal{A}_d^D}\Lambda_{\boldsymbol{\alpha},\mu_n} + \epsilon$. Moreover, note that

$$\begin{aligned}
\Lambda_\mu^{\mathrm{Oracle}} &= \Lambda_{\boldsymbol{\alpha}_\mu^{\mathrm{Oracle}},\mu} \\
&\geq \Lambda_{\boldsymbol{\alpha}_\mu^{\mathrm{Oracle}},\mu_n} - \widetilde{\mathcal{O}}\left(\frac{\overline{\varphi}_{2,D}\sqrt{D\log(|\mathcal{X}|/\delta)}}{\sqrt{n}} + \frac{\overline{\varphi}_{2,D}^2\log(|\mathcal{X}|/\delta)}{n}\right) \\
&\geq \min_{\boldsymbol{\alpha}\in\mathcal{A}_d^D}\Lambda_{\boldsymbol{\alpha},\mu_n} - \widetilde{\mathcal{O}}\left(\frac{\overline{\varphi}_{2,D}\sqrt{D\log(|\mathcal{X}|/\delta)}}{\sqrt{n}} + \frac{\overline{\varphi}_{2,D}^2\log(|\mathcal{X}|/\delta)}{n}\right) \\
&\geq \Lambda_{\boldsymbol{\alpha}^{(I)},\mu_n} - \epsilon - \widetilde{\mathcal{O}}\left(\frac{\overline{\varphi}_{2,D}\sqrt{D\log(|\mathcal{X}|/\delta)}}{\sqrt{n}} + \frac{\overline{\varphi}_{2,D}^2\log(|\mathcal{X}|/\delta)}{n}\right).
\end{aligned}$$

Replacing this relation in equation 20 we get the thesis.

□

D.4 Gradient method

E Proofs of section 5

Theorem 14. *Let $\mu(\cdot) = \mathcal{U}([-1,1])$. There is a constant C independent of d such that, for $D = 2d$ and $\boldsymbol{\varphi}_d(x) = [1,\ldots x^{d-1}]$, $\boldsymbol{\varphi}_D(x) = [1,\ldots,x^{2d-1}]$, we have $\Lambda_\mu^{Oracle} \leq C$.*

Proof. See Theorem 3.1 by Themistoclakis and Van Barel [2017] □

Algorithm 1 Subgradient Method

Require: Feature map $\boldsymbol{\varphi}_D$, d, Number I of iterations
Ensure: Sequence $\boldsymbol{\alpha}^* \in \mathcal{A}_d^D$

```
Compute φ̂_D from φ_D via Gram-Schmidt orthogonalization
Define loss as in equation (??):
        J(α) = ||M(α)||_∞
Initialize α^(0) ← [ones(d), zeros(D − d)]^⊤
for ℓ = 1 to I do
    Compute step size γ_ℓ = √(D−d) / (φ̂_{2,d} √(ℓ+1))
    Compute a subgradient g_ℓ ∈ ∂J(α^(ℓ−1))
    Update: α^(ℓ) = α^(ℓ−1) − γ_ℓ g_ℓ
    if α^(ℓ) ∉ A_d^D then
        Project: h^(ℓ) = Π_{H_d^D} α^(ℓ)
    end if
end for
return α* = α^(I)
```

Proposition 25. *Fix $\gamma > 0$. There is a function $f : [-1,1] \to \mathbb{R}$ such that, $\mathcal{E}_\infty(f) \overset{d}{\to} 0$ and under assumptions 1 and 2 for $\mu = \mathcal{U}([-1,1])$, with probability one,*

$$\lim_{d\to\infty}\lim_{n\to\infty} \|f(\cdot) - \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,BWR}\|_\infty = 0 \qquad \lim_{n\to\infty} \|f(\cdot) - \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,OLS}\|_\infty \gtrsim d^{1-\gamma}.$$

Most of the proof of this proposition is about in building the function, that we are calling $f(\cdot)$.

The construction of the function in this proof is going to be quite involved. The function is going to be a sum over n of terms of the form $\widetilde{f}_n(\cdot)$. The following notation will be used

1. Let d_n dimension of the basis function used at step n
2. Let $a_n = d_n^{-\gamma}$, for a parameter $\gamma > 0$ to be defined
3. Let h_n width of the mollifier
4. Let $M_n(\cdot) = M(\cdot/h_n)$, where $M(\cdot)$ is the standard mollifier, that is, a nonnegative function $M(\cdot) \in \mathcal{C}^\infty((-1,1))$ with integral one and compact support.
5. $f_n(\cdot) := \operatorname{sgn}(\overline{\boldsymbol{\varphi}}_{d_n}(\cdot)^\top \overline{\boldsymbol{\varphi}}_{d_n}(x_n))$, where x_n is such that
$$\|\overline{\boldsymbol{\varphi}}_{d_n}(\cdot)^\top \overline{\boldsymbol{\varphi}}_{d_n}(x_n)\|_{L^1} \geq \sup_{x\in(-1,1)} \|\overline{\boldsymbol{\varphi}}_{d_n}(\cdot)^\top \overline{\boldsymbol{\varphi}}_{d_n}(x)\|_{L^1} - 1.$$
6. $\widetilde{f}_n := f_n * M_n$

We are able to prove the following lemmas:

Lemma 9. *For every n,*

$$\|f_n - \widetilde{f}_n\|_{L^2} = \|f_n - f_n * M_n\|_{L^2} \leq 4\sqrt{h_n} d_n$$

Proof. In order to perform this proof, we need one result from mathematical analysis. In fact, call *bounded variation* a function $\mathcal{X} = (-1,1) \to \mathbb{R}$ such that the following norm is bounded

$$\|f\|_{BV} := \sup_{\{x_n\}_n \subset \mathcal{X}} \sum_n |f(x_{n+1}) - f(x_n)|.$$

A well-known characterization of this space Ambrosio et al. [2000] ensures that the former norm is equivalent to

$$\|f_n\|_{BV} \propto \|f\|_{L^1} + \|f'\|_{\mathcal{M}} \qquad \|f'\|_{\mathcal{M}} := \sup_{g \in \mathcal{C}^0(\mathcal{X}), \|g\|_\infty = 1} \int_{\mathcal{X}} g(x) f'(x) dx.$$

Now, we can proceed to the proof. First, note that by definition f_n is in the $BV((-1,1))$ class with $\|f_n\|_{BV} = \mathcal{O}(d_n)$. Indeed, $f_n(\cdot)$ takes only values in $\{-1,+1\}$, and can only jump between the two values when $\overline{\boldsymbol{\varphi}}_{d_n}(\cdot)^\top \overline{\boldsymbol{\varphi}}_{d_n}(x_n) = 0$, which happens at most d_n times, as the previous is a polynomial of degree d_n. At this point, by the properties of convolution,

$$\begin{aligned} f_n(y) - f_n * M_n(y) &= f_n * (M_n(y) - \delta(y)) \\ &= f_n' * \left(\int_{-1}^{y} M_n(t) - \delta(t)\, dt \right), \end{aligned}$$

Where we have moved the derivative in the first term. At this point, the properties of convolution allow us to say that for any pair of functions g_1, g_2, $\|g_1 * g_2\|_{L^2} \leq \|g_1\|_{\mathcal{M}} \|g_2\|_{L^2}$. Therefore, we have

$$\begin{aligned} \|f_n(\cdot) - f_n * M_n(\cdot)\|_{L^2} &\leq \|f_n'(\cdot)\|_{\mathcal{M}} \left\| \int_{-1}^{y} M_n(t) - \delta(t)\, dt \right\|_{L^2} \\ &\leq \underbrace{\|f_n(\cdot)\|_{BV}}_{\leq d_n} \left\| \int_{-1}^{y} M_n(t) - \delta(t)\, dt \right\|_{L^2}. \end{aligned}$$

At this point, note that by definition $M_n(t) \geq 0$, its integral is one and its support is contained in $(-h_n, h_n)$. Therefore,

$$\left| \int_{-1}^{y} M_n(t) - \delta(t)\, dt \right| \leq \begin{cases} 0 & y \geq h_n \\ 2 & -h_n < y < h_n \\ 0 & y \leq -h_n \end{cases},$$

so that its L^2 norm is bounded by $4\sqrt{h_n}$. This completes the proof.

□

Lemma 10. *For every $m \leq n$, and $s > 0$*

$$\|\widetilde{f}_m - \Pi^{\infty}_{d_{n+1},\infty} \widetilde{f}_m\|_\infty \leq \mathcal{O}(d_{n+1}^{-s} h_m^{-s}).$$

Proof. First, let us examine the smoothness of $\widetilde{f}_m$. Indeed, we have, for any $s > 0$

$$\begin{aligned} \|\widetilde{f}_m\|_{\mathcal{C}^s} &= \|f_m * M_m\|_{\mathcal{C}^s} \\ &\leq \|f_m\|_\infty \|M_m\|_{\mathcal{C}^s} \\ &= \|M_m\|_{\mathcal{C}^s} = \mathcal{O}(h_m^{-s}). \end{aligned}$$

Therefore, by Jackson's theorem, we have for any s,

$$\|\widetilde{f}_m - \Pi_{d_{n+1},\infty} \widetilde{f}_m\| \leq \mathcal{O}(d_{n+1}^{-s} \|\widetilde{f}_m\|_{\mathcal{C}^s}) = \mathcal{O}(d_{n+1}^{-s} h_m^{-s}).$$

□

Theorem 26. *For any $\gamma < 1/4$ there is f^* such that*

- $\lim_d \|f^* - \Pi_{d,\infty} f^*\|_\infty = 0$

- $\limsup_d \frac{\|f^*-\Pi_{d,\mu}f^*\|_\infty}{d^{1-\gamma}} > 0$.

Proof. Let

$$f^*(\cdot) = \sum_{n=1}^{\infty} a_n \widetilde{f}_n(\cdot).$$

First part

Fix $\varepsilon > 0$. As a_n goes to zero faster than exponentially and $\|\widetilde{f}_n(\cdot)\|_\infty \leq 1$, we can find n_0 such that

$$\left\| f^*(\cdot) - \sum_{n=1}^{n_0} a_n \widetilde{f}_n(\cdot) \right\|_\infty \leq \varepsilon/2.$$

Now, $\sum_{n=1}^{n_0} a_n \widetilde{f}_n(\cdot)$ is a finite sum of $C^\infty([-1,1])$ functions, so it is uniformly continuous, in particular. Therefore, by Stone-Weierstrass theorem, for sufficiently large d,

$$\left\| \sum_{n=1}^{n_0} a_n \widetilde{f}_n(\cdot) - \Pi_{d,\infty} \sum_{n=1}^{n_0} a_n \widetilde{f}_n(\cdot) \right\|_\infty \leq \varepsilon/2.$$

Putting the two results together, we have proved that, for sufficiently large d,

$$\begin{aligned}
\|f^* - \Pi_{d,\infty} f^*\|_\infty &\leq \left\| f^*(\cdot) - \Pi_{d,\infty} \sum_{n=1}^{n_0} a_n \widetilde{f}_n(\cdot) \right\|_\infty \\
&\leq \varepsilon/2 + \left\| \sum_{n=1}^{n_0} a_n \widetilde{f}_n(\cdot) - \Pi_{d,\infty} \sum_{n=1}^{n_0} a_n \widetilde{f}_n(\cdot) \right\|_\infty \\
&\leq \varepsilon.
\end{aligned}$$

Second part Let us fix $n = \ell$ and consider

$$\begin{aligned}
\|\Pi_{d_\ell,\mu} f^*\|_\infty &= \left\| \Pi_{d_\ell,\mu} \sum_{n=1}^{\infty} a_n \widetilde{f}_n(\cdot) \right\|_\infty \\
&= \left\| \Pi_{d_\ell,\mu} \sum_{n=1}^{\ell-1} a_n \widetilde{f}_n(\cdot) + \Pi_{d_\ell,\mu} a_\ell \widetilde{f}_\ell(\cdot) + \Pi_{d_\ell,\mu} \sum_{n=\ell+1}^{\infty} a_n \widetilde{f}_n(\cdot) \right\|_\infty \\
&\geq \underbrace{\left\| \Pi_{d_\ell,\mu} a_\ell \widetilde{f}_\ell(\cdot) \right\|_\infty}_{A} - \underbrace{\left\| \Pi_{d_\ell,\mu} \sum_{n=1}^{\ell-1} a_n \widetilde{f}_n(\cdot) \right\|_\infty}_{B} - \underbrace{\left\| \Pi_{d_\ell,\mu} \sum_{n=\ell+1}^{\infty} a_n \widetilde{f}_n(\cdot) \right\|_\infty}_{C}.
\end{aligned}$$

We are going to analyze the three terms separately.

(A) We start bounding the first term from below,

$$\begin{aligned}
A &= a_\ell \left\| \Pi_{d_\ell,\mu} \widetilde{f}_\ell(\cdot) \right\|_\infty \\
&\geq a_\ell \left\| \Pi_{d_\ell,\mu} f_\ell(\cdot) \right\|_\infty - a_\ell \left\| \Pi_{d_\ell,\mu} (\widetilde{f}_\ell(\cdot) - f_\ell(\cdot)) \right\|_\infty \\
&\geq \alpha_\ell \Lambda_{d_\ell,\mu} - \alpha_\ell \overline{\varphi}_{2,d_\ell} \|\boldsymbol{c}_{d_\ell}(\widetilde{f}_\ell(\cdot) - f_\ell(\cdot))\|_2 \\
&= \alpha_\ell \Lambda_{d_\ell,\mu} - \alpha_\ell \overline{\varphi}_{2,d_\ell} \|\widetilde{f}_\ell(\cdot) - f_\ell(\cdot)\|_{L^2} \\
&\geq \alpha_\ell \Lambda_{d_\ell,\mu} - 4\alpha_\ell \overline{\varphi}_{2,d_\ell} d_\ell \sqrt{h_\ell}.
\end{aligned}$$

Here, the second inequality comes from Cauchy-Schwartz, the sequent equality from Parseval's theorem and the last comes from lemma 9. Note that, for the polynomial basis, $\overline{\varphi}_{2,d_\ell} \approx \Lambda_{d_\ell,\mu} \approx d_\ell$, so we get

$$A \geq \Omega\left(\alpha_\ell d_\ell (1 - d_\ell \sqrt{h_\ell})\right)$$

(B) This term is

$$\begin{aligned}
B &= \left\| \Pi_{d_\ell,\mu} \sum_{n=1}^{\ell-1} a_n \widetilde{f}_n(\cdot) \right\|_\infty \\
&\leq \sum_{n=1}^{\ell-1} a_n \left\| \Pi_{d_\ell,\mu} \widetilde{f}_n(\cdot) \right\|_\infty \\
&\leq \sum_{n=1}^{\ell-1} a_n \left\| \Pi_{d_\ell,\mu} (\widetilde{f}_n(\cdot) - \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot)) \right\|_\infty + a_n \left\| \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot) \right\|_\infty .
\end{aligned}$$

The last passage holds as $\Pi_{d_\ell,\mu}\Pi_{d_\ell,\infty}\widetilde{f}_n(\cdot) = \Pi_{d_\ell,\infty}\widetilde{f}_n(\cdot)$. Now, we can apply lemma 10, as $n < \ell$, which ensures

$$\begin{aligned}
B &\leq \sum_{n=1}^{\ell-1} a_n \left\| \Pi_{d_\ell,\mu} (\widetilde{f}_n(\cdot) - \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot)) \right\|_\infty + a_n \left\| \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot) \right\|_\infty \\
&\leq \sum_{n=1}^{\ell-1} a_n \left\| \Pi_{d_\ell,\mu} (\widetilde{f}_n(\cdot) - \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot)) \right\|_\infty + a_n \left\| \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot) \right\|_\infty \\
&\leq \sum_{n=1}^{\ell-1} a_n \Lambda_\ell \left\| \widetilde{f}_n(\cdot) - \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot) \right\|_\infty + a_n \left\| \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot) \right\|_\infty \\
&\leq \sum_{n=1}^{\ell-1} a_n d_\ell^{-s+1} h_n^{-s} + a_n \left\| \Pi_{d_\ell,\infty} \widetilde{f}_n(\cdot) \right\|_\infty .
\end{aligned}$$

(C) The last term can be simply bounded due to the fact that $\|\widetilde{f}_n\|_\infty \leq 1$:

$$C \leq d_\ell \sum_{n=\ell+1}^{\infty} a_n.$$

Now, fix any $\gamma < 1/4$ and take

$$s = 2; \quad d_n = \exp(1/\gamma^n); \quad h_n = \exp(-1/(2\gamma^{n+1})); \quad a_n = \exp(-1/\gamma^{n-1}).$$

We get

$$\begin{aligned}
A &\geq \Omega\left(\alpha_\ell d_\ell (1 - d_\ell \sqrt{h_\ell})\right) \\
&\geq \Omega\left(\exp((1-\gamma)/\gamma^\ell)(1 - \exp(1/\gamma^n - 1/(4\gamma^{n+1})))\right) \\
&\geq \Omega\left(\exp((1-\gamma)/\gamma^\ell)(1 - \exp(1/\gamma^n \underbrace{(1 - 1/(4\gamma))}_{\leq 0}))\right) \\
&\geq \Omega\left(\exp((1-\gamma)/\gamma^\ell)\right) = \Omega(d_\ell^{1-\gamma}).
\end{aligned}$$

For term B, we have

$$\begin{aligned}
B &\leq \mathcal{O}\left(\sum_{n=1}^{\ell-1} a_n d_\ell^{-s+1} h_n^{-s} + a_n \left\|\Pi_{d_\ell,\infty}\widetilde{f}_n(\cdot)\right\|_\infty\right) \\
&\leq \mathcal{O}\left(\sum_{n=1}^{\ell-1} a_n d_\ell^{-s+1} h_n^{-s} + a_n\right) \\
&\leq \mathcal{O}\left(\sum_{n=1}^{\ell-1} a_n \exp((-s+1)/\gamma^\ell)\exp(s/(2\gamma^{n+1})) + a_n\right) \\
&\leq \mathcal{O}\left(\sum_{n=1}^{\ell-1} a_n \exp((-s+1)/\gamma^\ell)\exp(s/(2\gamma^{\ell})) + a_n\right) \\
&\leq \mathcal{O}\left(\sum_{n=1}^{\ell-1} a_n \underbrace{\exp((-s/2+1)/\gamma^\ell)}_{\leq 1} + a_n\right).
\end{aligned}$$

Last term:

$$C \leq \mathcal{O}\left(d_\ell \sum_{n=\ell+1}^{\infty} a_n\right) \leq \mathcal{O}\left(\exp(1/\gamma^\ell)\sum_{n=\ell+1}^{\infty}\exp(-1/\gamma^{n-1})\right) = \mathcal{O}\left(\sum_{m=0}^{\infty}\exp(-1/\gamma^m)\right).$$

Again, this term satisfies $C = \mathcal{O}(1)$, as the term $\exp(-1/\gamma^m)$ in the last sum decays faster than 2^{-m}. □

where the last passage holds as $s = 2$. Therefore we get $B \leq \mathcal{O}(\sum_{n=1}^{\ell-1} a_n) = \mathcal{O}(1)$, since a_n decays faster than 2^{-n} which already generates a convergent seqeuence.

All together, these passages prove

$$\|\Pi_{d_n,\mu} f^*\|_\infty \geq \Omega(d_n^{1-\gamma}).$$

Therefore, taking this d_n sequence entails $\limsup_{d\to\infty} \frac{\|f^*-\Pi_{d,\mu}f^*\|_\infty}{d^{1-\gamma}} > 0$.

Proof. (of proposition 15). Let $f = f^*$ defined before, for the specific value of $\gamma > 0$. Thanks to part one of theorem 26[4], assumption $\mathcal{E}_\infty(f) \xrightarrow{d} 0$ is satisfied:

$$\mathcal{E}_\infty(f) = \|f^* - \Pi_{d,\infty} f^*\|_\infty \xrightarrow{d} 0.$$

Then, we prove the two theses point by point. Point one: for fixed d, theorem 13 gives

$$\|f(\cdot) - \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,\mathrm{BWR}}\|_\infty \leq (1+\Lambda_\mu^{\mathrm{Oracle}})\mathcal{E}_\infty(f) + \widetilde{\mathcal{O}}\left(\frac{\overline{\varphi}_{2,D}\sqrt{D\log(|\mathcal{X}|/\delta)}}{\sqrt{n}} + \frac{\overline{\varphi}_{2,D}^2\log(|\mathcal{X}|/\delta)}{n}\right).$$

As $\mathcal{X}$ is $[-1,1]$ and the feature map is Lipschitz continuous, we can get rid of the $|\mathcal{X}|$ by a covering argument. As $n \to \infty$, the former gives

$$\lim_n \|f(\cdot) - \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,\mathrm{BWR}}\|_\infty \leq (1+\Lambda_\mu^{\mathrm{Oracle}})\mathcal{E}_\infty(f).$$

[4]formally, the result holds for $\gamma > 1/4$ but, for what we are trying to prove, the validity of the statement for γ implies its validity for every $\gamma' > \gamma$, therefore we can proceed w.l.o.g.

For $\mu = \mathcal{U}([-1,1])$, theorem 14 ensured that $\Lambda_\mu^{\text{Oracle}} < C$, a universal constant independent on d. Therefore,

$$\lim_d \lim_n \|f(\cdot) - \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,\text{BWR}}\|_\infty \leq \lim_n (1+C)\mathcal{E}_\infty(f) = 0.$$

Let us pass to the second thesis:

$$\lim_{n\to\infty} \|f(\cdot) - \boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,\text{OLS}}\|_\infty \gtrsim d^{1-\gamma}.$$

This follows from the fact that, for $n \to \infty$, $\boldsymbol{\varphi}_d(\cdot)^\top \widehat{\theta}_{n,\text{OLS}} \to \Pi_{d,\mu} f(\cdot)$ and that theorem 26 ensures $\limsup_d \frac{\|f^* - \Pi_{d,\mu} f^*\|_\infty}{d^{1-\gamma}} > 0$.

□