Regularization in Kernel Learning

Shahar Mendelson[1,2] Joseph Neeman[1,3]

February 29, 2008

Abstract

Under mild assumptions on the kernel, we obtain the best known error rates in a regularized learning scenario taking place in the corresponding reproducing kernel Hilbert space. The main novelty in the analysis is a proof that one can use a regularization term that grows significantly slower than the standard quadratic growth in the RKHS norm.

1 Introduction

Let F be a family of functions from a probability space (Ω, μ) to $\mathbb{R}$. A classical problem of Learning Theory is the following: we set ν to be an (unknown) probability measure on $\Omega \times \mathbb{R}$ whose marginal distribution on Ω is μ. Given n independent samples $(X_1, Y_1) \ldots (X_n, Y_n) \in \Omega \times \mathbb{R}$ distributed according to ν, our task is to find a function $\hat{f} \in F$ such that

$$\mathbb{E}(\hat{f}(X_1) - Y_1)^2 - \inf_{f \in F} \mathbb{E}(f(X_1) - Y_1)^2 \tag{1.1}$$

is very small. In other words, we want to approximate the distribution ν by a function from F as closely as possible. Specifically, we want to find a method of choosing $\hat{f}$ as a function of the sample $(X_i, Y_i)_{i=1}^n$ such that, with high probability, (1.1) is smaller than a function of n that tends to zero as n grows.

[1] Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia and Department of Mathematics, Technion, I.I.T, Haifa 32000, Israel.

SM was supported in part by an Australian Research Council Discovery grant DP0559465 and by an Israel Science Foundation grant 666/06.

[2] Email: shahar.mendelson@anu.edu.au

[3] Email: joeneeman@gmail.com

A widely-used approach for solving this problem is to consider a function $\hat{f} \in F$ that minimizes the functional

$$\sum_{i=1}^{n}(f(X_i) - Y_i)^2.$$

over all $f \in F$. Such a function is called an empirical minimizer and its properties have been widely studied (see, for example [3]). It turns out that the complexity and geometry of F play a large part in determining whether (1.1) is small. Roughly speaking, if F is a small family of functions, then (1.1) will be, with high probability, a rapidly decreasing function of n.

Of course, there is a disadvantage to having a small family of functions, namely that $\inf_{f \in F} \mathbb{E}(f(X_1) - Y_1)^2$ becomes larger as F becomes smaller. This trade-off is known as the *bias-variance* problem. The expression (1.1) is known as the *sample error* and $\inf_{f \in F} \mathbb{E}(f(X_1) - Y_1)^2$ is called the *approximation error.*

One major issue that needs to be addressed when using the empirical minimization algorithm is overfitting. Since all the information that one has is on the behavior of the minimizer on the sample, there is no way of distinguishing a "simple" minimizer from a more complicated one. The *regularized learning model* is a method of solving the bias-variance problem while addressing the overfitting problem. We take F to be a very large function class (so that the approximation error is small) and consider a function $\hat{f}$ that minimizes the functional

$$\sum_{i=1}^{n}(f(X_i) - Y_i)^2 + \gamma_n(f)$$

where $\gamma_n(f)$ measures, in some sense, the "complexity" of the function of f and, for a fixed f, $\gamma_n \to 0$ as $n \to \infty$. Thus, if two functions have the same empirical behavior, the algorithm will choose the simpler function of the two.

A common example of the regularized learning problem, and the situation we will be considering in this article is the case where the class of functions is a Reproducing Kernel Hilbert Space (RKHS), defined below - and which will be denoted throughout this article by H. All the error bounds in this situation were restricted to a regularization term of the form $\gamma_n(f) = \eta_n \|f\|_H^2$, and typically the aim was to make η_n as small as possible. It was not believed that one can improve the power of $\|f\|_H$ in the regularization process. Doing just that is the main goal of this article.

One can motivate the regularized learning model by looking at it as a collection of empirical minimization problems. Indeed, let B_H be the unit ball of the space H and consider the empirical minimization problem in rB_H for some $r > 0$. As r increases, the approximation error for rB_H decreases and its sample error increases. We could achieve a small total error by choosing the right value of r and performing empirical minimization in rB_H. The role of the regularization term $\gamma_n(f)$ is to force the algorithm to choose the right value of r for empirical minimization. We will explain later why this motivation can be made rigorous, and that the regularization problem may be solved by a solution to a hierarchy of minimization problems.

It should be clear from this motivation that the choice of γ_n is critical for the success of the regularized learning model. There has been some significant work done recently on finding explicit formulas for γ_n that provide low error rates with high probability. Of particular importance are the results of Caponetto and De Vito [6]; and Smale and Zhou [26], which use operator-theoretic techniques to bound the error rate. The point of comparison for our result will be that of Smale and Zhou: let $T_K : L_2(\Omega, \mu) \to L_2(\Omega, \mu)$ be the integral operator associated with the kernel $K : \Omega \times \Omega \to \mathbb{R}$, defined by

$$(T_K f)(x) = \int K(x, y) f(y)\, dy.$$

where Ω is a compact subset of $\mathbb{R}^d$. It is well known that this is a compact, positive, trace-class operator and that the RKHS is $H = T_K^{1/2} L_2$ with the inner product

$$\langle f, g\rangle_H = \langle T_K^{-1/2} f, T_K^{-1/2} g\rangle_2.$$

Theorem 1.1 *Fix $\theta > 0$ and assume that the regression function $\mathbb{E}(Y|X)$ belongs to the range of T_K^{θ}. For any $x > 0$, define the regularization parameter by*

$$\gamma_n(f) = \begin{cases} cx \left(\frac{\|T_K^{-\theta}\mathbb{E}(Y|X)\|^2}{n} \right)^{1/(1+2\theta)} \|f\|_H^2, & \theta \geq \frac{1}{2} \\ \frac{cx}{\sqrt{n}} \|f\|_H^2, & \theta < \frac{1}{2}. \end{cases}$$

Then, with probability at least $1 - \exp(-x)$, the regularized minimizer satisfies

$$\mathbb{E}(\hat{f}(X) - Y)^2 - \inf_{f \in F} \mathbb{E}(f(X) - Y^2) \leq \begin{cases} cx \left(\frac{\|T_K^{-\theta}\mathbb{E}(Y|X)\|^{1/\theta}}{n} \right)^{\theta/(1+2\theta)}, & \theta \geq \frac{1}{2} \\ cx \left(\frac{1 + \|T_K^{-\theta}\mathbb{E}(Y|X)\|}{n} \right)^{\theta/2}, & \theta < \frac{1}{2}. \end{cases}$$

The approach we take is significantly different to that which has been used before in this context. As we mentioned, our goal is to not only improve the way the regularization term depends on the sample size n, but to explain why a quadratic dependence on $\|f\|_H$ is pessimistic and can be improved considerably. The starting point of our analysis is the notion of isomorphic coordinate projections, introduced in the context of Learning Theory in [3].

Suppose F is a family of functions for which the infimum $\inf_{f \in F} \mathbb{E}(f(X) - Y)^2$ is achieved; call the minimizer f^* and define the excess loss function to be, for any $f \in F$,

$$\mathcal{L}_f(X,Y) = (f(X) - Y)^2 - (f^*(X) - Y)^2.$$

Denote by P the conditional expectation with respect to the sample:

$$P\mathcal{L}_f = \mathbb{E}(\mathcal{L}_f | X_1, Y_1, \ldots, X_n, Y_n)$$

and let $P_n \mathcal{L}_f = \sum_{i=1}^n \mathcal{L}_f(X_i, Y_i)$. One can show that there is some (small) number ρ_n such that, with probability at least $1 - e^{-x}$, every $f \in F$ satisfies

$$\frac{1}{2} P_n \mathcal{L}_f - \rho_n \leq P\mathcal{L}_f \leq 2 P_n \mathcal{L}_f + \rho_n. \tag{1.2}$$

This is a useful approach for bounding the error of the empirical minimizer. Indeed, it is not hard to see that it implies

$$\mathbb{E}(\hat{f}(X) - Y)^2 - \inf_{f \in F} \mathbb{E}(f(X) - Y)^2 = P\mathcal{L}_{\hat{f}} \leq \rho_n.$$

It turns out that this “isomorphic coordinate projection” approach applies to regularized learning as well as to empirical minimization. The main result in this direction is due to Bartlett [1] and implies that if every ball rB_H satisfies an almost-isomorphic condition then it is possible to establish a regularized learning bound.

Theorem 1.2 *[1] For each $f \in H$, denote by $\mathcal{L}_f$ the loss of f relative to the ball $\|f\|B_H$:*

$$\mathcal{L}_f(X,Y) = (f(X) - Y)^2 - (f^*(X) - Y)^2$$

where $f^* = \operatorname{argmin}_{\|g\| \leq \|f\|} \mathbb{E}(g(X) - Y)^2$. Under some conditions on $\gamma_n(\cdot)$, if for every $f \in F$,

$$\frac{1}{2} P_n \mathcal{L}_f - \gamma_n(f) \leq P\mathcal{L}_f \leq 2 P_n \mathcal{L}_f + \gamma_n(f)$$

then the regularized minimizer satisfies

$$\mathbb{E}(\hat{f}(X)-Y)^2 \leq \inf_{f \in F} \left((f(X)-Y)^2 + c\gamma_n(c'f)\right),$$

where c and c' are absolute constants.

Thus, if one can establish sharp "isomorphic coordinate projections" type estimates for every excess loss class $\{\mathcal{L}_f : f \in rB_H\}$ this would yield regularization bounds.

It is important to emphasize that although at first glance, the problem of obtaining isomorphic bounds for kernel classes has been solved in the past (based, for example, on estimates from [18, 2]), this is far from being the case. The isomorphic bounds for kernel classes have been studied for the base class $F = B_H$ (i.e. $r = 1$), while the essential ingredient required for our analysis (and which determines the regularization parameter) is the way in which these bounds scale with the radius r. In all the previous isomorphic results obtained in the context of kernel classes this was not important and thus never addressed. Moreover, the analysis used to obtain those results would give a suboptimal estimate as a function of r, and as we will explain later, will be of no help to us in an attempt to improve the way $\gamma_n(f)$ depends on $\|f\|_H$.

Our analysis will show that the standard regularization bounds, that grow like r^2 where $r = \|f\|_H$, are very pessimistic and may be improved considerably. Moreover, if we set the regularization term as $\eta_n \nu(\|f\|_H)$, we will establish the best known bounds on η_n as well (both results will require mild assumptions on the kernel).

There are two reasons for the improved bounds. First of all, the ability to employ the "isomorphic" approach allows one to use localization techniques. Thus, the effective complexity of the excess loss class is caused only from excess loss functions with a relatively small variance. Thanks to the geometry of rB_H, that happens to be a rather small subset of the excess loss class. This approach, presented in Section 3, is enough to give improved estimates on η_n, but still leaves one with a regularization term that grows like r^2.

To remove the r^2 regularization term one has to use a more sophisticated analysis (and additional assumptions on the kernel). As a starting point, one has to understand the source of the r^2 term. The "trivial" reason for this term is the quadratic loss function. Indeed, to obtain isomorphic type estimates one has to analyze the localized empirical process

$$\sup_{\{f \in rB_H :\ \mathbb{E}\mathcal{L}_f \le \lambda\}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_f(X_i) - \mathbb{E}\mathcal{L}_f \right| ,$$

and standard methods of analyzing this quantity involve contraction inequalities. Since the only a-priori L_∞ estimate on $\mathcal{L}_f$ is $\|f\|_\infty^2$, one will suffer one factor of r as a consequence of the contraction inequality and another one from the "complexity" of rB_H. In Section 4 we will present a way of bypassing this loose method of analysis. To that end, we shall present a general bound on the empirical process indexed by the localized square excess loss class associated with a base class consisting of linear functionals on ℓ_2 of norm at most r. We will use this result to show that if the eigenvalue vector of the integral operator T_K belongs to a weak ℓ_p space $\ell_{p,\infty}$ for some $0 < p < 1$, then one can obtain an isomorphic bound with ρ_n that scales like

$$\max\left\{ \theta^{2/1+p}, \theta^{2/p} \right\} ,$$

for $\theta \sim r^p n^{-1/2} \log n$. This translates to a regularization term of

$$\max\left\{ r^{2p/1+p} \left(\frac{\log^2 n}{n} \right)^{1/1+p}, \frac{r^2}{n} \right\} ,$$

where again, $r = \|f\|_H$.

Although with this result one still has a regularization term that grows like r^2, this is a considerable improvement to the previous result. Because it decays faster as a function of the sample size n, the r^2/n term seems superfluous – because one would expect it to be dominated by the first term. And indeed, in Section 5 we will show that it can be removed: under the same assumption on the decay of the eigenvalues of T_K as above one may use a regularization term (up to logarithmic term) of

$$\frac{r^{2p/1+p}}{n^{1/1+p}} ,$$

which is the best known dependency on r and n.

We will end this introduction with the formulation of this, our main result.

Assumption. Assume that the eigenvalues of the integral operator T_K satisfy that $(\lambda_n)_{n=1}^{\infty} \in \ell_{p,\infty}$ for some $0 < p < 1$ and that $\|K(x,x)\|_\infty \le 1$.

Assume further that there is a constant A for which the eigenfunctions $(\varphi_n)_{n\geq 1}$ of T_K satisfy that $\sup_n \|\varphi_n\|_\infty \leq A < \infty$.

Theorem A. Under the assumption above, there exist constants c_1, c_2 and c_3 that depend only on A, p and $\|(\lambda_i)\|_{p,\infty}$ and a constant c_Y that depends only on $\|Y\|_\infty$ for which the following holds. Let

$$\tilde{V}(f,u) = c_3(1+u+c_Y \ln n + \ln\log(\|f\|_H + e)) \left(\frac{(\|f\|_H + 1)^p \log n}{\sqrt{n}}\right)^{2/(1+p)}.$$

If $n \geq N_0 = N_0(\|Y\|_\infty, p)$ and $c_1 \log\log n \leq u \leq c_2(\log n)^{2/(1-p)}$, then with probability at least $1-\exp(-u/2)$, every minimizer $\hat{f}$ of

$$P_n \ell_f + \kappa_1 \tilde{V}(f,u)$$

satisfies that

$$P\ell_{\hat{f}} \leq \inf_{f\in H} P\ell_f + \kappa_2 \tilde{V}(f,u),$$

where κ_1 and κ_2 are absolute constants and $\ell_f = (f-Y)^2$ is the squared loss function.

2 Preliminaries

We begin with a word about notation. We will denote absolute constants (that is, fixed, positive numbers) by $c, c_1, ...$ etc. Their value may change from line to line. Absolute constants whose value will remain unchanged are denoted by $\kappa_1, \kappa_2, ...$ By $c(a)$ we mean that the constant c depends only on the parameter a. We denote $a \sim b$ if there exist absolute constants c_1 and c_2 such that $c_1 a \leq b \leq c_2 b$, and $a \sim_p b$ if the equivalence constants depend on the parameter p.

Arguably the most important tool in modern empirical processes theory is Talagrand's concentration inequality for an empirical process indexed by a class of uniformly bounded functions [28, 15]. The version of this concentration results we shall use here is due to Massart [17].

Theorem 2.1 *There exists an absolute constant C for which the following holds. Let F be a class of functions defined on (Ω, μ) such that for every $f \in F$, $\|f\|_\infty \leq b$ and $\mathbb{E}f = 0$. Let $X_1, ..., X_n$ be independent random variables distributed according to μ and set $\sigma^2 = n \sup_{f\in F} \mathbb{E}f^2$. Define*

$$Z = \sup_{f\in F} \sum_{i=1}^n f(X_i) \text{ and } \bar{Z} = \sup_{f\in F} \left|\sum_{i=1}^n f(X_i)\right|.$$

Then, for every $x > 0$ and every $\rho > 0$,

$$Pr\left(\left\{Z \geq (1+\rho)\mathbb{E}Z + \sigma\sqrt{Cx} + C(1+\rho^{-1})bx\right\}\right) \leq e^{-x},$$
$$Pr\left(\left\{Z \leq (1-\rho)\mathbb{E}Z - \sigma\sqrt{Cx} - C(1+\rho^{-1})bx\right\}\right) \leq e^{-x},$$

and the same inequalities hold for $\bar{Z}$.

Throughout this article we denote by $\ell(x,y) = (x-y)^2$ the squared loss function. When f is a function $\Omega \to \mathbb{R}$ and Y is some target random variable, we denote $\ell_f = (f-Y)^2$. If F is a class of functions let $\mathcal{L}_f = (f-Y)^2 - (f^*-Y)^2$, where $f^* = \text{argmin}_{f\in F}\mathbb{E}\ell(f,Y)$. Of course, we assume that this minimizer exists and is unique, which is the case, for example, if F is compact and convex.

For a class of functions F on a probability space (Ω,μ) we set

$$\|P_n - P\|_F = \sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right|,$$

where $(X_i)_{i=1}$ are independent, distributed according to μ.

Define for any $\lambda \geq 0$ the localized excess loss class

$$\mathcal{L}_\lambda = \{\mathcal{L}_f : \mathbb{E}\mathcal{L}_f \leq \lambda\},$$

and set

$$V = \text{star}(\mathcal{L}_F, 0) = \{\theta\mathcal{L}_f : 0 \leq \theta \leq 1,\ f \in F\},$$
$$V_\lambda = \{\theta\mathcal{L}_f : 0 \leq \theta \leq 1,\ \mathbb{E}(\theta\mathcal{L}_f) \leq \lambda\} = \{h \in \text{star}(\mathcal{L}_F, 0) : \mathbb{E}h \leq \lambda\}$$

(where for a set T, $\text{star}(T,0) = \{\theta t :\ 0 \leq \theta \leq 1,\ t \in T\}$ is the star-shaped hull of T and 0).

The following "isomorphic" result is similar in nature to the one proved in [3]. The bound from Theorem 2.2 normally leads to an estimate on the error of the empirical minimizer, but in [4] and here it will serve a different goal. This isomorphic result will enable us to control the solution of the regularized learning problem in the context of kernel learning.

Theorem 2.2 *There exists an absolute constant c for which the following holds. Let $\mathcal{L}_F$ be a squared loss class associated with a convex class F and a random variable Y. If $b = \max\{\sup_{f\in F}\|f\|_\infty, \|Y\|_\infty\}$ and $\mathbb{E}\|P_n - P\|_{V_\lambda} \leq \lambda/8$, then with probability $1-\exp(-u)$, for every $f \in F$*

$$\frac{1}{2}P_n\mathcal{L}_f - \frac{\lambda}{2} - c(1+b^2)\frac{u}{n} \leq P\mathcal{L}_f \leq 2P_n\mathcal{L}_f + \frac{\lambda}{2} + c(1+b^2)\frac{u}{n} \qquad (2.1)$$

Proof. By Talagrand's inequality, there exists an absolute constant C such that, with probability at least $1 - e^{-u}$,

$$\|P_n - P\|_{V_\lambda} \leq 2\mathbb{E}\|P_n - P\|_{V_\lambda} + \left(\frac{Cu}{n}\right)^{1/2} \sup_{g \in V_\lambda} \sqrt{\operatorname{Var} g} + \frac{Cbu}{n}.$$

It is standard to verify (see, for example, [16]), that there exists an absolute constant C such that, for a convex class F, every $\mathcal{L}_f \in \mathcal{L}_F$ satisfies $\mathbb{E}\mathcal{L}_f^2 \leq Cb^2 \mathbb{E}\mathcal{L}_f$. Thus, every $g \in V_\lambda$ satisfies $\operatorname{Var} g \leq Cb^2\lambda$. Set

$$\alpha = \max\left\{\lambda, 25C\frac{(1+b^2)u}{n}\right\}$$

and note that, because V is star-shaped and $\alpha \geq \lambda$, $\mathbb{E}\|P_n - P\|_{V_\alpha} \leq \alpha/8$. Therefore, with probability at least $1 - e^{-u}$,

$$\begin{aligned} \|P_n - P\|_{V_\alpha} &\leq \frac{\alpha}{4} + \left(C\frac{b^2\alpha u}{n}\right)^{1/2} + \frac{Cbu}{n} \\ &\leq \frac{\alpha}{4} + \frac{\alpha}{5} + \frac{\alpha}{25} \\ &\leq \frac{\alpha}{2}. \end{aligned} \tag{2.2}$$

Consider the event in which (2.2) holds. Fix some $\mathcal{L}_f \in \mathcal{L}_F$. If $P\mathcal{L}_f \leq \alpha$ then $\mathcal{L}_f \in V_\alpha$ and so

$$P_n\mathcal{L}_f - \frac{\alpha}{2} \leq P\mathcal{L}_f \leq P_n\mathcal{L}_f + \frac{\alpha}{2}$$

and (2.1) holds. If, on the other hand, $P\mathcal{L}_f = \beta > \alpha$ then let $g = \frac{\alpha}{\beta}\mathcal{L}_f$ and note that $g \in V_\alpha$. Thus, by (2.2),

$$\frac{1}{2}Pg = Pg - \frac{\alpha}{2} \leq P_n g \leq Pg + \frac{\alpha}{2} \leq 2Pg.$$

Since $\mathcal{L}_f$ is a constant multiple of g,

$$\frac{1}{2}P\mathcal{L}_f \leq P_n\mathcal{L}_f \leq 2P\mathcal{L}_f$$

and so (2.1) holds once again.

To conclude, (2.2) implies that (2.1) holds for all $\mathcal{L}_f \in \mathcal{L}_F$. Thus (2.1) holds with probability at least $1 - e^{-u}$. ■

Remark 2.3 *The claim of Theorem 2.2 holds under milder assumptions. Note that the assumption that F is convex is there to ensure that $P\ell_f$ attains a unique minimum in F, and that the excess loss class satisfies a Bernstein type condition: that for every $f \in F$, $\mathbb{E}\mathcal{L}_f^2 \leq C\mathbb{E}\mathcal{L}_f$. One can show that if F is convex then for any function $f \in F$, $\mathbb{E}\mathcal{L}_f^2 \leq c\|f\|_\infty^2\mathbb{E}\mathcal{L}_f$. Hence, if F is convex and $G \subset F$ that contains the minimizer in F of $P\ell_f$, the analog of Theorem 2.2 will be true for $\{\mathcal{L}_g : g \in G\}$.*

The first part in our analysis will be to show that this isomorphic information can be used to derive estimates in regularized learning.

2.1 From Isomorphic information to Regularized Learning

The regularized learning model provides a method for learning in a very large class of functions without suffering a large statistical error. As we mentioned in the introduction, obtaining an "isomorphic" result for a hierarchy of classes can lead to estimates in the regularized learning model. This approach was introduced in [1] and was formulated in the way we will use here in [4]. Since this last article has not yet appeared, we present a proof of the result we need in an appendix.

Let F be a class of functions and suppose there is a collection of subsets $\{F_r; r \geq 1\}$ with the following properties:

1. $\{F_r : r \geq 1\}$ is monotone (that is, whenever $r \leq s$, $F_r \subseteq F_s$);
2. for every $r \geq 1$, there exists a unique element $f_r^* \in F_r$ such that $P\ell_{f_r^*} = \inf_{f \in F_r} P\ell_f$;
3. the map $r \to P\ell_{f_r^*}$ is continuous;
4. for every $r_0 \geq 1$, $\bigcap_{r > r_0} F_r = F_{r_0}$; and
5. $\bigcup_{r \geq 1} F_r = F$.

Definition 2.4 *Given a class of functions F, we say that $\{F_r; r \geq 0\}$ is an* ordered, parameterized hierarchy of F *if conditions 1-5 are satisfied. Define, for $f \in F$,*

$$r(f) = \inf\{r \geq 1; f \in F_r\}.$$

Note that from the semi-continuity property of an ordered, parameterized hierarchy (property 4), it follows that $f \in F_{r(f)}$ for all $f \in F$.

From the second property of an ordered, parameterized hierarchy, we can define, for $r \geq 1$ and $f \in F_r$, $\mathcal{L}_{r,f} = (f - Y)^2 - (f_r^* - Y)^2$. That is, $\mathcal{L}_{r,f}$ is the excess loss function with respect to the class F_r.

Theorem 2.5 *There exist absolute constants κ_1 and κ_2 such that the following holds. Suppose that $\{F_r; r \geq 1\}$ is an ordered, parameterized hierarchy and that $\rho_n(r,u)$ is a continuous function (possibly depending on the sample) that is increasing in both r and u. Suppose also that for every $r \geq 1$ and every $u > 0$, with probability at least $1 - \exp(-u)$,*

$$\frac{1}{2} P_n \mathcal{L}_{r,f} - \rho_n(r,u) \leq P\mathcal{L}_{r,f} \leq 2P_n\mathcal{L}_{r,f} + \rho_n(r,u)$$

for all $f \in F_r$.

Then for every $u > 0$, with probability at least $1-\exp(-u)$, any function $\hat{f} \in F$ that minimizes the functional

$$P_n \ell_f + \kappa_1 \rho_n(2r(f), \theta(r(f), u))$$

also satisfies

$$P\ell_{\hat{f}} \leq \inf_{f \in F} \Big(P\ell_f + \kappa_2 \rho_n(2r(f), \theta(r(f), u)) \Big)$$

where

$$\theta(r,x) = x + \ln \frac{\pi^2}{6} + 2 \ln \left(1 + \frac{P\ell_{f_1^*}}{\rho_n(1, x + \log(\pi^2/6))} + \log r \right).$$

Remark 2.6 *In fact, the proof of Theorem 2.5 reveals something slightly stronger: if $\tilde{\rho}_n(r,u)$ is a continuous, increasing function in both variables such that*

$$\tilde{\rho}_n(r,u) \geq \rho_n(2r, \theta(r,u))$$

for every r, u and n then every function $\hat{f}$ that minimizes the functional

$$P_n \ell_f + \kappa_1 \tilde{\rho}_n(r,u)$$

satisfies

$$P\ell_{\hat{f}} \leq \inf_{f \in F} \Big(P\ell_f + \kappa_2 \tilde{\rho}_n(r,u) \Big).$$

In other words, we can always regularize with a larger regularization term; we will obtain a correspondingly larger error bound. We will use this fact later on.

The conclusion of Theorem 2.5 can be reformulated in a way that makes the traditional distinction between the approximation and sample errors more explicit. We begin by defining an approximation error term by

$$\mathcal{A}(r) = \inf_{f \in F_r} P\ell_f.$$

Then $\mathcal{A}(r)-\inf_{f\in F} P\ell_f$ tends to zero as $r\to\infty$ and the rate of this convergence measures how well the ordered, parameterized hierarchy approximates Y. Smale and Zhou [25] study this approximation error in a variety of contexts, including the case in which we are interested: when F_r is the ball of radius $r-1$ in a reproducing kernel Hilbert space.

Corollary 2.7 *Under the assumptions of Theorem 2.5, with probability at least* $1-\exp(-u)$*,*

$$P\ell_{\hat{f}} \leq \inf_{r\geq 1}\Big(\mathcal{A}(r)+\kappa_2\rho_n(2r,\theta(r,u))\Big).$$

Proof. Let $u>0$, fix $\varepsilon>0$ and choose an $s\geq 1$ such that

$$\mathcal{A}(s)+\kappa_2\rho_n(2s,\theta(s,u)) \leq \inf_{r\geq 1}\Big(\mathcal{A}(r)+\kappa_2\rho_n(2r,\theta(r,u))\Big)+\frac{\varepsilon}{2}.$$

Consider $g\in F_s$ such that $P\mathcal{L}_g \leq A(s)+\varepsilon/2$. Since ρ_n is increasing in both its arguments,

$$P\mathcal{L}_g+\kappa_2\rho_n(2r(g),\theta(r(g),u)) \leq \inf_{r\geq 1}\Big(\mathcal{A}(r)+\kappa_2\rho_n(2r,\theta(r,u))\Big)+\varepsilon.$$

But we can find such a function g for every $\varepsilon>0$. Therefore

$$\inf_{f\in F}\Big(P\mathcal{L}_f+\kappa_2\rho_n(2r(f),\theta(r(f),u))\Big) \leq \inf_{r\geq 1}\Big(\mathcal{A}(r)+\kappa_2\rho_n(2r,\theta(r,u))\Big)$$

and the conclusion follows from Theorem 2.5. ■

3 Regularization in Kernel classes

The case that we will be interested in is when F_r is a multiple of the unit ball of a reproducing kernel Hilbert space (RKHS). For more details on properties of a RKHS that are relevant in the context of Learning Theory we refer the reader, for example, to [7].

Let Ω be a compact set, consider $K:\Omega\times\Omega\to\mathbb{R}$, a positive definite, continuous function and without loss of generality, we will assume that $\|K\|_\infty\leq 1$. Let T_K be the corresponding integral operator, $T_K: L_2(\mu)\to L_2(\mu)$, defined by

$$(T_Kf)(x)=\int_\Omega K(x,y)f(y)d\mu(y).$$

By Mercer's Theorem [7], there is an orthonormal basis of eigenfunctions $(\varphi_i)_{i=1}^{\infty}$ of T_K, corresponding to the eigenvalues $(\lambda_i)_{i=1}^{\infty}$ arranged in a non increasing order, such that

$$K(x,y)=\sum_{i=1}^{\infty}\lambda_i\varphi_i(x)\varphi_i(y),$$

both in L_2 and $\mu \times \mu$-almost surely.

The RKHS, which will be denoted throughout by H, can be identified with linear functionals in ℓ_2. Indeed, consider $\Phi(x)=\sum_{i=1}^{\infty}\sqrt{\lambda_i}\varphi_i(x)e_i : \Omega \to \ell_2$. For every $t \in \ell_2$, $f_t(x)=\langle \Phi(x),t\rangle$, and $\|f_t\|_H=\|t\|_{\ell_2}$. In the reverse direction, from the definition of the RKHS one can verify that each $h \in H$ is of the form f_t for some $t \in \ell_2$. Hence, to study properties of a subset of H it is enough to study the corresponding set of linear functionals, as a set $T \subset \ell_2$ uniquely determines $F_T=\{f_t : t \in T\}$. Here, we will be mostly concerned with $T=rB_2$, corresponding to $F=rB_H$, where B_H is the unit ball of the RKHS. In this case, the measure endowed on ℓ_2 is given by $\Phi(Z)$, where Z is distributed in Ω according to μ.

3.1 Classes of linear functionals — the L_∞ approach

Our first approach to the problem of regularized learning in an RKHS will lead to a regularization term of $\|f\|_H^2$. As we said in the introduction, this is over-regularization, which is an artifact of the analysis of the learning problem. It stems from the way the L_∞ bound on functions in $\mathcal{L}_F$ is used, and since the only way to bound $\|\mathcal{L}_f\|_{L_\infty}$ is by $\|\mathcal{L}_f\|_{L_\infty} \leq c\|f\|_H^2$. In this section we will use this (loose) approach, but still obtain better error estimates than those previously known — though still using a regularization term of $\|f\|_H^2$. We will obtain considerably better results in the following sections.

The idea we will use is to obtain an isomorphic result for the hierarchy $F_r=rB_H$ (in our ℓ_2 representation, F_r corresponds to rB_2). We then use Corollary 2.7 for the function ρ_n given by the isomorphic analysis.

In our presentation, we will study the following, more general, situation. Let $T \subset \ell_2$ be a compact, convex, symmetric set and consider a random vector X on ℓ_2. Denote by $f_t=\langle t,\cdot\rangle$ the linear functional defined by t and put

$$D=\{t:\mathbb{E}f_t^2 \leq 1\}=\{t:\mathbb{E}\langle t,X\rangle^2 \leq 1\}.$$

Thus, D represents the L_2 unit ball in the parameter space ℓ_2.

Our first, L_∞-based approach to the problem of learning in an RKHS relies on the following bound, which was implicit in [18].

Theorem 3.1 *There exist constants c and c' depending only on $\|Y\|_\infty$ for which the following holds. Let $V_{r,\lambda} = \{\alpha \mathcal{L}_f ; 0 \le \alpha \le 1, f \in rB_H, \mathbb{E}\mathcal{L}_f \le \lambda\}$. Then for every $r \ge 1$ and every $\lambda > 0$,*

$$\mathbb{E}\|P - P_n\|_{V_{r,\lambda}} \le cr\mathbb{E} \sup_{\{t \in rB_2 \cap \sqrt{\lambda}D\}} \left| \frac{1}{n} \sum_{i=1}^n g_i f_t(X_i) \right|$$

where the g_i are independent standard Gaussian variables. In the case where $r = 1$,

$$\mathbb{E} \sup_{\{t \in B_2 \cap \sqrt{\lambda}D\}} \left| \sum_{i=1}^n g_i f_t(X_i) \right| \le c' \left(\frac{1}{n} \sum_{i=1}^\infty \min\{\lambda, \lambda_i\} \right)^{1/2} .$$

The proof of the first part of Theorem 3.1 uses a comparison theorem, relating the Gaussian process $t \to \sum_{i=1}^n g_i \mathcal{L}_{f_t}(X_i, Y_i)$, conditioned on $(X_i, Y_i)_{i=1}^n$, to the conditioned Gaussian process $t \to \sum_{i=1}^n g_i f_t(X_i)$. This is done using an L_∞ bound, since

$$\begin{aligned}\sum_{i=1}^n \left(\mathcal{L}_{f_t} - \mathcal{L}_{f_s}\right)^2 (X_i, Y_i) &= \sum_{i=1}^n (f_t - f_s)^2(X_i) \cdot \left(\left(f_t + f_s\right)(X_i) - 2Y_i\right)^2 \\ &\le 4(r + \|Y\|_\infty)^2 \sum_{i=1}^n (f_t - f_s)^2(X_i),\end{aligned}$$

which will turn out to be the main source of the quadratic regularization term $\|f\|_H^2$.

From Theorem 3.1 one obtains

Corollary 3.2 *There exists a constant c, depending only on $\|Y\|_\infty$ such that, if $x > 0$ satisfies that*

$$x \ge c \left(\frac{1}{n} \sum_{i=1}^\infty \min\{x, \lambda_i\} \right)^{1/2}$$

then, for all $r \ge 1$,

$$\frac{\lambda}{8} \ge \mathbb{E}\|P - P_n\|_{V_{r,\lambda}}$$

where $\lambda = r^2 x$.

Proof. Define

$$\psi_r(x) = r\mathbb{E} \sup_{\{t \in rB_2 \cap \sqrt{x}D\}} \left| \sum_{i=1}^{n} g_i f_t(X_i) \right|.$$

Then $\psi_1(x) \leq c'x$ by the second part of Theorem 3.1 and it is easy to check that $\psi_r(x) = r^2\psi_1(xr^{-2})$ for any x and r. That is, $\psi_r(r^2x) = r^2\psi_1(x) \leq c'r^2x$. The claim now follows from the first part of Theorem 3.1. ■

With this Corollary and Theorem 2.2, we can obtain an isomorphic condition on the unit ball of an RKHS using information on the decay of the eigenvalues. For the sake of concreteness, we will make the following assumption on this rate of decay; this assumption will allow us to compute an error bound explicitly.

Definition 3.3 *For $0 < p < 1$ define*

$$\|(\lambda_i)\|_{p,\infty} = \sup_{x>0} x^p |\{\lambda_i \geq x\}|.$$

Hence, for any $x > 0$,

$$|\{\lambda_i \geq x\}| \leq \|(\lambda_i)\|_{p,\infty} x^{-p}. \tag{3.1}$$

Assumption 3.1 *Let K be a kernel on a compact probability space $(\Omega \times \Omega, \mu \times \mu)$ where μ is a Borel measure and $\Omega \subset \mathbb{R}^d$. Assume that the eigenvalues of the kernel satisfy that $(\lambda_n)_{n=1}^{\infty} \in \ell_{p,\infty}$ for some $0 < p < 1$ and that $\|K(x,x)\|_\infty \leq 1$.*

Since $\int K(x,x)d\mu(x) = \sum_{i=1}^{\infty} \lambda_i$, then $(\lambda_i) \in \ell_{1,\infty}$ when $K(x,x) \in L_1(\mu)$. The stronger Assumption 3.1 is satisfied under some smoothness condition on the kernel. For example, if the kernel K belongs to some Besov space $B^{\alpha}_{2,\infty}$ (in particular, this is the case if $\alpha \in \mathbb{N}$ and $K \in \mathcal{C}^{\alpha}(\Omega \times \Omega)$) then, by Theorem 4.1 of [5] (see also [14]), the sequence (λ_i) belongs to $\ell_{p,\infty}$ for

$$p = \frac{1}{\alpha/d + 1/2}.$$

In fact, the result of [5] is slightly stronger — the sequence $\lambda_n n^{-1/p}$ tends to 0 with n — but we will not need this strengthening. The L_∞ assumption on $K(x,x)$ is only to simplify the presentation and any uniform bound instead of 1 would do.

The assumption on the rate of decay of the eigenvalues allows us to obtain the following bound:

Lemma 3.4 *For $0 < p \leq 1$ there is a constant c_p depending only on p such that for all $x > 0$ and all $r > 0$,*

$$\sum_{i=1}^{\infty} \min\{x, r^2\lambda_i\} \leq c_p \|(\lambda_i)\|_{p,\infty} x^{1-p} r^{2p}.$$

Proof. It suffices to prove the lemma for $r = 1$ and the result will follow for all r by homogeneity. Set $N_x = |\{\lambda_i \geq x\}|$ and observe that for all $x > 0$,

$$\sum_{i=1}^{\infty} \min\{x, \lambda_i\} = xN_x + \sum_{i=N_x+1}^{\infty} \lambda_i \leq \|(\lambda_i)\|_{p,\infty} x^{1-p} + \sum_{i=N_x+1}^{\infty} \lambda_i.$$

To estimate the second term, let $\{a_j\}_{j=0}^{\infty}$ be any decreasing sequence with $a_0 = x$. Then

$$\begin{aligned} \sum_{i=N_x+1}^{\infty} \lambda_i &= \sum_{\lambda_i < x} \lambda_i \\ &\leq \sum_{j=0}^{\infty} a_j |\{\lambda_i \geq a_{j+1}\}| \\ &\leq \|(\lambda_i)\|_{p,\infty} \sum_{j=0}^{\infty} a_j a_{j+1}^{-p}. \end{aligned}$$

Now set $a_j = x2^{-j}$ and note that

$$\sum_{i=N_x+1}^{\infty} \lambda_i \leq 2\|(\lambda_i)\|_{p,\infty} \sum_{j=1}^{\infty} a_j^{1-p} \leq c_p x^{1-p} \|(\lambda_i)\|_{p,\infty},$$

as required. ■

Corollary 3.5 *Let K be a kernel that satisfies Assumption 3.1 for some $0 < p \leq 1$. There exists a constant c_p depending only on p such that if $z = c_p \left(\frac{\|(\lambda_i)\|_{p,\infty}}{n}\right)^{1/(1+p)}$ then, for all $r > 1$,*

$$\frac{\lambda}{8} \geq \mathbb{E}\|P - P_n\|_{V_{r,\lambda}}$$

where $\lambda = r^2 z$.

Lemma 3.6 *Let H be the reproducing kernel Hilbert space associated to a kernel K, set $F = H$, and define, for every $r \geq 1$, $F_r = (r-1)B_H$ where B_H is the unit ball of H. Then $\{F_r; r \geq 1\}$ is an ordered, parameterized hierarchy and $r(f) = \|f\| + 1$.*

Proof. The first, fourth and fifth properties of an ordered, parameterized hierarchy are immediate. The second property follows from the fact that B_H is convex and compact with respect to the L_2 norm. For the third property, fix $1 \leq q < r < s$ and let $\beta = \frac{r-1}{q-1}$ and $\alpha = \frac{r-1}{s-1}$. Note that $\alpha f_q^* \in F_r$ and $\beta f_s^* \in F_r$. Thus,

$$0 \leq P\ell_{f_r^*} - P\ell_{f_s^*} \leq P\ell_{\alpha f_s^*} - P\ell_{f_s^*} = (\alpha^2 - 1)P(f_s^*)^2 + 2(1-\alpha)Pf_s^*Y.$$

As $s \to r$, the right hand side tends to zero (because the candidates for f_s^* are uniformly bounded in L_2) and so $r \to P\ell_{f_r^*}$ is upper semi-continuous (the same argument works for $r = 1$). In the other direction,

$$0 \leq P\ell_{f_q^*} - P\ell_{f_r^*} \leq (\beta^2 - 1)P(f_r^*)^2 + 2(1-\beta)Pf_r^*Y$$

and the right hand side tends to zero for the same reason as before. ■

Combining Theorem 2.2 with Corollaries 3.2 and 2.7, we obtain the following error bound for regularized learning in an RKHS.

Theorem 3.7 *There exist absolute constants κ_1 and κ_2, constants c_Y and c'_Y depending only on $\|Y\|_\infty$ and a constant c_p depending only on p such that the following holds. Let K be a kernel satisfying Assumption 3.1 and define*

$$\rho_n(r,u) = c_p r^2 \left(\frac{\|(\lambda_i)\|_{p,\infty}}{n}\right)^{1/(1+p)} + c_Y(1+r^2)\frac{u}{n}.$$

Then for every $u > 0$, with probability at least $1 - \exp(-u)$, any function $\hat{f} \in F$ that minimizes the functional

$$P_n\ell_f + \kappa_1 \tilde{\rho}_n(r(f), u)$$

also satisfies

$$P\ell_{\hat{f}} \leq \inf_{r \geq 1}(\mathcal{A}(r) + \kappa_2 \tilde{\rho}_n(r,u))$$

where

$$\tilde{\rho}_n(r,u) = \rho_n\left(2r, u + \ln\frac{\pi^2}{6} + 2\ln(1 + c'_Y n + \log r)\right).$$

In particular,

$$P\ell_{\hat{f}} \leq \inf_{r\geq 1}\left(\mathcal{A}(r) + c\left(\frac{r^2}{n^{1+1/p}} + \frac{1+r^2}{n}\left(u + \log n + \log\log(r+e)\right)\right)\right),$$

where $c = c(p, \|Y\|_\infty, \|(\lambda_i)\|_{p,\infty})$.

Proof. By Theorem 2.2 and Corollary 3.2, the function $\rho_n(r,x)$ satisfies the condition of Theorem 2.5 (where we set $c_Y = c\|Y\|_\infty$). Then we can apply Corollary 2.7 to obtain the result. Since $0 \in F_r$ for any $r > 0$ then $P\ell_{f_1^*} \leq P\ell_0 = \|Y\|^2_{L_2(\mu)}$ and $\rho_n(1, u + \ln(\pi^2/6)) \geq c''_Y/n$ so that

$$\frac{P\ell_{f_1^*}}{\rho_n(1, x + \ln(\pi^2/6))} \leq c'_Y n,$$

to which we apply Remark 2.6. ■

Let us compare the estimate on the regularization term and the resulting error rate that follows from this theorem to previously obtained bounds on regularized learning in an RKHS. Since all of the results we consider have exponentially good confidence, we will simplify this comparison by ignoring the confidence term and focusing on the decay of the error bound as the sample size increases.

In 2002, Cucker and Smale [7] used covering numbers to bound the sample error of the regularized minimizer by

$$\frac{c_Y\sqrt{\ln n}(\gamma + c)}{\gamma^2\sqrt{n}}$$

where γ is the regularization parameter. Noting that the optimal γ tends to zero as the sample size increases, we will ignore the γ term in the numerator. By Lemma 2 of [8], the approximation error with a regularization term of γ is at least $\mathcal{A}(c/\gamma + 1)$ (Cucker and Smale give a more detailed analysis of the approximation error, which we will not use for the sake of simplicity. In order to prevent this simplification from biasing our comparison, note that we are using a *lower* bound on the approximation error in the result of [8]).

With the substitution $r = c/\gamma + 1$, the total error bound of [8] becomes

$$P\ell_{\hat{f}} \leq \inf_{r\geq 1}\left(\mathcal{A}(r) + cr^2\left(\frac{\ln n}{n}\right)^{1/2}\right).$$

On the other hand, our bound is

$$P\ell_{\hat{f}} \leq \inf_{r\geq 1}\left(\mathcal{A}(r) + cr^2\left(\frac{1}{n}\right)^{1/(1+p)} + c'(1+r^2)\frac{\ln n + \ln(1+\log r)}{n}\right).$$

Note that the $n^{-1}r^2 \ln \log r$ term is asymptotically insignificant. Indeed, for the optimal $r = r(n)$, our bound tends to $\lim_{r\to\infty} \mathcal{A}(r)$ with n. In particular, $n^{-1/(1+p)}r^2(n) \to 0$ and so $\ln \log r(n) \ll \ln n$. By eliminating the $\ln \log r$ term, our result is directly comparable to that of [8]. In the (worst possible) case where $p = 1$, we gain by removing a relatively insignificant factor of $\sqrt{\log n}$. For smaller p, however, we improve Cucker and Smale's result by a polynomial factor of n. Given that it is common in Machine Learning for the kernel associated to an RKHS to have some smoothness properties, this is a significant improvement.

It is well known that by making an assumption on the regression function $\mathbb{E}(Y|X)$, it is possible to bound the approximation error and thereby obtain to bound the total error in terms of the sample size. A result of this sort was first obtained in [8] and improved in [26]. In particular, Corollary 5 of [26] gives the following bound: suppose that $\mathbb{E}(Y|X)$ is in the range of T_K^σ for some $0 < \sigma < 1$. Then for sufficiently large n,

$$P\ell_{\hat{f}} - \inf_{f \in F} P\ell_f \lesssim \begin{cases} \left(\frac{1}{n}\right)^{\sigma/(1+2\sigma)}, & \text{if } \sigma \geq \frac{1}{2} \\ \left(\frac{1}{n}\right)^{\sigma/2}, & \text{if } \sigma < \frac{1}{2}, \end{cases} \tag{3.2}$$

(where the regularization term used to obtain this result is $\sim r^2(1/n)^{1/(1+2\sigma)}$ if $\sigma \geq 1/2$ and $\sim r^2(1/n)^{1/2}$ otherwise).

In comparison, our regularization term is of the order of $\sim r^2/n^{1/1+p}$. To obtain the resulting error bound we first consider the case where $\sigma \geq \frac{1}{2}$. Then $\mathbb{E}(Y|X) \in F$ and so $\mathcal{A}(r) - \inf_{f\in F} P\ell_f$ is zero for sufficiently large r. Then we have shown that, for sufficiently large n,

$$P\ell_{\hat{f}} - \inf_{f \in F} P\ell_f \lesssim \left(\frac{1}{n}\right)^{1/(1+p)}$$

which is a significant improvement on (3.2) even in the case $p = 1$.

To deal with the case $\sigma < \frac{1}{2}$, we will use the following theorem of [25]:

Theorem 3.8 *Let A be a compact, symmetric and strictly positive operator on a separable Hilbert space H. Then, for any $0 < \sigma < s$, any $r > 0$ and any a in the range of A^σ,*

$$\inf_{\|A^{-s}b\| \leq r} \|a - b\| \leq \left(\frac{1}{r}\right)^{\sigma/(s-\sigma)} \|A^{-\sigma}a\|^{s/(s-\sigma)}.$$

In particular, we can apply this with $H = L_2$, $A = T_K$, $s = \frac{1}{2}$ and $a = \mathbb{E}(Y|X)$ to obtain

$$\begin{aligned}\mathcal{A}(r-1) - \inf_{f \in F} P\ell_f &= \inf_{\|f\|_F \leq r} \|\mathbb{E}(Y|X) - f\|_2^2 \\ &\leq \left(\frac{1}{r}\right)^{4\sigma/(1-2\sigma)} \|L_K^{-\sigma}\mathbb{E}(Y|X)\|^{2/(1-2\sigma)}.\end{aligned}$$

Set $k = 4\sigma/(1-2\sigma)$. Then we can choose $r = n^{\frac{1}{(1+p)(2+k)}}$ and our error bound becomes

$$\begin{aligned}P\ell_{\hat{f}} - \inf_{f \in F} P\ell_f &\lesssim \mathcal{A}\left(n^{\frac{1}{(1+p)(2+k)}}\right) + n^{\frac{2}{(1+p)(2+k)}} \left(\frac{1}{n}\right)^{1/(p+1)} \\ &\lesssim \left(\frac{1}{n}\right)^{\frac{k}{(1+p)(2+k)}} \\ &= \left(\frac{1}{n}\right)^{2\sigma/(1+p)}. \qquad (3.3)\end{aligned}$$

Once again, this is an improvement over (3.2) even in the case $p = 1$.

4 Towards a smaller regularization parameter

The bound (3.3) would be substantially improved if we could remove the r^2 term and replace it by a smaller power of r — which is the main novelty in this article. As we mentioned before, the most significant source for this improvement comes from bypassing L_∞-based bounds. In recent years there has been considerable progress made on bounding various empirical processes that are indexed by sets that are either not bounded or very weakly bounded in L_∞. Most of these results were motivated by questions in Asymptotic Geometric Analysis, most notably, sampling from an isotropic, log-concave measure (e.g. [24, 13, 21]) and the approximate reconstruction problem [19, 12]. The fact that such an approach is called for here seems strange because we are dealing with a learning problem relative to a class of uniformly bounded functions, so it would seem that there is no reason to employ techniques designed to handle an unbounded situation. Even more so, because in a standard learning analysis the way the error bounds depend on the L_∞ diameter of the class is usually of no real importance. In contrast, here, the way the isomorphic results scale with the L_∞ bound is extremely important because one is trying to obtain a result for the entire hierarchy, and the L_∞ diameter of F_r is directly linked to the hierarchy parameter r. Thus, the standard, and very loose approach which is commonly used in a single class situation can cause real damage in our case because the regularization term will be strongly influenced by the way the L_∞ diameter enters into the bounds.

To see where one can improve upon the standard L_∞ analysis (in a very "hand waving" way), let us return to the localized Gaussian process indexed by $\{t : \mathbb{E}\mathcal{L}_{f_t} \leq \lambda\} \cap rB_2$, conditioned on the data (X_i, Y_i), that is,

$$t \to \sum_{i=1}^{n} g_i \mathcal{L}_{f_t}(X_i, Y_i) = \sum_{i=1}^{n} g_i \langle t - \beta^*, X_i \rangle \left(\langle t + \beta^*, X_i \rangle - 2Y_i \right),$$

where f_{β^*} minimizes the loss in rB_2. For every t the variance of each conditioned Gaussian variable satisfies

$$\sigma^2 (\sum_{i=1}^{n} g_i \mathcal{L}_{f_t}(X_i, Y_i)) = \sum_{i=1}^{n} \langle t - \beta^*, X_i \rangle^2 \left(\langle t + \beta^*, X_i \rangle - 2Y_i \right)^2.$$

Consider some t for which $\mathbb{E}\mathcal{L}_{f_t} \leq \lambda$. One can show that in this case, $\|t - \beta^*\| \leq \sqrt{\lambda}$ (see Lemma 4.1 below). Now, if one has a very strong concentration phenomenon and if $D = B_2$ then

$$\begin{aligned} \left(\langle \frac{t + \beta^*}{2}, X_i \rangle - Y_i \right)^2 &= \left(\langle \frac{t - \beta^*}{2}, X_i \rangle + \left(\langle \beta^*, X_i \rangle - Y_i \right) \right)^2 \\ &\approx_c \lambda + \mathbb{E}\ell_{f_{\beta^*}}. \end{aligned}$$

Since the expected loss of the best in the class only decreases with r, this term is of the order of λ, rather than a factor that grows quadratically in r, which is the estimate that results from the L_∞ approach. This at least hints to the fact that the L_∞ approach is likely to lead to very loose estimates.

Despite the fact that above paragraph is totally unjustified as stated and very optimistic, it turns out that this scenario is very close to the actual situation (although the proof requires a rather delicate analysis).

4.1 Further preliminaries

For technical reasons, we will make an additional assumption on the eigenfunctions of the kernel. We should emphasize that it is possible that this assumption may not be necessary to obtain the improved regularization term, though we were not able to remove it here and it has a crucial role in our analysis.

Assumption 4.1 *Let K be a kernel on a compact probability space $(\Omega \times \Omega, \mu \times \mu)$ with $\Omega \subset \mathbb{R}^d$. Assume that there is a constant A for which the eigenfunctions of K satisfy that* $\sup_n \|\varphi_n\|_\infty \leq A < \infty$.

One case is which this assumption is satisfied is when K is a translation invariant kernel (i.e. $K(x,y) = k(x-y)$ for some function k), Ω is a compact group and μ is the Haar measure on Ω. In this case all the eigenfunctions are characters of the group, and thus uniformly bounded in L_∞.

Recall that the feature map Φ defines an isometry between an RKHS and ℓ_2. Let $T \subset \ell_2$ be a centrally symmetric, convex, compact subset of ℓ_2. The first step in our analysis is to relate the localized sets $\mathcal{L}_\lambda$ (corresponding to the class $\{f_t : t \in T\}$) to subsets of T. Since this fact appeared implicitly in several places (see, for example [20], Cor. 3.4) and in more general situations - for example, loss functions that are uniformly convex rather than the squared loss, we omit its proof.

Lemma 4.1 *Let* $\beta^* = \operatorname{argmin}_{t \in T} \mathbb{E}\ell_{f_t}$. *For every* $\lambda > 0$,

$$\{t - \beta^* : t \in T, \mathcal{L}_{f_t} \in \mathcal{L}_\lambda\} \subset 2\sqrt{\lambda}D \cap 2T.$$

Lemma 4.1 shows that it is sufficient to consider the complexity of the sets $\sqrt{\lambda}D \cap T$. The complexity parameters we shall use come from a generic chaining argument (defined below), and thus a significant part of our analysis will be based on covering numbers.

Definition 4.2 *Let $A, B \subset \ell_2$. Denote by $N(A,B)$ the smallest number of translates of B needed to cover A. If εB is a ball of radius ε with respect to some norm then $N(A, \varepsilon B)$ is the minimal cardinality of an ε-cover of A with respect to the that norm. If (A,d) is a metric space (rather than a normed one), we denote the cardinality of a minimal ε-cover of A by $N(A, \varepsilon, d)$.*

The generic chaining mechanism (see [29] for the most recent survey on this topic) is used to relate probabilistic properties of a random process indexed by a metric space to the metric structure of the underlying space. This mechanism originated in the study of Gaussian processes $t \to X_t$ where it was proved that $\mathbb{E}\sup_{t \in T} X_t$ is equivalent to a metric invariant of (T,d), for $d(s,t) = (\mathbb{E}|X_s - X_t|^2)^{1/2}$. This so-called *majorizing measures* Theorem (in which the upper bound of the equivalence was proved by Fernique [10] and the lower by Talagrand [27]) was later developed to a more general theory with many interesting applications [29]. The metric invariant that is at the heart of this theory is the γ_2 functional.

Let (T,d) be a metric space. An *admissible sequence* of T is a collection of subsets of T, $\{T_s : s \geq 0\}$, such that for every $s \geq 1$, $|T_s| = 2^{2^s}$ and $|T_0| = 1$.

Definition 4.3 *For a metric space (T,d) define*

$$\gamma_2(T,d) = \inf \sup_{t \in T} \sum_{s=0}^{\infty} 2^{s/2} d(t, T_s),$$

where the infimum is taken with respect to all admissible sequences of T.

Definition 4.4 *A random process $t \rightarrow X_t$ indexed by a metric space (T,d) is subgaussian relative to d if for every $s,t \in T$ and every $u \geq 1$,*

$$Pr\left(|X_s - X_t| \geq u d(s,t)\right) \leq 2 \exp\left(-\frac{u^2}{2}\right).$$

The generic chaining mechanism can be used to show that if $\{X_t : t \in (T,d)\}$ is subgaussian then there is an absolute constant c such that for every $t_0 \in T$,

$$\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}| \leq c\gamma_2(T,d),$$

and similar bounds hold with high probability.

Note that one choice for sets T_s that constitute a potential (yet, usually suboptimal) admissible sequence are ε_s-covers of T, where each ε_s is selected in a way that ensures that $N(T, \varepsilon_s, d) \leq 2^{2^s}$. Then, an easy computation [29] shows that

$$\gamma_2(T,d) \leq c \int_0^{\mathrm{diam}(T,d)} \sqrt{\log N(T,\varepsilon,d)} d\varepsilon, \tag{4.1}$$

where c is an absolute constant. This is a generalization of Dudley's entropy integral (see for example [9, 29]), used in the study of Gaussian processes. As will be explained later, this integral bound can be improved under certain assumptions on geometry of T if d is endowed by a norm.

The metric d we will focus on here is a random one and depends on the sample $X_1,...,X_n \subset \ell_2$. For every $X_1,...,X_n$ set

$$d_{\infty,n}(f,g) = \max_{1 \leq i \leq n} |f(X_i) - g(X_i)|.$$

In our case, $f = f_s$ and $g = g_t$, and thus $d_{\infty,n}$ defines a random norm on a projection of ℓ_2 — since $\max_{1 \leq i \leq n} |f_s(X_i) - g_t(X_i)| = \max_{1 \leq i \leq n} |\langle X_i, s - t \rangle|$. Next, let $U_n(T) = (\mathbb{E}\gamma_2^2(T, d_{\infty,n}))^{1/2}$ and for every $x > 0$ set

$$\phi_n(x) = \frac{U_n(K_x)}{\sqrt{n}} \cdot \max\left(\sqrt{x}, \sqrt{\mathbb{E}\mathcal{L}_{\beta^*}}, \frac{U_n(K_x)}{\sqrt{n}}\right),$$

where $K_x = T \cap \sqrt{x} D \subset \ell_2$ and β^* is the parameter in T for which $\inf_{t \in T} \mathbb{E}\mathcal{L}_{f_t}$ is attained.

Recall that

$$\mathcal{L}_\lambda = \{\mathcal{L}_f : \mathbb{E}\mathcal{L}_f \leq \lambda\},$$

and that

$$V_\lambda = \{\theta \mathcal{L}_f : 0 \leq \theta \leq 1,\ \mathbb{E}(\theta \mathcal{L}_f) \leq \lambda\} = \{h \in \text{star}(\mathcal{L}_F, 0) : \mathbb{E}h \leq \lambda\}.$$

From Theorem 2.2 it is clear that in order to obtain a useful "isomorphic" result one has to bound $\mathbb{E}\|P_n - P\|_{V_\lambda}$ as a function of λ; this is done in the following theorem. Since it is a modification of a result that was proved in [4] we will only present an outline of its proof.

Theorem 4.5 *There exists an absolute constant c for which the following holds. If T and X are as above then for every $\lambda > 0$,*

$$\mathbb{E}\|P_n - P\|_{V_\lambda} \leq c \sum_{i=0}^{\infty} 2^{-i} \phi_n(2^{i+1}\lambda).$$

Lemma 4.6 *For every $\lambda > 0$,*

$$\mathbb{E}\|P_n - P\|_{V_\lambda} \leq 2 \sum_{i=0}^{\infty} 2^{-i} \mathbb{E}\|P_n - P\|_{\mathcal{L}_{2^{i+1}\lambda}}.$$

Proof. Note that for every $\lambda > 0$,

$$\begin{aligned} W_\lambda &= \{\theta \mathcal{L}_f : 0 \leq \theta \leq 1,\ \mathbb{E}(\theta \mathcal{L}_f) \leq \lambda,\ \mathbb{E}\mathcal{L}_f \geq \lambda\} \\ &= \left\{\frac{t\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} : \mathbb{E}\mathcal{L}_f \geq \lambda,\ 0 \leq t \leq \lambda\right\} \\ &= \bigcup_{i=0}^{\infty} \left\{\frac{t\mathcal{L}_f}{\mathbb{E}\mathcal{L}_f} : 2^i\lambda \leq \mathbb{E}\mathcal{L}_f \leq 2^{i+1}\lambda,\ 0 \leq t \leq \lambda\right\} \equiv \bigcup_{i=0}^{\infty} W_{i,\lambda}. \end{aligned}$$

If $t\mathcal{L}_f / \mathbb{E}\mathcal{L}_f \in W_{i,\lambda}$ then $t/\mathbb{E}\mathcal{L}_f \leq 2^{-i}$ and $\mathcal{L}_f \in \mathcal{L}_{2^{i+1}\lambda}$. Thus, $\|P_n - P\|_{W_{i,\lambda}} \leq 2^{-i}\|P_n - P\|_{\mathcal{L}_{2^{i+1}\lambda}}$.

Finally, let $W_{0,\lambda} = \text{star}(\mathcal{L}_\lambda, 0)$. Note that $\|P_n - P\|_{W_{0,\lambda}} \leq \|P_n - P\|_{\mathcal{L}_\lambda}$, and that $V_\lambda \subset W_0 \cup W_{0,\lambda}$, from which our claim follows. ■

Outline of the proof of Theorem 4.5. Fix $\lambda > 0$. First of all, one can verify that the Bernoulli process indexed by $\mathcal{L}_\lambda$, given by $t \rightarrow \sum_{i=1}^n \varepsilon_i \mathcal{L}_{f_t}(X_i, Y_i)$ conditioned on $(X_i, Y_i)_{i=1}^n$ is subgaussian with respect to the metric

$$d(f_{t_1}, f_{t_2}) = d_{\infty,n}(f_{t_1}, f_{t_2}) \left(\sup_{v \in \sqrt{\lambda} D \cap T} \sum_{i=1}^n \langle X_i, v \rangle^2 + \sum_{i=1}^n \mathcal{L}_{\beta^*}(X_i, Y_i) \right)^{1/2}$$

Hence, if we set $K = \sqrt{\lambda} D \cap T$, then by the Giné-Zinn symmetrization method [11] followed by a generic chaining argument,

$$\mathbb{E}\|P_n - P\|_{\mathcal{L}_\lambda} \leq \frac{c_1}{n} \mathbb{E} \left(\gamma_2(K, d_{\infty,n}) \left(\sup_{t \in K} \sum_{i=1}^n \langle t, X_i \rangle^2 + \sum_{i=1}^n \mathcal{L}_{\beta^*}(X_i, Y_i) \right)^{1/2} \right).$$

Moreover, one can show (see, for example, [12]) that if H is a class of functions then

$$\mathbb{E} \sup_{h \in H} \left| \sum_{i=1}^n h^2(X_i) - \mathbb{E}h^2 \right| \leq c_2 \max\left\{ \sqrt{n} \sigma_H U_n(H), U_n^2(H) \right\},$$

where $\sigma_H^2 = \sup_{h \in H} \mathbb{E}h^2$. In particular, for $H = \{\langle t, \cdot \rangle : t \in K\}$,

$$\mathbb{E} \sup_{t \in K} \sum_{i=1}^n \langle t, X_i \rangle^2 \leq n\lambda + c_2 \max\{\sqrt{n\lambda} U_n(K), U_n^2(K)\},$$

because $\mathbb{E}\langle t, \cdot \rangle^2 \leq \lambda$. Now, a straightforward computation shows that

$$\mathbb{E}\|P_n - P\|_{\mathcal{L}_\lambda} \leq \phi_n(\lambda).$$

To conclude the proof, note that by Lemma 4.6 it is possible to estimate $\mathbb{E}\|P_n - P\|_{V_\lambda}$ using $\mathbb{E}\|P_n - P\|_{\mathcal{L}_{2^i\lambda}}$. ■

Observe that the sets T we will be interested in are rB_2 since they are the images of rB_H in ℓ_2. The rest of this section will be devoted to finding a bound on $\phi_n(x)$ for these sets T.

4.2 Controlling ϕ_n for $T = rB_2$

It is clear that ϕ_n is determined by the structure of the sets $K_{x,r} = \sqrt{x} D \cap 2rB_2 \subset \ell_2$. To study the metric properties of these sets we first have to identify D.

Consider the random variable Z on Ω distributed according to μ and let $X = \Phi(Z) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\varphi_i(Z)e_i \in \ell_2$ be the random feature map. Clearly,

$$D = \{\beta \in \ell_2 : \mathbb{E}\big\langle \beta, X \big\rangle^2 \le 1\} = \{\beta \in \ell_2 : \mathbb{E}\big\langle \beta, \Phi(Z) \big\rangle^2 \le 1\}.$$

Since $(\varphi_i)_{i=1}^{\infty}$ is an orthonormal system in $L_2(\mu)$ then

$$\mathbb{E}\big\langle \beta, \Phi(Z) \big\rangle^2 = \mathbb{E} \sum_{i,j} \beta_i \beta_j \sqrt{\lambda_i \lambda_j} \varphi_i(Z)\varphi_j(Z) = \sum_{i=1}^{\infty} \lambda_i \beta_i^2.$$

Hence, D is an ellipsoid in ℓ_2 with the standard basis $(e_i)_{i=1}^{\infty}$ as principal directions, and lengths $1/\sqrt{\lambda_i}$.

It is straightforward to verify that for every $x, r > 0$ there is an ellipsoid $\mathcal{E}_{x,r}$, such that $K_{x,r} = 2rB_2 \cap \sqrt{x}D$ satisfies $\frac{1}{2}\mathcal{E}_{x,r} \subset K_{x,r} \subset \mathcal{E}_{x,r}$. The principal directions of $K_{x,r}$ and $\mathcal{E}_{x,r}$ coincide and the principal lengths of $\mathcal{E}_{x,r}$ are

$$c \min\left\{\sqrt{\frac{x}{\lambda_i}}, r\right\},$$

where c is an absolute constant.

The structure of the ellipsoids $\mathcal{E}_{x,r}$ indicates that it should be possible to obtain a sub-linear dependency on the radius r and the fact that we were not able to do so in Section 3.1 is an artifact of the suboptimal analysis that was used there. The sub-linearity occurs because for $\alpha > 1$, $\mathcal{E}_{x,\alpha r}$ is much smaller than $\alpha\mathcal{E}_{x,r}$; since it is an intersection body, it only grows in some directions, and the number of directions in which it grows decreases quickly with r.

Now that we have identified the intersection body we are ready to estimate

$$U_n = \left(\mathbb{E}\gamma_2^2(\mathcal{E}_{x,r}, d_{\infty,n})\right)^{1/2}.$$

Theorem 4.7 *There exists an absolute constant c for which the following holds. Suppose* $\sup_n \|\varphi_n\|_\infty \le A$ *and set*

$$Q(x,r) = A\left(\sum_{i=1}^{\infty} \min\{x, r^2\lambda_i\}\right)^{1/2}.$$

Then

$$\left(\mathbb{E}\gamma_2^2(\mathcal{E}_{x,r}, d_{\infty,n})\right)^{1/2} \le cQ \log n.$$

Before proving the Theorem we need two additional facts. The first is an improved "Dudley entropy integral" bound, due to Talagrand.

Theorem 4.8 *[29] There exists an absolute constant c for which the following holds. If $\mathcal{E} \subset \ell_2^m$ is an ellipsoid and B is the unit ball of some norm $\| \ \|$ on $\mathbb{R}^m$ then*

$$\gamma_2(\mathcal{E}, \| \ \|) \leq c \left(\int_0^\infty \varepsilon \log N(\mathcal{E}, \varepsilon B) d\varepsilon \right)^{1/2}.$$

Another standard fact we need is the dual Sudakov inequality [22].

Lemma 4.9 *There exists an absolute constant c for which the following holds. Let B_E be a unit ball of some norm on $\mathbb{R}^m$. Then, for every $\varepsilon > 0$,*

$$\log N(B_2^m, \varepsilon B_E) \leq c \left(\frac{\mathbb{E}\|G\|_E}{\varepsilon} \right)^2,$$

where $G = (g_1, ..., g_m)$ is a standard Gaussian vector on $\mathbb{R}^m$.

Proof of Theorem 4.7. Fix $X_1, ..., X_n$ and note that in order to bound $\gamma_2(\mathcal{E}_{x,r}, d_{\infty,n})$ it suffices to consider the projection of the (infinite dimensional) ellipsoid $\mathcal{E}_{x,r}$ onto the subspace spanned by $X_1, ..., X_n$. Hence, one can apply Lemma 4.9. Set $\|v\|_E = \max_{1 \leq i \leq n} |\langle v, X_i \rangle|$ and let B_E be the unit ball $\{v \in \ell_2 : \|v\|_E \leq 1\}$. Consider the ellipsoid $\mathcal{E}_{x,r} \subset \ell_2$ with principal directions $(e_i)_{i=1}^\infty$ and lengths $\theta_i = c_1 \min \left\{ \sqrt{x/\lambda_i}, r \right\}$. Let T be the operator $Te_i = \theta_i e_i$. Thus, $TB_2 = \mathcal{E}_{x,r}$, for every $\varepsilon > 0$,

$$N(TB_2, \varepsilon B_E) = N(B_2, \varepsilon T^{-1} B_E),$$

and $v \in \varepsilon T^{-1} B_E$ if and only if $\max_{1 \leq i \leq n} |\langle v, T^* X_i \rangle| = \max_{1 \leq i \leq n} |\langle v, TX_i \rangle| \leq \varepsilon$. Hence, if we set $W_i = TX_i$ and $B_{\bar{E}} = \{v : \max_{1 \leq i \leq n} |\langle v, W_i \rangle| \leq 1\}$ then

$$N(TB_2, \varepsilon B_E) = N(B_2, \varepsilon B_{\bar{E}}) = N(B_2^n, \varepsilon B_{\bar{E}}),$$

where here we mean by B_2^n the unit ball in the subspace of ℓ_2 spanned by $(W_i)_{i=1}^n$.

Let G be a standard Gaussian vector on $\mathbb{R}^n$. Then, by Slepian's Lemma [9, 23],

$$\mathbb{E}\|G\|_{\bar{E}} = \mathbb{E} \max_{1 \leq i \leq n} |\langle G, TX_i \rangle| \leq c_2 \sqrt{\log n} \max_{1 \leq i \leq n} \|TX_i\|_2.$$

Since T is a diagonal operator and $X_j = \sum_{i=1}^\infty \sqrt{\lambda_i} \varphi_i(Z_j) e_i$ then

$$\|TX_j\|_2^2 = \sum_{i=1}^\infty \theta_i^2 \lambda_i \varphi_i^2(Z_j) \leq A^2 \sum_{i=1}^\infty \theta_i^2 \lambda_i = A^2 \sum_{i=1}^\infty \min \left\{ x, r^2 \lambda_i \right\}.$$

Hence, setting

$$Q(x,r) = A\left(\sum_{i=1}^{\infty} \min\left\{x, r^2\lambda_i\right\}\right)^{1/2},$$

it is evident that

$$\mathbb{E}\|G\|_{\bar{E}} \leq c_2\sqrt{\log n}Q, \tag{4.2}$$

and by Lemma 4.9, for every $\varepsilon > 0$

$$\log N(B_2^n, \varepsilon B_{\bar{E}}) \leq c_3\frac{Q^2\log n}{\varepsilon^2}.$$

In particular, the diameter of B_2^n with respect to the norm $\| \ \|_{\bar{E}}$ is at most $cQ\sqrt{\log n}$, and we denote this diameter by D_2.

This estimate on the covering numbers will be used for "large" scales of ε. For smaller scales we need a different argument. Applying a volumetric estimate (see, e.g. [23]), for every norm $\| \ \|_X$ on $\mathbb{R}^n$ and every $\varepsilon > 0$, $N(B_X, \varepsilon B_X) \leq (5/\varepsilon)^n$. Thus, for every $0 < \varepsilon < \delta$,

$$\begin{aligned}\log N(B_2^n, \varepsilon B_{\bar{E}}) &\leq \log N(B_2^n, \delta B_{\bar{E}}) + \log N(\delta B_{\bar{E}}, \varepsilon B_{\bar{E}}) \\ &\leq c_3\frac{Q^2\log n}{\delta^2} + n\log\left(\frac{\delta}{\varepsilon}\right).\end{aligned}$$

If we take $\delta^2 = c_3Q^2\frac{\log n}{n}$ it follows that for $\varepsilon \leq c_4Q\sqrt{\log n/n} = \varepsilon_0$,

$$\log N(B_2^n, \varepsilon B_{\bar{E}}) \leq n\log(\varepsilon_0/\varepsilon).$$

Now, by Theorem 4.8, for every $X_1, ..., X_n$,

$$\begin{aligned}\gamma_2^2(\mathcal{E}_{x,r}, d_{\infty,n}) &\leq c_5\int_0^\infty \varepsilon\log N(TB_2, \varepsilon B_E)d\varepsilon = c_5\int_0^\infty \varepsilon\log N(B_2^n, \varepsilon B_{\bar{E}})d\varepsilon \\ &\leq c_6\int_0^{\varepsilon_0} n\varepsilon\log\left(\frac{\varepsilon_0}{\varepsilon}\right)d\varepsilon + c_6\int_{\varepsilon_0}^{D_2}\frac{Q^2\log n}{\varepsilon}d\varepsilon.\end{aligned}$$

Using the change of variables $\eta = \varepsilon/\varepsilon_0$, the first integral is bounded by $c_6n\varepsilon_0^2\int_0^1 \eta\log(\eta^{-1})d\eta = c_7Q^2\log n$. Noting that $\varepsilon_0 = c_8D_2n^{-1/2}$, the second integral is just

$$c_7Q^2\log n(\log D_2 - \log\varepsilon_0) = c_7Q^2\log n\left(\frac{1}{2}\log n - \log c_8\right) = c_9Q^2\log^2 n.$$

■

Now we will bound $\phi_n(x)$ using a parameter that describes the decay of the eigenvalues (λ_i). By Assumption 3.1, the sequence of eigenvalues has a bounded weak ℓ_p norm for some $0 < p < 1$, implying that for all $x > 0$,

$$|\{\lambda_i \geq x\}| \leq \|(\lambda_i)\|_{p,\infty} x^{-p}. \tag{4.3}$$

Set $\tilde{Q}^2(x,r) = c_p A^2 x^{1-p} r^{2p} \|(\lambda_i)\|_{p,\infty}$ and define the function $\tilde{U}_n(x,r)$ by

$$\tilde{U}_n(x,r) = c_p' \tilde{Q}(x,r) \log n,$$

where c_p' is an appropriate constant that depends only on p. Then, by Lemma 3.4, $U_n(\mathcal{E}_{x,r}) \leq \tilde{U}_n(x,r)$ and setting

$$\tilde{\phi}_n(x,r) = \frac{\tilde{U}_n(x,r)}{\sqrt{n}} \cdot \max\left(\sqrt{x}, \sqrt{\mathbb{E}\mathcal{L}_{\beta^*}}, \frac{\tilde{U}_n(x,r)}{\sqrt{n}}\right),$$

it follows that for $T = rB_2$, $\phi_n(x) \leq \tilde{\phi}_n(x,r)$.

Lemma 4.10 *Suppose that K satisfies Assumption 3.1 and Assumption 4.1. Then there exists a constant c_p depending only on p for which the following holds. Let $T_r = rB_2$ and set V_r to be the star-shaped hull of $\{\mathcal{L}_f : f \in T_r\}$. If $V_{r,\lambda} = \{\mathcal{L}_f \in V_r : \mathbb{E}\mathcal{L}_f \leq \lambda\}$ then*

$$\mathbb{E}\|P_n - P\|_{V_{r,\lambda}} \leq c_p \tilde{\phi}_n(\lambda, r).$$

Proof. In view of Theorem 4.5, it is enough to show that the sum

$$\sum_{i=0}^{\infty} 2^{-i} \tilde{\phi}_n(2^{i+1}\lambda, r)$$

is dominated by a multiple of the first term in the sum.

For any $\alpha \geq 1$ and any $x > 0$, it is evident from the definition of $\tilde{U}_n$ that

$$\tilde{U}_n(\alpha x, r) \leq \alpha^{1/2-p/2} \tilde{U}_n(x,r);$$

therefore one can verify that $\tilde{\phi}_n(\alpha x, r) \leq \alpha^{1-p/2} \tilde{\phi}_n(x,r)$. In particular,

$$\sum_{i=0}^{\infty} 2^{-i} \tilde{\phi}_n(2^{i+1}\lambda, r) \leq 2^{1-p/2} \sum_{i=0}^{\infty} 2^{-ip/2} \tilde{\phi}_n(\lambda, r) \leq c_p \tilde{\phi}_n(\lambda, r).$$

■

Let us pause and explain why this analysis indeed yields a far better result than the L_∞ approach. We will show later that the dominant factor in $\mathbb{E}\|P_n - P\|_{V_{r,\lambda}}$ is $\tilde{U}_n/\sqrt{n}$, which is, up to a logarithmic term and appropriate constants,

$$A\left(\frac{1}{n}\sum_{i=1}^{\infty}\min\{x, r^2\lambda_i\}\right)^{1/2} = (*).$$

In comparison, the L_∞ approach leads to a bound of the order of

$$r\left(\frac{1}{n}\sum_{i=1}^{\infty}\min\{x, r^2\lambda_i\}\right)^{1/2} = (**)$$

on $\mathbb{E}\|P_n - P\|_{V_{r,\lambda}}$ — which is considerably larger as r grows to infinity.

If x is a "fixed point" of $(**)$ (as required in the "isomorphic" result on Theorem 2.2) then

$$\left(\frac{1}{n}\sum_{i=1}^{\infty}\min\left\{\frac{x}{r^2}, \lambda_i\right\}\right)^{1/2} = c\frac{x}{r^2},$$

and thus x scales quadratically in r. On the other hand, the fixed point of $(*)$ satisfies

$$rA\left(\frac{1}{n}\sum_{i=1}^{\infty}\min\left\{\frac{x}{r^2}, \lambda_i\right\}\right)^{1/2} = cx.$$

Hence, if (λ_i) decays quickly, the fixed point will scale like a smaller power of r — in the worst case, linearly in r.

The estimate on the fixed point in the alternative approach we presented in this section is the following:

Theorem 4.11 *There exists a constant $c_{p,Y}$ depending only on p and $\|Y\|_{L_2}$ such that the following holds. If Assumptions 3.1 and 4.1 are satisfied then for every $r > 1$, if*

$$\Theta = \frac{A\|(\lambda_i)\|_{p,\infty}^{1/2} r^p \log n}{\sqrt{n}}$$

and

$$\lambda \geq c_{p,Y}\max\{\Theta^{2/(1+p)}, \Theta^{2/p}\},$$

then one has

$$\mathbb{E}\|P_n - P\|_{V_{\lambda,r}} \leq \lambda/8.$$

Proof. Fix $r > 1$. From the definition of $\tilde{\phi}_n$ it suffices to find x for which $\tilde{U}_n(x,r)/\sqrt{n} \leq c_Y \min\{x, \sqrt{x}\}$, where $c_Y \leq c_1 \min\{1, (\mathbb{E}\mathcal{L}_{\beta^*})^{-1/2}\}$, for a suitable absolute constant c_1. Note that since $\beta = 0$ is a potential minimizer, $c \leq c_1(1 + (\mathbb{E}Y^2)^{1/2})$.

The definition of Θ ensures that $\tilde{U}_n(x,r)/\sqrt{n} = c'_p x^{1/2-p/2}\Theta$. To have $\tilde{U}_n(x,r)/\sqrt{n} \leq cx$, therefore, it is enough to have $x \geq (c_{p,Y}\Theta)^{2/(1+p)}$. Similarly, to have $\tilde{U}_n(x,r)/\sqrt{n} \leq cx^{1/2}$ it is enough that $cx \geq (c_{p,Y}\Theta)^{2/p}$. ■

Corollary 4.12 *There exist a constant $c_{p,Y}$ depending only on p and $\|Y\|_\infty$ such that the following holds. Suppose that Assumptions 3.1 and 4.1 hold. Let*

$$\Theta = \frac{A\|(\lambda_i)\|_{p,\infty}^{1/2} r^p \log n}{\sqrt{n}}$$

and

$$\rho_n(r,u) = c_{p,Y}(1+u)\max\left\{\Theta^{2/(1+p)}, \frac{r^2}{n}\right\}.$$

Then the function ρ_n is a legal function in the sense of Theorem 2.5.

In particular, for every $u > 0$, with probability at least $1 - \exp(-u)$, any function $\hat{f} \in F$ that minimizes the functional

$$P_n \ell_f + \kappa_1 \tilde{\rho}_n(r(f), u)$$

also satisfies

$$P\ell_{\hat{f}} \leq \inf_{r \geq 1}(\mathcal{A}(r) + \kappa_2 \tilde{\rho}_n(r,u))$$

where

$$\tilde{\rho}_n(r,u) = \rho_n\left(2r, u + \ln\frac{\pi^2}{6} + 2\ln(1 + c'_Y n + \log r)\right).$$

Let us examine the sample error term in order to compare it with our previous result and with the result of [8]. For a fixed u and r, the dependency on n is similar to our previous result; the worst term is $\sim (\log^2 n/n)^{1/(1+p)}$. The dependency on r is more interesting: there is one term that grows like r^2, while other grows polynomially in r with an exponent between zero and one.

The feature of this new bound that makes it better than our previous one is the fact that the term with the worst asymptotic behavior in n has the best asymptotic behavior in r. Indeed, the r^2 term in $\rho_n(r,u)$ has a dependence in n that scales like $1/n$, a much better rate than in the previous section. The significance of this is the suggestion that a regularization term of $\|f\|_H^2$

will result in over-regularization when n is large. In fact, a similar study to the one that leads to the estimate in (3.3) shows that Corollary 4.12 is indeed far better. In the following section we will show that one can improve Corollary 4.12 even further by completely removing the r^2 term.

5 Removing the r^2 term

The function ρ_n from Corollary 4.12 is almost the function we would have liked to have. Its leading term is $\Theta^{2/(1+p)} \sim (r^{2p}n^{-1}\log^2 n)^{1/(1+p)}$ while the other term scales like r^2/n and is dominant only for very large values of r. Here, we will show that the latter does not influence the minimization problem we are interested in and can be removed. Since some of the technical details of the proof of that observation are rather tedious and have already been presented in previous sections, certain parts of the argument will only be outlined.

Let us return to Theorem 2.2. The isomorphic condition we have established there holds in the set $F = rB_H$ with the functional

$$\psi(f,u) = c_{p,Y}\left(\max\left\{\Theta^{2/(1+p)}, \Theta^{2/p}\right\} + c_Y(1+u)\frac{\|f\|_\infty^2}{n}\right).$$

That is, for every $u > 0$, with probability at least $1-\exp(-u)$, for every $f \in F$,

$$\frac{1}{2}P_n\mathcal{L}_f - \psi(f,u) \le P\mathcal{L}_f \le 2P_n\mathcal{L}_f + \psi(f,u).$$

Consider the minimization problem one faces when performing regularized learning. The problem is always to minimize a functional $\hat{\Lambda} = P_n\ell_f + \kappa_1 V_n$, hoping that the minimizer $\hat{f}$ will satisfy that

$$P\ell_{\hat{f}} \le \inf_f \Lambda(f) = \inf_f \left(P\ell_f + \kappa_2 V_n\right),$$

where the functional $V_n : H \times \mathbb{R}_+ \to \mathbb{R}_+$ is nonnegative. In addition, all of the functionals we are interested in have the property that, for a fixed $f \in H$ and $u \in \mathbb{R}_+$, $V_n(f,u)$ tends to zero as $n \to \infty$.

We will specify our choice for the functional V_n later, but as a starting point, observe that since $f = 0$ is a potential minimizer, then (assuming that $\|Y\|_\infty \le 1$), any minimizer of $\hat{\Lambda}$ will satisfy that $\hat{\Lambda}(\hat{f}) \le \hat{\Lambda}(0) \le 1 + V_n(0)$, and the same will hold for Λ. Since $V_n(0)$ tends to zero as n grows, we can take n sufficiently large (depending on $\|Y\|_\infty$) to ensure that $v_n(0) \le 1$. Therefore, for these values of n any minimizer $\hat{f}$ of $\hat{\Lambda}$ satisfies

$$\hat{\Lambda}(\hat{f}) \le 2$$

and any minimizer f^* of Λ satisfies

$$\Lambda(f^*) \leq 2.$$

Thus,

$$\{f : f \text{ minimizes } \Lambda\} \subset \{f : \mathbb{E}(f - Y)^2 \leq 2\} \subset \{f : \mathbb{E}f^2 \leq 9\},$$

and

$$\{f : f \text{ minimizes } \hat{\Lambda}\} \subset \{f : \hat{\Lambda}(f) \leq 2\} \subset \{f : P_n f^2 \leq 9\}.$$

Having this in mind, we will decompose H into two subsets. The first one, H_1, will contain $\{f : \mathbb{E}f^2 \leq 9\}$. In addition, we will show that $\bar{F}_r = H_1 \cap rB_H$ is an ordered, parameterized hierarchy of H_1 and that the assumptions of Theorem 2.5 will be satisfied with respect to a functional $V(r, x)$ for which the dominant term is $\Theta^{2/(1+p)}$.

Thus, by Theorem 2.5, with high probability, any minimizer of $\hat{\Lambda}$ in H_1 will satisfy

$$P\ell_{\hat{f}} \leq \inf_{f \in H_1} \left(P\ell_f + \kappa_2 \tilde{V}(\|f\|_H, u) \right), \tag{5.1}$$

where $\tilde{V}$ is defined in a similar way to $\tilde{\rho}_n$ in Corollary 4.12.

The next step will be to extend the result beyond H_1 to H. Indeed, since $\{f : \mathbb{E}f^2 \leq 9\} \subset H_1$ then the infimum in H of the RHS of (5.1) is actually attained in H_1. Hence, the infimum in (5.1) is really over all functions in H. To conclude this line of reasoning, we will then show that with high probability, every empirical minimizer of $\hat{\Lambda}$ is in H_1, by proving that if $f \in H \backslash H_1$ then $P_n f^2 \geq 9$.

The correct decomposition of H is attained using the following estimate on the ratio between the $\|f\|_H$ and $\|f\|_\infty$ for any function in H.

Lemma 5.1 *Suppose that Assumptions 3.1 and 4.1 are satisfied. There is a constant $\kappa_3 = \kappa_3(A, p, \|(\lambda_i)\|_{p,\infty})$ such that, for every $f \in H$*

$$\mathbb{E}f^2 \geq \kappa_3 \left(\frac{\|f\|_\infty}{\|f\|_H^p} \right)^{2/(1-p)}.$$

Proof. Recall that $\|K(x,x)\|_\infty \leq 1$ and let $r > 0$. Set $f(x) = \sum_{i=1}^\infty t_i \sqrt{\lambda_i} \varphi_i(x)$ where $\|t\|_2 = r$, and observe that since $\|K(x,x)\|_\infty \leq 1$ then $\alpha = \|f\|_\infty \leq r$. Also, since $\|(\lambda_i)\|_{p,\infty} < \infty$ and $(\lambda_i)_{i=1}^\infty$ is nonnegative and non-increasing then for every i, $\lambda_i \leq (\|(\lambda_i)\|_{p,\infty}/i)^{1/p}$.

Fix N be named later and observe that

$$\begin{aligned}\|f\|_\infty &\leq A\left(\sum_{i=1}^N |t_i|\sqrt{\lambda_i} + r\left(\sum_{N+1}^\infty \lambda_i\right)^{1/2}\right)\\ &\leq A\left(\sum_{i=1}^N |t_i|\sqrt{\lambda_i} + r\|(\lambda_i)\|_{p,\infty}^{1/2p}\left(\frac{1}{N}\right)^{(1-p)/2p}\right)\\ &\leq A\sum_{i=1}^N |t_i|\sqrt{\lambda_i} + \frac{\|f\|_\infty}{2},\end{aligned}$$

provided that $N^{(1-p)/2p} \geq 2Ar\|(\lambda_i)\|_{p,\infty}^{1/2p}/\alpha$. Hence, $A\sum_{i=1}^N t_i\sqrt{\lambda_i} \geq \alpha/2$. Note that r/α is bounded below by $1/\|K\|_\infty$ and so we can choose an integer N such that

$$\frac{2Ar\|(\lambda_i)\|_{p,\infty}^{1/2p}}{\alpha} \leq N^{(1-p)/2p} \leq \frac{cAr\|(\lambda_i)\|_{p,\infty}^{1/2p}}{\alpha}$$

for some constant c depending on $\|K\|_\infty$, p and $\|(\lambda_i)\|_{p,\infty}$. Clearly, for any $v \in \mathbb{R}^N$, $\|v\|_{\ell_2^N} \geq \|v\|_{\ell_1^N}/\sqrt{N}$, and thus,

$$\sum_{i=1}^N t_i^2\lambda_i \geq c'\frac{\alpha^2}{N} = c_1\left(\frac{\alpha}{r^p}\right)^{2/(1-p)},$$

where c_1 is a constant depending on K, A, p and $\|(\lambda_i)\|_{p,\infty}$.

On the other hand, since $(\varphi_i)_{i=1}^\infty$ is an orthonormal family,

$$\begin{aligned}\mathbb{E}f^2 &= \mathbb{E}\sum_{i,j} t_i t_i\sqrt{\lambda_i\lambda_j}\varphi_i\varphi_j \geq \sum_{i=1}^N t_i^2\lambda_i \geq c_1\left(\frac{\alpha}{r^p}\right)^{2/(1-p)}\\ &= c_1\left(\frac{\|f\|_\infty}{\|f\|_H^p}\right)^{2/(1-p)}.\end{aligned}$$

■

Let

$$H_1 = \{0\} \cup \left\{f : \kappa_3\left(\frac{\|f\|_\infty}{\|f\|_H^p}\right)^{2/(1-p)} \leq 50\right\}.$$

Since the set of minimizers of any functional Λ we will be interested in is contained in $\{f : \mathbb{E}f^2 \leq 9\}$ then by Lemma 5.1, the set of such minimizers is contained in H_1.

The set H_1 has additional properties. There is a constant c, depending on p and κ_3, such that on H_1,

$$\|f\|_\infty \leq c\|f\|_H^p. \tag{5.2}$$

Moreover, for every $r > 0$, if one considers $\bar{F}_r = H_1 \cap rB_H$ then the minimizer of $P\ell_f$ in $F_r = rB_H$ actually belongs to $\bar{F}_r$ (again, by comparing to $f = 0$). Therefore, it is straightforward to show that $\bar{F}_r$ is an ordered, parameterized hierarchy of H_1 with $r(f) = \|f\|_H + 1$, implying that one can obtain the desired isomorphic result on H_1, with the $\|f\|_\infty^2/n$ term replaced by $\|f\|_H^{2p}/n$.

Indeed, we can combine Theorem 2.2 with (5.2) and the fact that the localized averages $\mathbb{E}\|P_n - P\|$ indexed by $\{\mathrm{star}(\mathcal{L}_{\bar{F}_r}, 0) : \mathbb{E}h \leq \lambda\}$ are smaller than the localized averages indexed by the larger set $\{\mathrm{star}(\mathcal{L}_{F_r}, 0) : \mathbb{E}h \leq \lambda\}$ to show that for every $r \geq 1$, with probability at least $1 - \exp(-u)$, for every $f \in \bar{F}_r$,

$$\frac{1}{2}P_n\mathcal{L}_{r,f} - \frac{\lambda}{2} - c(1 + r^{2p})\frac{u}{n} \leq P\mathcal{L}_{r,f} \leq 2P_n\mathcal{L}_{r,f} + \frac{\lambda}{2} + c(1 + r^{2p})\frac{u}{n},$$

where $\mathcal{L}_{r,f}$ is the excess loss associated with f relative to $\bar{F}_r$.

Using Theorem 4.11, one obtains the following:

Corollary 5.2 *There exists a constant κ_4' that depends on $p, A, \|(\lambda_i)\|_{p,\infty}$ and $\|Y\|_\infty$ for which the following holds. If $\Upsilon = r^p/\sqrt{n}$ then the function*

$$V'(r,u) = \kappa_4'(1+u)\max\{(\Upsilon \log n)^{2/(1+p)}, (\Upsilon \log n)^{2/p}, \Upsilon^2\},$$

is a legal function in the sense of Theorem 2.5 for the hierarchy $\{\bar{F}_r : r > 0\}$.

In particular, if we set $\hat{\Lambda}'(f,x) = P_n\ell_f + \kappa_1\tilde{V}'(f,u)$, then with probability at least $1 - \exp(-u)$, every f that minimizes $\hat{\Lambda}'$ in H_1 also satisfies

$$P\ell_{\hat{f}} \leq \inf_{f \in H}\left(P\ell_f + \kappa_2\tilde{V}'(r(f),u)\right)$$

where $\tilde{V}'$ is defined analogously to $\tilde{\rho}_n$ in Corollary 4.12.

Next we will show that the $(\Upsilon \log n)^{2/p}$ and Υ^2 terms are non-essential. Indeed, for sufficiently large n, the minimal value in H of $\hat{\Lambda}$ will be at most 2 (by comparing it to $f = 0$). Hence, if $f \in H$ satisfies that $\kappa_5'\kappa_1\Upsilon\log n \geq 2$ (i.e., if $\|f\|_H \geq \kappa_5(n/\log^2 n)^{1/2p}$) then it is not a potential minimizer of $\hat{\Lambda}'$ in H. Therefore, on the set of potential minimizers, $\Upsilon \log n \leq c$, where c depends on κ_1, κ_4' and p. Hence, on this set of minimizers, we can bound

$$V'(r,u) \leq \kappa_4(1+u)(\Upsilon \log n)^{2/(1+p)}.$$

Denoting the right hand side by $V(r,u)$, we can invoke Remark 2.6 to show that $V(r,u)$ is a valid functional.

Note that we can increase H by adding every function $f \in H$ for which $\|f\|_H \geq (n/\log^2 n)^{(1/2p)}$; we have already argued that such functions cannot minimize $\hat{\Lambda}$.

To conclude, if

$$H_1' = H_1 \cup \{f : \|f\|_H \geq \kappa_5 (n/\log^2 n)^{1/2p}\}$$

then with probability at least $1-\exp(-u)$, every f that minimizes

$$P_n \ell_f + \kappa_1 \tilde{V}(r(f),u),$$

in H_1' also satisfies

$$P\ell_{\hat{f}} \leq \inf_{f \in H} \left(P\ell_f + \kappa_2 \tilde{V}(r(f),u) \right).$$

Next, let us consider the set $H_2 = H \backslash H_1'$. Clearly, each function in H_2 satisfies that $\|f\|_H \leq c_1 \|f\|_\infty^{1/p}$ and $\mathbb{E}f^2 \geq 50$. We will show that with high probability, any $f \in H_2$ satisfies that $P_n f^2 \geq 9$, and thus it is not a potential minimizer to $\hat{\Lambda}$ in H.

Lemma 5.3 *There exist a constant κ_6 that depends on A, p, and $\|(\lambda_i)\|_{p,\infty}$ and an absolute constant κ_7 for which the following holds. If $0 \in F$ and $F \subset \kappa_6 (n/\log^2 n)^{1/2p} B_H$ then for every $u > 0$, with probability at least $1-\exp(-u)$, for every $f \in F$,*

$$P_n f^2 \geq \frac{1}{2}\mathbb{E}f^2 - 1 - \kappa_7 (1 + \|F\|_\infty^2) \frac{u}{n},$$

where $\|F\|_\infty = \sup_{f \in F} \|f\|_\infty$.

Proof. Apply Theorem 4.11 with $Y \equiv 0$, noting that in this case, $\mathcal{L}_f = f^2$. It follows that we can set

$$W_{x,r} = \{f^2 : \|f\|_H \leq r,\ \mathbb{E}f^2 \leq x\}$$

and $\mathbb{E}\|P_n - P\|_{W_{\lambda,r}} \leq \lambda/8$ provided that

$$\lambda \geq c_1 \max\{(\Upsilon \log n)^{2/(1+p)}, (\Upsilon \log n)^{2/p}\},$$

where c_1 depends on A, p and $\|(\lambda_i)\|_{p,\infty}$. We will apply this fact for $\lambda = 2$. That is, we need to ensure that r is chosen in a way such that

$$c_1 \max\{(\Upsilon \log n)^{2/(1+p)}, (\Upsilon \log n)^{2/p}\} \leq 2,$$

which is the case, for example, if $r \leq c_2 (n/\log^2 n)^{1/2p}$.

The result now follows from Theorem 2.2. ■

Set $r_H = \kappa_6 (n/\log^2 n)^{1/2p}$ and observe that we may assume that $H_2 \subset r_H B_H$. Indeed, this could be done by increasing κ_5 and noting that $H_2 \subset \kappa_5 (n/\log^2 n)^{1/2p} B_H$.

The final preliminary step we take is to decompose H_2 into L_∞ shells in the following way. Fix $u > 0$ and set r_0 such that $\kappa_7 u(1+r_0^2)/n < 9$. Put $(r_i)_{i=0}^m$, $r_i = 2^i r_0$, and r_m is the first that exceeds r_H. Thus, $m \leq c_1(\log n + \log u)$. Let

$$B = \left\{ f : \|f\|_\infty \geq \kappa_8 \|f\|_H \left(\frac{u}{n}\right)^{(1-p)/2p} \right\} \tag{5.3}$$

where κ_8 is some constant to be named later. We will consider the sets $F_0 = H_2 \cap r_0 B_\infty$ and

$$F_i = H_2 \cap \{f : r_i \leq \|f\|_\infty \leq r_{i+1}\} \cap B.$$

Since $\bigcup_{i=0}^m (H_2 \cap \{f : r_i \leq \|f\|_\infty \leq r_{i+1}\}) = H_2$, any $f \in H_2 \backslash \bigcup_{i=0}^m F_i$ satisfies that

$$\|f\|_\infty \leq \kappa_8 \|f\|_H \left(\frac{u}{n}\right)^{(1-p)/2p},$$

and because $\|f\|_H \leq r_H$, then

$$\|f\|_\infty \leq \kappa_6 \kappa_8 \left(\frac{n}{\log^2 n}\right)^{1/2p} \cdot \left(\frac{u}{n}\right)^{(1-p)/2p} = c_1 u^{(1-p)/2p} \frac{n^{1/2}}{\log^{1/p} n}.$$

Therefore,

$$\frac{\|H_2 \backslash \bigcup_{i=0}^m F_i\|_\infty^2}{n} \leq c_1^2 \frac{u^{(1-p)/p}}{\log^{2/p} n}.$$

Lemma 5.4 *There exist constants c_1 and c_2 depending only on A, p and $\|(\lambda_i)\|_{p,\infty}$ for which the following holds. Fix n and $0 < u < c_1 n$ and perform the above decomposition. For every $0 \leq i \leq m$, with probability at least $1 - \exp(-u)$, every $f \in F_i$ satisfies that $P_n f^2 \geq 9$. Also, if $u \leq c_2(\log n)^{2/(1-p)}$ then with probability $1-\exp(-u)$, for every $f \in H_2 \backslash \bigcup_{i=0}^m F_i$, $P_n f^2 \geq 9$.*

Proof. First, fix $1 \leq i \leq m$ and apply Lemma 5.3 to the set F_i. For every $f \in F_i$, $\|f\|_\infty \leq \|F_i\|_\infty \leq 2\|f\|_\infty$, and thus, with probability at least $1 - \exp(-u)$,

$$P_n f^2 \geq \frac{1}{2}\mathbb{E} f^2 - 1 - \kappa_7 \frac{u(1+\|F_i\|_\infty^2)}{n} \geq \frac{1}{2}\mathbb{E} f^2 - 1 - 2\kappa_7 \frac{u(1+\|f\|_\infty^2)}{n}.$$

On the other hand, for every $f \in B$,

$$\frac{1}{4}\mathbb{E}f^2 \geq \frac{\kappa_3}{4}\left(\frac{\|f\|_\infty}{\|f\|_H^p}\right)^{2/(1-p)} \geq 2\kappa_7\frac{u\|f\|_\infty^2}{n}$$

provided that $\kappa_8 \geq (8\kappa_7/\kappa_3)^{(1-p)/2p}$. Therefore, with probability at least $1-\exp(-u)$, for every $f \in F_i$,

$$P_nf^2 \geq \frac{1}{4}\mathbb{E}f^2 - 1 - \frac{2\kappa_7 u}{n} \geq 10 - \frac{2\kappa_7}{c_1} \geq 9,$$

for a suitably large choice of c_1.

Turning to F_0, since $\kappa_7\frac{u(1+\|F_0\|_\infty^2)}{n} \leq 9$ then by Lemma 5.3, with probability at least $1-\exp(-u)$, for every $f \in F_0$,

$$P_nf^2 \geq \frac{1}{2}\mathbb{E}f^2 - 1 - \kappa_7\frac{u(1+r_0^2)}{n} \geq 9.$$

Finally, since $n^{-1}\|H_2\backslash\bigcup_{i=0}^m F_i\|_\infty^2 \leq c\frac{u^{(1-p)/p}}{\log^{2/p} n}$, then for our choice of u,

$$\kappa_7 u\frac{\|H_2\backslash\bigcup_{i=0}^m F_i\|_\infty^2}{n} \leq 9,$$

from which our claim follows using the same argument as for F_0. ■

Now we can prove our main result, which is the second part of the following claim, and was formulated as Theorem A in the introduction.

Corollary 5.5 *If Assumptions 3.1 and 4.1 are satisfied then there exist constants c_1, c_2 and c_3 that depend only on A, p and $\|(\lambda_i)\|_{p,\infty}$ a constant N_0 that depends on $\|Y\|_\infty$ and on p and a constant c_Y that depends only on $\|Y\|_\infty$ for which the following holds.*

If $n \geq N_0$, $c_1\log\log n \leq u \leq c_2(\log n)^{2/(1-p)}$, then with probability at least $1-\exp(-u/2)$, for every $f \in H_2$, $P_nf^2 \geq 9$. Thus, all the minimizers in H of

$$P_n\ell_f + \kappa_1\tilde{V}(f,u) \tag{5.4}$$

belong to H_1. In particular, for such values of u, with probability at least $1-2\exp(-u/2)$, every minimizer $\hat{f}$ in H of (5.4) satisfies that

$$P\ell_{\hat{f}} \leq \inf_{f\in H} P\ell_f + \kappa_2\tilde{V}(f,u)$$

where

$$\tilde{V}(f,u) = c_3(1+u+c_Y\ln n+\ln\log(\|f\|_H+e))\left(\frac{(\|f\|_H+1)^p\log n}{\sqrt{n}}\right)^{2/(1+p)}.$$

A Proofs

The starting point in the proof of Theorem 2.5 is the following theorem by Bartlett [1].

Theorem A.1 *Suppose that $\{F_r; r \geq 1\}$ is an ordered, parameterized hierarchy and that $\rho_n(r)$ is a positive, continuous, increasing function. If, for all $r \geq 1$ and all $f \in F_r$,*

$$\frac{1}{2} P_n \mathcal{L}_{r,f} - \rho_n(r) \leq P\mathcal{L}_{r,f} \leq 2P_n\mathcal{L}_{r,f} + \rho_n(r) \tag{A.1}$$

then

$$P\ell_{\hat{f}} \leq \inf_{f \in F} (P\ell_f + c_1 \rho_n(r(f)))$$

where $\hat{f}$ is any function that minimizes the functional $P_n\ell_f + c_2\rho_n(r(f))$.

Proof of Theorem 2.5. Let $(r_i)_{i=1}^{\infty}$ be an increasing sequence (to be determined later) such that $r_1 = 1$ and $r_i \to \infty$ as $i \to \infty$. Define, for each $i \geq 1$, $u_i = u + \ln(\pi^2/6) + 2\ln i$. Then

$$\sum_{i=0}^{\infty} e^{-u_i} = e^{-u}$$

and so, by the union bound, with probability at least $1 - e^{-u}$, for every $i \geq 1$,

$$\frac{1}{2} P_n \mathcal{L}_{r_i,f} - \rho_n(r_i, u_i) \leq P\mathcal{L}_{r_i,f} \leq 2P_n\mathcal{L}_{r_i,f} + \rho_n(r_i, u_j).$$

If we only cared about a sequence of r_i, this would be enough for our result. However, we need an almost-isomorphic condition for all $r \geq 1$ and so the next step must be to find an almost-isomorphic condition for F_r when $r \in [r_{j-1}, r_j]$. In one direction, we have

$$\begin{aligned}
P\mathcal{L}_{r,f} &= P\mathcal{L}_{r_j,f} - P\mathcal{L}_{r_j,f_r^*} \\
&\leq 2P_n\mathcal{L}_{r_j,f} + \rho_n(r_j, u_j) - P\mathcal{L}_{r_j,f_r^*} \\
&= 2P_n\mathcal{L}_{r,f} + 2P_n\mathcal{L}_{r_j,f_r^*} + \rho_n(r_j, u_j) - P\mathcal{L}_{r_j,f_r^*} \\
&\leq 2P_n\mathcal{L}_{r,f} + 5\rho_n(r_j, u_j) + 3P\mathcal{L}_{r_j,f_r^*} \\
&\leq 2P_n\mathcal{L}_{r,f} + 5\rho_n(r_j, u_j) + 3P\mathcal{L}_{r_j,f_{r_{j-1}}^*}
\end{aligned} \tag{A.2}$$

while in the other direction, we get

$$\begin{aligned}
2P\mathcal{L}_{r,f} &= 2P\mathcal{L}_{r_j,f} - 2P\mathcal{L}_{r_j,f_r^*} \\
&\geq P_n\mathcal{L}_{r_j,f} - 2\rho_n(r_j,u_j) - 2P\mathcal{L}_{r_j,f_r^*} \\
&= P_n\mathcal{L}_{r,f} + P_n\mathcal{L}_{r_j,f_r^*} - 2\rho_n(r_j,u_j) - 2P\mathcal{L}_{r_j,f_r^*} \\
&\geq P_n\mathcal{L}_{r,f} - \frac{5}{2}\rho_n(r_j,u_j) - \frac{3}{2}P\mathcal{L}_{r_j,f_r^*} \\
&\geq P_n\mathcal{L}_{r,f} - \frac{5}{2}\rho_n(r_j,u_j) - \frac{3}{2}P\mathcal{L}_{r_j,f_{r_{j-1}}^*} \qquad \text{(A.3)}
\end{aligned}$$

Now we can choose our sequence r_i: recall that $r_1 = 1$ and set r_i, for all $i \geq 2$, to be the largest number satisfying both

$$\begin{aligned}
r_i &\leq 2r_{i-1} \\
P\mathcal{L}_{r_j,f_{r_{i-1}}^*} &\leq \rho_n(r_i,u_i). \qquad \text{(A.4)}
\end{aligned}$$

Note that choosing the largest number is not a problem because both $\rho_n(r,u)$ and $P\mathcal{L}_{r,f_{r_{j-1}}^*}$ are continuous functions of r; that is, the supremum of the set of r satisfying (A.4) is attained.

Our choice of r_i ensures that, for all $i \geq 1$,

$$i \leq \frac{P\ell(f_{r_1}^*,Y)}{\rho_n(r_1,u_1)} - \frac{P\ell(f_{r_i}^*,Y)}{\rho_n(r_i,u_i)} + \log(2r_i) \leq \frac{P\ell(f_{r_1}^*,Y)}{\rho_n(r_1,u_1)} + \log(2r_i). \qquad \text{(A.5)}$$

Indeed, for $i = 1$ this is trivial. For larger i we can proceed by induction: our definition of r_i ensures that either $r_i = 2r_{i-1}$ or $P\ell(f_{r_{i-1}}^*,Y) = P\ell(f_{r_i}^*,Y) + \rho_n(r_i,u_i)$. In the first case, $\log r_i = \log r_{i-1} + 1$ and the inductive step follows. In the second case, assuming that

$$i - 1 \leq \frac{P\ell(f_{r_1}^*,Y)}{\rho_n(r_1,u_1)} - \frac{P\ell(f_{r_{i-1}}^*,Y)}{\rho_n(r_{i-1},u_{i-1})} + \log r_{i-1}$$

then

$$\begin{aligned}
i &\leq \frac{P\ell(f_{r_1}^*,Y)}{\rho_n(r_1,u_1)} - \frac{P\ell(f_{r_{i-1}}^*,Y)}{\rho_n(r_{i-1},u_{i-1})} + 1 + \log(2r_i) \\
&\leq \frac{P\ell(f_{r_1}^*,Y)}{\rho_n(r_1,u_1)} - \frac{P\ell(f_{r_{i-1}}^*,Y)}{\rho_n(r_i,u_i)} + 1 + \log(2r_i) \\
&= \frac{P\ell(f_{r_1}^*,Y)}{\rho_n(r_1,u_1)} - \frac{P\ell(f_{r_i}^*,Y)}{\rho_n(r_i,u_i)} + \log(2r_i)
\end{aligned}$$

which proves (A.5) by induction. In particular, for any $i \geq 1$ and any $r \geq r_i$, $u_i \leq \theta(r,u)$. Therefore

$$\rho_n(r_i,u_i) \leq \rho_n(2r,\theta(r,u))$$

for any $r \in [r_{i-1}, r_i]$.

Note that (A.5) implies that the sequence r_i tends to infinity with i. Then by (A.2), (A.3) and (A.4), with probability at least $1 - e^{-u}$, for all $r \geq 1$ and all $f \in F_r$,

$$\frac{1}{2}P_n\mathcal{L}_{r,f} - 4\rho_n(2r,\theta(r,u)) \leq P\mathcal{L}_{r,f} \leq 2P_n\mathcal{L}_{r,f} + 8\rho_n(2r,\theta(r,u)).$$

We conclude the proof by applying Theorem A.1. ■

References

[1] P. L. Bartlett, *Fast rates for estimation error and oracle inequalities for model selection*, Econometric Theory, 24(2), 2008.

[2] P. L. Bartlett, O. Bousquet and S. Mendelson, Local Rademacher Complexities, Ann. Stat. 33(4), 1497-1537, 2005.

[3] P. L. Bartlett and S. Mendelson, *Empirical Minimization*, Probability Theory and Related Fields, 135, 311–344, 2006.

[4] P. L. Bartlett and S. Mendelson, work in progress.

[5] M. Sh. Birman and M. Z. Solomyak, *Estimates of singular numbers of integral operators*, Russian Mathematical Surveys 32, 15–89, 1977.

[6] A. Caponnetto and E. De Vito, *Optimal rates for regularized least-squares algorithm*, Foundation of Computational Mathematics, 7(3), 331-368, 2007.

[7] F. Cucker and S. Smale, *On the Mathematical Foundations of Learning*, Bulletin of the American Mathematical Society, 39, 1–49, 2002.

[8] F. Cucker and S. Smale, *Best Choices for Regularization Parameters in Learning Theory: On the Bias-Variance Problem*, Foundations of Computational Mathematics, 2, 413–428, 2002.

[9] R.M. Dudley *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics 63, Cambridge University Press 1999.

[10] FERNIQUE, X. Régularité des trajectoires des fonctiones aléatoires gaussiennes, Ecole d'Eté de Probabilités de St-Flour 1974, Lecture Notes in Mathematics 480, Springer-Verlag 1975, 1–96.

[11] E. Giné and J. Zinn, Some limit theorems for empirical processes, Ann. Probab. 12(4), 929-989, 1984.

[12] O. Guedon, S. Mendelson, A. Pajor and N. Tomczak-Jaegermann, Subspaces and orthogonal decompositions generated by bounded orthogonal systems, Positivity, 11(2), 269-283, 2007.

[13] O. Guédon, M. Rudelson, L_p moments of random vectors via majorizing measures, Adv. Math. 208(2), 798-823, 2007.

[14] H. Konig, *Eigenvalue Distribution of Compact Operators*, Springer Verlag, 1986.

[15] M. Ledoux, *The Concentration of Measure Phenomenon*, American Mathematical Society, 2001.

[16] W. S. Lee, P. L. Barlett and R. C. Williamson, *The Importance of Convexity in Learning with Squared Loss*, IEEE Transactions on Information Theory, 44(5), 1974–1980, 1996.

[17] P. Massart: About the constants in Talagrand's concentration inequality for empirical processes, Ann. Probab. 28(2) 863-884, 2000.

[18] S. Mendelson, *Estimating the performance of kernel classes*, Journal of Machine Learning Research, 4, 759–771, 2003.

[19] S. Mendelson, A. Pajor and N. Tomczak-Jaegermann, *Reconstruction and subgaussian operators in Asymptotic Geometric Analysis*, Geometric and Functional Analysis, 17(4), 1248-1282, 2007.

[20] S. Mendelson, Obtaining fast error rates in nonconvex situations, Journal of Complexity, in press.

[21] S. Mendelson, On weakly bounded empirical processes, Math. Ann. 340(2), 293-314, 2008.

[22] A. Pajor, N. Tomczak-Jaegermann, Remarques sur les nombres d'entropie d'umopérteur et de son transposé, C. R. Acad. Sci Paris 301, 743-746, 1985.

[23] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press, 1989.

[24] M. Rudelson, Random vectors in the isotropic position, J. Funct. Anal. 164, 60-72, 1999.

[25] S. Smale and D. X. Zhou, *Estimating the approximation error in learning theory*, Analysis and Applications, 1, 17–41, 2003.

[26] S. Smale and D. X. Zhou, *Learning Theory Estimates via Integral Operators and Their Approximations*, Constructive Approximation, Springer, 2007.

[27] Talagrand, M. Regularity of Gaussian processes, Acta Math. 159 (1987), 99–149.

[28] M. Talagrand, Sharper bounds for Gaussian and empirical processes, Ann. Probab. 22(1), 28-76, 1994.

[29] M. Talagrand, *The Generic Chaining*, Springer, 2005.