Journal of Machine Learning Research 4 (2003) 759-771 Submitted 2/03; Published 10/03

On the Performance of Kernel Classes

Shahar Mendelson SHAHAR.MENDELSON@ANU.EDU.AU
RSISE, The Australian National University,
Canberra, ACT 0200, Australia

Editors: Thore Graepel and Ralf Herbrich

Abstract

We present sharp bounds on the localized Rademacher averages of the unit ball in a reproducing kernel Hilbert space in terms of the eigenvalues of the integral operator associated with the kernel. We use this result to estimate the performance of the empirical minimization algorithm when the base class is the unit ball of the reproducing kernel Hilbert space.

1. Introduction

In this article we investigate the connections between the random averages associated with kernel classes and the spectrum of the integral operator $T_K : L_2(\Omega,\mu) \to L_2(\Omega,\mu)$, which is defined by

$$(T_K f)(x) = \int K(x,y) f(y) d\mu(y),$$

where (Ω,μ) is a probability space.

The kernel K is used to generate a Hilbert space, known as a reproducing kernel Hilbert space, whose unit ball is the class of functions we investigate.

Recall that if K is a positive definite function $K : \Omega \times \Omega \to \mathbb{R}$, then by Mercer's Theorem there is an orthonormal basis $(\phi_i)_{i=1}^\infty$ of $L_2(\mu)$ such that $\mu\times\mu$ almost surely, $K(x,y) = \sum_{i=1}^\infty \lambda_i \phi_i(x)\phi_i(y)$, where $(\lambda_i)_{i=1}^\infty$ is the sequence of eigenvalues of T_K (arranged in a non-increasing order) and ϕ_i is the eigenvector corresponding to λ_i.

Let H_K be the set of functions of the form $\sum_{i=1}^\infty a_i K(x_i,\cdot)$, where $x_i \in \Omega$ and $a_i \in \mathbb{R}$ satisfy that $\sum_{i,j=1}^\infty a_i a_j K(x_i,x_j) \le 1$. One can show that this so-called kernel class H_K is the unit ball in the reproducing kernel Hilbert space defined by the integral operator, and that for every $f \in H_K$, $\|f\|_\infty \le \|K\|_\infty$.

An alternative way to define the reproducing kernel Hilbert space is via the *feature map*. Indeed, if we define $\Phi : \Omega \to \ell_2$ by $\Phi(x) = \left(\sqrt{\lambda_i}\phi_i(x)\right)_{i=1}^\infty$, then

$$H_K = \left\{ f(\cdot) = \langle \beta, \Phi(\cdot)\rangle_{\ell_2} \,\middle|\, \|\beta\|_{\ell_2} \le 1 \right\}.$$

In other words, the feature map is a way of embedding the space Ω in ℓ_2 and H_K can be represented by the unit ball in ℓ_2, where each $\beta \in \ell_2$ acts as a functional on the image of the Ω via the feature map. Moreover,

$$\|\Phi(x)\|_{\ell_2}^2 = \sum_{i=1}^\infty \lambda_i \phi_i^2(x) = K(x,x),$$

and thus, if $\|K\|_\infty < \infty$ then $\{\Phi(x) : x \in \Omega\}$ is a bounded subset of ℓ_2. We refer the reader to Cucker and Smale (2002) for more details on reproducing kernel Hilbert spaces and their connections to Learning Theory.

One of the main goals in Statistical Learning Theory is to establish sharp bounds (which hold with high probability) on the expectation of the excess loss of the function produced by a learning algorithm, based on the random sample. Here, we focus on one particular algorithm - empirical minimization. For the sake of simplicity, the learning model we investigate is the noise-free one, in which the target function one wishes to learn is deterministic, though the same result can be derived in the noisy case, and with an identical proof. In the noise-free scenario, the learner attempts to construct an "almost optimal" approximation to an unknown target function T in a given base class H using the empirical data $(X_i, T(X_i))_{i=1}^n$, where $(X_i)_{i=1}^n$ are independent data points sampled according to a fixed but unknown probability measure μ on Ω. The way one measures the "almost optimality" is via the loss functional. Here, we focus on the squared loss; Recall that the squared loss class associated with a target T and the base class H is the set of all functions of the form $\ell_h = (h-T)^2 - (P_H T - T)^2$, where $P_H T$ is the nearest point to T in H with respect to the $L_2(\mu)$ norm (that is, the best approximation of T in the class H with respect to the L_2 structure endowed by the underlying measure μ).

Given a sample $(X_1, ..., X_n)$, $(T(X_1), ..., T(X_n))$, the empirical minimization algorithm produces a function $\hat{f} = \ell_h$, which is a loss function that minimizes $\sum_{i=1}^n \ell_h(X_i)$. Note that in order to find the empirical minimizer, it suffices to minimize $\sum_{i=1}^n (h-T)^2(X_i)$, which is possible because the set values of $(T(X_i))_{i=1}^n$ are known to the learner.

One general method of obtaining bounds on $\mathbb{E}_\mu \hat{f} = \int \hat{f}(x) d\mu(x)$ is based on the fact that the random empirical structure on the loss class is comparable with the actual structure endowed by μ. For example, the bounds based on the uniform law of large numbers imply that for every $f \in F$, $|\mathbb{E}_\mu f - n^{-1}\sum_{i=1}^n f(X_i)|$ is small. This additive notion of similarity of the two structures is too restrictive, both because one has to control the difference uniformly over the entire class and because of the additive nature of the estimate. It is possible to obtain better bounds, based on a multiplicative notion of similarity (Bartlett and Mendelson, 2003) which uses the so-called *localized averages*. The localized averages is a function that measures the richness of a class of functions with respect to a given probability measure. Roughly speaking, the localized average at scale r is the expectation of the supremum of the empirical process $|\mathbb{E}_\mu f - n^{-1}\sum_{i=1}^n f(X_i)|$, indexed by the functions in the class with expectation smaller than r. This parameter can be used to "filter out" functions which have a large expectation (and thus are of little significance from the learner's point of view, because the empirical minimization algorithm is unlikely to select them) and to identify the scale at which the function class becomes "intrinsically rich", that is, the set of functions whose expectation is smaller than that scale is too rich to enable a useful comparison between the random empirical structure endowed by the empirical means and the one endowed by μ. To simplify notation, denote

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

Our main result is motivated by the fact that under mild assumptions on the class, one can estimate the error of the empirical minimizer as a function of $\mathbb{E}\|\mu_n - \mu\|_{F_r}$, where $F_r = \{f \in F : \mathbb{E}f = r\}$. To that end, recall the following definition:

Definition 1.1 *A class F is called a Bernstein class of type 1 with respect to the measure μ if there is some constant B such that for any $f \in F$, $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)$. A class is called star-shaped around 0 if for every $f \in F$ and $0 \leq t \leq 1$, $tf \in F$.*

It is straightforward to see that if F is a Bernstein class then its star-shaped hull with 0, defined by $\{tf : 0 \leq t \leq 1,\ f \in F\}$ is also a Bernstein class and with the same type and constant.

The need for the star-shape assumption is to ensure a certain regularity of the class. The idea is that if the class is star-shaped, its "richness" increases as the scale shrinks. Indeed, any function encountered at radius r will have a scaled version at any scale $r' < r$. Hence, one can think of the class as composed of shells which get filled as r decreases, and at a critical value of r the shell becomes too large to enable a a useful comparison between the two structures - the empirical and the real.

The following theorem exhibits the connection between $\|\mu_n - \mu\|_{F_r}$ and the expected loss of the empirical minimizer.

Theorem 1.2 *(Bartlett and Mendelson, 2003) There exists an absolute constant C for which the following holds. Let F be a class of functions bounded by b which is star-shaped around 0 and has Bernstein type 1 with a constant B. Given $0 < \varepsilon < 1$ and $r > 0$, if $\mathbb{E}\|\mu_n - \mu\|_{F_r} \leq (1-\alpha)r\varepsilon$, then with probability larger than*

$$1 - \exp\left(-C\alpha^2\varepsilon^2 n \min\left\{\frac{r}{B}, \frac{r}{b}\right\}\right),$$

the empirical minimizer satisfies that

$$\mathbb{E}_\mu \hat{f} \leq \max\left\{\frac{\mathbb{E}_\sigma \hat{f}}{1-\varepsilon}, r\right\},$$

where $\mathbb{E}_\sigma h = \frac{1}{n}\sum_{i=1}^n h(X_i)$ and $F_r = \{f \in F : \mathbb{E}f = r\}$.

Hence, the critical scale at which the class becomes "too rich" to handle via this line of argumentation is when $\mathbb{E}\|\mu_n - \mu\|_{F_r} \sim r$. Let us mention that an estimate on the function $\mathbb{E}\|\mu_n - \mu\|_{F_r}$ can sometimes lead to a better error bound via a direct analysis of the empirical minimization process (Bartlett and Mendelson, 2003), which makes the problem of estimating the localized averages even more important.

Unfortunately, obtaining bounds on the localized averages is not an easy task in general. The main result in this article is a sharp estimate on the localized averages of the squared loss class associated with a kernel base class, given as a function of the eigenvalues of the integral operator T_K.

Theorem 1.3 *There is an absolute constant C for which the following holds. Let K be a kernel such that $\|K\|_\infty \leq 1$ and let $(\lambda_i)_{i=1}^\infty$ be the spectrum of T_K (arranged in a non-increasing order). Set H_K to be the kernel class, let $T : \Omega \to [0,1]$ and put F to be the squared loss class. Then, for every $r \geq 1/n$,*

$$\mathbb{E}\|\mu_n - \mu\|_{V_r} \leq \frac{C}{\sqrt{n}}\psi(r),$$

where $\psi(r) = \left(\sum_{i=1}^\infty \min\{r, \lambda_i\}\right)^{1/2}$, $V = \{tf : 0 \leq t \leq 1,\ f \in F\}$ and $V_r = \{f \in V : \mathbb{E}f = r\}$.

For example, if $\lambda_i = e^{-i}$ then it is easy to verify that for every $0 < r < 1$,

$$\sum_{i=1}^{\infty} \min\{r, \lambda_i\} \leq cr \log(2/r),$$

where c is a suitable absolute constant. Thus, $\|\mu_n - \mu\|_{V_r} \leq r/4$ for $r \geq \frac{C \log n}{n}$, and by Theorem 1.2, with probability at least $1 - \frac{1}{n^c}$,

$$\mathbb{E}_\mu \hat{f} \leq \max\left\{\mathbb{E}_n \hat{f}, \frac{C \log n}{n}\right\} \leq \frac{C \log n}{n},$$

where the last inequality holds because the empirical expectation of the empirical minimizer is non-positive.

It is interesting to compare the resulting error bound to the estimates established by Zhang (2003), who showed that for a large family of convex loss functions, the error rate of the same problem we tackle is cr, where c is a constant depending on some parameters of the problem (e.g., on $\|K\|_\infty$), and thus under very mild assumptions can be considered as an absolute constant, while the factor r is determined as follows. Let $D_\lambda = \sum_{i=1}^{\infty} \lambda_i/(\lambda_i + \lambda)$, and let r be such that $D_r \geq c_1$ and $r/D_r \geq c_2 t/n$, where c_1 and c_2 are constants with similar properties to c. Then, with probability larger than $1 - 4\exp(-t)$, $\mathbb{E}_\mu \hat{f}$ is bounded by cr. It is easy to check that for every $\lambda > 0$,

$$\frac{\psi^2(\lambda)}{2\lambda} \leq D_\lambda \leq \frac{\psi^2(\lambda)}{\lambda}.$$

Therefore, the conditions on r are equivalent to

$$\psi(r) \geq k_1\sqrt{r}, \quad n^{-1/2}\psi(r) \leq r/k_2\sqrt{t},$$

and Zhang's result can be recovered by the previous theorem, with a slightly different tail estimate. The difficulty in Zhang's approach is that the proof of the error bound depends heavily on the structure of the Hilbert space, while here, the error bound follows from completely general principles, and the only place where the geometry of the class appears is in the estimate of the localized averages.

Note that the bound we obtain on the localized averages is not data-dependent. It differs from the worst case analysis which can be established via the shattering dimension, (see, e.g., Mendelson and Schechtman, 2003), because the underlying measure has a strong influence on the bound; indeed, a change of measure yields a different integral operator and thus a different spectrum. Data dependent error bounds, which involve the spectrum of the kernel matrix $(K(X_i, X_j))_{i,j=1}^n$ where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ were recently developed by Bartlett et al. (2003) (see also Lugosi and Wegkamp, 2003).

Let us mention that the learning model we investigate is not the only one used in the context of kernel classes. In the so-called regularized loss method, rather than restricting the base class to the unit ball of the reproducing kernel Hilbert space, a larger loss is assigned to functions which have a large norm in that space. This model will not be discussed in this article, but we refer the reader to Cucker and Smale (2003) for new results in that direction which are based on entropy estimates.

The article is organized as follows. Firstly we show that for a kernel class H_K,

$$\mathbb{E} \sup_{\{f \in H_K : \mathbb{E}f^2 \leq r\}} \left|\sum_{i=1}^{n} \varepsilon_i f(X_i)\right|,$$

is determined by the spectrum of the integral operator associated with the kernel (and the measure μ according to which $(X_i)_{i=1}^n$ are distributed). Next, we prove a general result which bounds the localized averages of the squared loss class in terms of those of the base class, as long as the latter is convex. In particular, if F is the squared loss class associated with a kernel class, we estimate $\mathbb{E}\sup_{\{f\in F:\mathbb{E}f\le r\}}|\sum_{i=1}^n \varepsilon_i f(X_i)|$. Finally, we use the technique of *peeling* to bound the localized averages of the star-shaped hull of the loss class, which yields the main result, as well as the promised error bound.

1.1 Technical Preliminaries

Below, we present some preliminary results on empirical, Rademacher and Gaussian processes we require in the sequel.

For $T\subset\mathbb{R}^n$, let $\{X_t : t\in T\}$ be the Gaussian process indexed by T whose covariance structure is given by the inner product in $\mathbb{R}^n$. Hence, for every $t\in T$, $X_t=\sum_{i=1}^n g_i t_i$, where $(g_i)_{i=1}^n$ are independent standard Gaussian variables.

The following comparison theorem for Gaussian processes originated in the work of Slepian, and is due to Fernique (see Pisier, 1989). We formulate the claim only for a finite indexing set, but it can be easily extended to more general indexing sets.

Lemma 1.4 *Let $\{Z_i, 1\le i\le m\}$ and $\{Y_i, 1\le i\le m\}$ be two Gaussian processes which satisfy that, for every i,j,*

$$\|Z_i-Z_j\|_2\le\|Y_i-Y_j\|_2.$$

Then

$$\mathbb{E}\sup_i Z_i\le\mathbb{E}\sup_i Y_i.$$

Let μ be a probability measure on Ω and set $(X_i)_{i=1}^n$ to be independent random variables distributed according to μ. Denote

$$\left\|\sum_{i=1}^n \varepsilon_i\delta_{X_i}\right\|_F=\sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|,$$

where $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables. Similarly, one can define $\|\sum_{i=1}^n g_i\delta_{X_i}\|_F$ where $(g_i)_{i=1}^n$ are, as above, independent standard Gaussian variables.

The following is a well known symmetrization claim showing that $\mathbb{E}\|\mu-\mu_n\|_F$ and $\mathbb{E}\|\sum_{i=1}^n\varepsilon_i\delta_{X_i}\|_F$ are essentially equivalent.

Lemma 1.5 *(van der Vaart and Wellner, 1996, Milman and Schechtman, 1986) Let μ be a probability measure and set F to be a class of functions. Then*

$$\begin{aligned}\mathbb{E}\|\mu_n-\mu\|_F &\le \frac{2}{n}\mathbb{E}\left\|\sum_{i=1}^n\varepsilon_i\delta_{X_i}\right\|_F\\ &\le 4\,\mathbb{E}\|\mu_n-\mu\|_F+2\left|\sup_{f\in F}\mathbb{E}_\mu f\right|\cdot\mathbb{E}_\varepsilon\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\right|,\end{aligned}$$

and

$$\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i\delta_{X_i}\right\|_F \le C\,\mathbb{E}\left\|\sum_{i=1}^{n}g_i\delta_{X_i}\right\|_F,$$

where C is an absolute constant.

We end this introduction with two concentration inequalities which are at the heart of the proofs we present.

The first is the well known Bernstein's inequality (Massart, 2000, van der Vaart and Wellner, 1996).

Theorem 1.6 *Let μ be a probability measure on Ω and let $X_1,\ldots,X_n$ be independent random variables distributed according to μ. Given a function $f:\Omega\to\mathbb{R}$, set $Z=\sum_{i=1}^n f(X_i)$, let $b=\|f\|_\infty$ and put $v=n\mathbb{E}_\mu f^2$. Then,*

$$Pr\left\{|Z-\mathbb{E}_\mu Z|\ge x\right\}\le 2e^{-\frac{x^2}{2(v+bx/3)}}.$$

The following is a version of Talagrand's inequality, which is a "functional" version of Theorem 1.6. The version we use is from Bousquet (2002).

Theorem 1.7 *Let F be a class of functions on a probability space (Ω,μ), such that for every $f\in F$, $\|f\|_\infty\le 1$ and $\mathbb{E}_\mu f=0$. Let $X_1,\ldots,X_n$ be independent random variables distributed according to μ and set*

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

If $\sigma^2\ge n\sup_{f\in F}\mathrm{var}(f)$ and $v=n\sigma^2+2\mathbb{E}Z$, then for every $x>0$

$$Pr\left(\{Z\ge \mathbb{E}Z+x\}\right)\le \exp\left(-vh\left(\frac{x}{v}\right)\right),$$

and

$$Pr\left(\{Z\le \mathbb{E}Z-x\}\right)\le \exp\left(-vh\left(\frac{x}{v}\right)\right),$$

where $h(x)=(1+x)\log(1+x)-x$. In particular,

$$Pr\left(\{|Z-\mathbb{E}Z|\ge x\}\right)\le 2\exp\left(-\frac{x^2}{2v+\frac{3x}{2}}\right).$$

Throughout this article, all absolute constants are denoted by C or c. Their value may change from line to line, or even within the same line. We denote by C_b a constant which depends only on b.

2. The Localized Averages of a Kernel Class

In this section we investigate the connections between the localized averages of a kernel class (with respect to a fixed probability measure μ) and the eigenvalues of the integral operator $T_K: L_2(\mu)\to L_2(\mu)$ associated with the kernel and μ, which is defined by

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

Theorem 2.1 *There are absolute constants C and c for which the following holds. Let $(\lambda_i)_{i=1}^{\infty}$ be the non-increasing sequence of eigenvalues of the integral operator T_K, put H_K to be the unit ball of the reproducing kernel Hilbert space, and for $r > 0$ set*

$$\psi(r) = \left(\sum_{j=1}^{\infty} \min\{\lambda_j, r\}\right)^{\frac{1}{2}}.$$

If $\lambda_1 \geq 1/n$, then for every $r \geq 1/n$,

$$c\psi(r) \leq \frac{1}{\sqrt{n}} \mathbb{E} \sup_{\{f \in H_K : \, \mathbb{E}_\mu f^2 \leq r\}} \left| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \leq C\psi(r),$$

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

The first part of the proof will be to show that all the L_p norms of the random variable $\sup_{f \in F} |\sum_{i=1}^{n} \varepsilon_i f(X_i)|$ are equivalent. This equivalence follows from a general principle based on concentration.

Lemma 2.2 *For every $b > 0$ there exists a constant c_b for which the following holds. Let F be a class of functions bounded by b, set $\sigma_F^2 = \sup_{f \in F} \mathrm{var}(f)$ and assume that $n\sigma_F^2 \geq 1$. Then,*

$$\mathbb{E} \, \|\mu_n - \mu\|_F \geq \frac{c_b \sigma_F}{\sqrt{n}}.$$

Proof. Without loss of generality, assume that $b = 1$ and that $\sigma_F^2 = \mathrm{var}(g)$ for some $g \in F$. Let $Y = \sum_{i=1}^{n} (g(X_i) - \mathbb{E}g)$ and set $v = n\mathrm{var}(g)$. By Bernstein's inequality there exists an absolute constant K such that

$$\mathbb{E}Y^2 \chi_{\{|Y| \geq K\sqrt{v}\}} \leq \frac{v}{4},$$

where $\chi_{\{\cdot\}}$ denotes the indicator function. Indeed, since $v = n \cdot \mathrm{var}(g) \geq 1$, then for every integer k,

$$\mathbb{E}Y^2 \chi_{\{|Y| \geq k\sqrt{v}\}} = \sum_{m=k}^{\infty} \mathbb{E}Y^2 \chi_{\{\{m\sqrt{v} \leq |Y| \leq (m+1)\sqrt{v}\}} \leq 2v \sum_{m=k}^{\infty} (m+1) e^{-c'm},$$

where c' is an absolute constant. Thus, the assertion follows by taking k sufficiently large. Since $\mathbb{E}Y^2 \chi_{\{|Y| \leq \sqrt{v}/2\}} \leq v/4$ then

$$\begin{aligned} v = \mathbb{E}Y^2 &\leq \frac{v}{4} + \mathbb{E}Y^2 \chi_{\{\sqrt{v}/2 \leq |Y| \leq K\sqrt{v}\}} + \frac{v}{4} \\ &\leq \frac{v}{2} + K^2 v \cdot Pr\left(\left\{\frac{\sqrt{v}}{2} \leq |Y| \leq K\sqrt{v}\right\}\right), \end{aligned}$$

and thus

$$Pr\left(\left\{\|\mu_n - \mu\|_F \geq \frac{\sigma_F}{2\sqrt{n}}\right\}\right) \geq Pr\left(\left\{\frac{\sqrt{v}}{2} \leq |Y| \leq K\sqrt{v}\right\}\right) \geq c,$$

which implies that

$$\mathbb{E}\|\mu_n - \mu\|_F \geq \frac{c\sigma_F}{\sqrt{n}}$$

for another absolute constant c. ■

Lemma 2.3 *Let Z be a nonnegative random variable which satisfies that there is some constant c, such that for every integer m,*

$$Pr(\{|Z-\mathbb{E}Z| \geq m\mathbb{E}Z\}) \leq 2e^{-cm}.$$

Then, for every $1 < p < \infty$ there is a constant c_p which depends only on p and c, such that

$$c_p(\mathbb{E}Z^p)^{\frac{1}{p}} \leq \mathbb{E}Z \leq (\mathbb{E}Z^p)^{\frac{1}{p}}.$$

Proof. By Hölder's inequality, the L_p norm is larger than the L_1 norm, and the upper bound is evident. For the lower one, fix some $1 < p < \infty$ and set $a = \mathbb{E}Z$. Clearly,

$$\mathbb{E}Z^p = \mathbb{E}Z^p\chi_{\{Z<a\}} + \sum_{m=0}^{\infty} \mathbb{E}Z^p\chi_{\{(m+1)a \leq Z < (m+2)a\}}.$$

Since Z has an exponential tail, $Pr(\{Z \geq (m+1)a\}) \leq 2e^{-cm}$, and thus

$$\mathbb{E}Z^p \leq a^p + 2a^p \sum_{m=0}^{\infty} (m+2)^p e^{-cm},$$

proving that $c_p(\mathbb{E}Z^p)^{1/p} \leq \mathbb{E}Z$.

■

Corollary 2.4 *For every $1 < p < \infty$ there is a constant c_p depending only on p for which the following holds. Let F be a class of functions bounded by 1 and let n be such that $\sigma_F^2 \geq 1/n$, where $\sigma_F^2 = \sup_{f \in F} \mathrm{var}(f)$. Then, for every $1 \leq p < \infty$*

$$c_p \left(\mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i \delta_{X_i} \right\|_F^p \right)^{\frac{1}{p}} \leq \mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i \delta_{X_i} \right\|_F \leq \left(\mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i \delta_{X_i} \right\|_F^p \right)^{\frac{1}{p}}$$

Proof. Denote by $\mathbb{E}$ the expectation with respect to the product measure $\nu^n = (\varepsilon \otimes \mu)^n$, set $Y_i = (\varepsilon_i, X_i)$ put $u(Y_i) = \varepsilon_i f(X_i)$ and let

$$Z = \sup_{u \in U} \left| \sum_{i=1}^{n} u(Y_i) \right|.$$

Observe that for every $u \in U$, $\mathbb{E}u = 0$, and that by Theorem 1.7

$$Pr(\{|Z-\mathbb{E}Z| \geq x\}) \leq 2\exp\left(-\frac{x^2}{2v + \frac{2\mathbb{E}Z}{3}}\right),$$

for any $v \geq n\sigma_U^2 + 2\mathbb{E}Z$. Since $\sigma_U \geq \sigma_F$, then by Lemma 2.2 and Lemma 1.5,

$$1 \leq n\sigma_F^2 \leq n\sigma_U^2 \leq c(\mathbb{E}Z)^2,$$

and thus

$$Pr(\{|Z-\mathbb{E}Z| \geq m\mathbb{E}Z\}) \leq 2\exp(-c'm)$$

where c' is an absolute constant. Now the assertion follows from Lemma 2.3. ■

Proof of Theorem 2.1 Let H_K be the kernel class and set $H_r = \{f : f \in H_K,\ \mathbb{E}f^2 \leq r\}$. Note that the indexing set H_r is an intersection of a ball and an ellipsoid, making the computation of the L_2 norm of $\sup_{f \in H_r} |\sum_{i=1}^n \varepsilon_i f(X_i)|$ possible. Indeed, if $\mathcal{E}$ and $\mathcal{E}'$ are ellipsoids with the same principal directions and axes $(a_i)_{i=1}^\infty$ and $(b_i)_{i=1}^\infty$ respectively, then the ellipsoid $\mathcal{B}$ whose principal directions are the same as $\mathcal{E}$ and its axes are $(\min\{a_i, b_i\})_{i=1}^\infty$ satisfies that $\mathcal{B} \subset \mathcal{E} \cap \mathcal{E}' \subset \sqrt{2}\mathcal{B}$. Therefore, one can replace the set $\mathcal{E} \cap \mathcal{E}'$ indexing the Rademacher process by $\mathcal{B}$, losing only a multiplicative factor. In our case, denote $B(r) = \{f | \mathbb{E}_\mu f^2 \leq r\}$. It follows that $f \in H_K$ is also in $B(r)$ if and only if its representing vector β satisfies that $\sum_{i=1}^\infty \beta_i^2 \lambda_i \leq r$. Hence, as a subset of ℓ_2,

$$H_K \cap B(r) = \left\{ \beta \,\middle|\, \sum_{i=1}^\infty \beta_i^2 \leq 1,\ \sum_{i=1}^\infty \beta_i^2 \lambda_i \leq r \right\},$$

implying that if $\mathcal{B} \subset \ell_2$ is defined as $\{\beta | \sum_{i=1}^\infty \mu_i \beta_i^2 \leq 1\}$, where $\mu_i = (\min\{1, r/\lambda_i\})^{-1}$, then

$$\mathcal{B} \subset H_r \subset \sqrt{2}\mathcal{B}.$$

Next, one can compute the L_2 norm of the supremum of the process indexed by $\mathcal{B}$. Indeed,

$$\begin{aligned}
\mathbb{E} \sup_{\beta \in \mathcal{B}} \left| \left\langle \beta, \sum_{j=1}^n \varepsilon_j \Phi(X_j) \right\rangle \right|^2 &= \mathbb{E} \sup_{\beta \in \mathcal{B}} \left| \left\langle \sum_{i=1}^\infty \sqrt{\mu_i} \beta_i e_i, \sum_{i=1}^\infty \sqrt{\frac{\lambda_i}{\mu_i}} \left(\sum_{j=1}^n \varepsilon_j \phi_i(X_j) \right) e_i \right\rangle \right|^2 \\
= \mathbb{E} \sum_{i=1}^\infty \frac{\lambda_i}{\mu_i} \left(\sum_{j=1}^\infty \varepsilon_j \phi_i(X_j) \right)^2 &= \mathbb{E}_\mu \sum_{i,j} \frac{\lambda_i}{\mu_i} \phi_i^2(X_j) = n \sum_{i=1}^\infty \frac{\lambda_i}{\mu_i}.
\end{aligned}$$

Finally, to show the the L_1 and the L_2 norms of $\sup_{f \in H_r} |\sum_{i=1}^n \varepsilon_i f(X_i)|$ are equivalent, it suffices to prove that

$$\sup_{f \in H_r} \mathrm{var}(\varepsilon \cdot \mathrm{f}(\mathrm{X})) = \sup_{f \in H_r} \mathbb{E} f^2 \geq \frac{1}{n},$$

where ε is a Rademacher random variable and X is distributed according to μ. To that end, let ϕ_1 be the eigenfunction of T_K associated with the largest eigenvalue λ_1, and set $g = \sqrt{T_K}\phi_1 = \lambda_1 \phi_1$. It is evident that $g \in H_K$ and that $\mathbb{E}_\mu g^2 = \lambda_1 \geq 1/n$. If $h = tg$ for an appropriate selection of $0 < t \leq 1$, then $h \in H_K$ as a convex combination of g and 0, and $\mathbb{E}_\mu h^2 \leq r$, implying that $\sup_{f \in H_r} \mathbb{E} f^2 \geq 1/n$. Thus, all the L_p norms of $\| \sum_{i=1}^n \varepsilon_i \delta_{X_i} \|_{\mathcal{B}}$ are equivalent, which completes the proof. ∎

Remark 2.5 *Note that the assumption that $\lambda_1 \geq 1/n$ is needed only for the lower bound. The upper estimate holds without that assumption.*

3. The Localized Averages of Loss Classes

Here, we establish bounds on the localized averages of the loss class associated with a kernel class using the main result of the previous section. In fact, the estimate we present is completely general, and is not restricted to kernel classes, but for arbitrary p-loss classes associated with a convex base class for $1 < p < \infty$. For the sake of simplicity, the results are formulated and proved for the squared loss case. Let us mention that all the assertions in this section hold for the agnostic (noisy) learning scenario for the squared loss, and with the same proofs.

Formally, let H be a convex class of functions bounded by b and set $T : \Omega \to [0,b]$ to be the target function. For every $h \in H$, recall that the squared loss function associated with h is $\ell_h = (h-T)^2 - (P_H T - T)^2$, where $P_H T$ is the metric projection on T onto H (that is, the nearest point to T in H with respect to the $L_2(\mu)$ norm), and that the loss class is $F = \{\ell_h : h \in H\}$.

The first lemma we present is standard and its proof is omitted.

Lemma 3.1 *Let F be the squared loss class associated with a target T and a convex class H, and set $b = \max\{\sup_{h\in H}\|h\|_\infty, \|T\|_\infty\}$. For every $\sigma \in \Omega^n$, $\inf_{h\in H}\mathbb{E}_\sigma \ell_h \leq 0$ and for every $h,h' \in H$ and any $x \in \Omega$,*

$$\left(\ell_h - \ell_{h'}\right)^2(x) \leq 16b^2(h-h')^2(x).$$

The second preliminary result we require was proved by Mendelson (2002) for a more general class of loss functionals, based on the notion of uniform convexity.

Lemma 3.2 *Let H, T and b be as in Lemma 3.1. Then there are constants C_b, C'_b and C''_b(which depend only on b) such that for every $h \in H$,*

$$\mathbb{E}\ell_h^2 \leq C_b\|h - P_H T\|_{L_2(\mu)}^2 \leq C'_b \mathbb{E}\ell_h.$$

In particular, for every $r > 0$,

$$\{h : \mathbb{E}\ell_h \leq r\} \subset \left\{h : \|h - P_H T\|_{L_2(\mu)}^2 \leq C''_b r\right\}.$$

Now, one has to show that the expectation of supremum of the Gaussian process indexed by $\{\ell_h : \mathbb{E}\ell_h \leq r\}$ can be controlled using the localized Gaussian averages associated with H.

Theorem 3.3 *Let F, H, T and b as in Lemma 3.1. Then, there are constants C_b and C'_b which depend only on b, such that for every $r > 0$,*

$$\mathbb{E}\sup_{\{h\in H:\mathbb{E}\ell_h\leq r\}}\left|\sum_{i=1}^n g_i \ell_h(X_i)\right| \leq C_b \mathbb{E}\sup_{\left\{h\in 2H:\mathbb{E}h^2\leq C'_b r\right\}}\left|\sum_{i=1}^n g_i h(X_i)\right|.$$

Proof. For every fixed $\sigma = (x_1,\ldots,x_n)$ we will apply Lemma 1.4 and compare the process indexed by

$$\{(\ell_h(x_1),\ldots,\ell_h(x_n))\}$$

where h ranges over $V = \{h \in H : \mathbb{E}\ell_h \leq r\}$, and the process indexed by

$$V' = \left\{(h(x_1),\ldots,h(x_n)) : h \in 2H, \mathbb{E}h^2 \leq Cr\right\}$$

for an appropriate absolute constant C. By Lemma 3.1, for every $h,h' \in H$,

$$\begin{aligned}\sum_{i=1}^n \left(\ell_h(x_i) - \ell_{h'}(x_i)\right)^2 &\leq C_b \sum_{i=1}^n (h-h')^2(x_i) \\ &= C_b \sum_{i=1}^n \left(\left(h - P_H T\right)(x_i) - \left(h' - P_H T\right)(x_i)\right)^2,\end{aligned}$$

and thus, by Lemma 1.4

$$\mathbb{E}\sup_{h\in V}\left|\sum_{i=1}^{n} g_i \ell_h(x_i)\right| \leq C_b \mathbb{E}\sup_{h\in V}\left|\sum_{i=1}^{n} g_i (h-P_H T)(x_i)\right|.$$

Applying Lemma 3.2, $V \subset \{h \in H : \mathbb{E}(h-P_H T)^2 \leq C_b' r\}$, and by setting $V'' = \{h - P_H T : h \in H,\ \mathbb{E}(h-P_H T)^2 \leq C_b' r\}$ it follows that

$$\mathbb{E}\sup_{h\in V}\left|\sum_{i=1}^{n} g_i (h-P_H T)(x_i)\right| \leq \mathbb{E}\sup_{u\in V''}\left|\sum_{i=1}^{n} g_i u(x_i)\right|.$$

To complete the proof, observe that H is convex and symmetric, and thus $h - P_H T \in 2H$, and $V'' \subset V'$. ■

Theorem 3.4 *Let H_K be a kernel class, put T to be a target function bounded by 1 and set F to be the squared loss class. Let μ be a probability measure and put $(\lambda_i)_{i=1}^{\infty}$ to be the sequence of eigenvalues of the integral operator associated with K and μ (arranged in a non-increasing order). Then, for every $r \geq 1/n$,*

$$\frac{1}{n}\mathbb{E}\sup_{\{f\in F:\mathbb{E}f\leq r\}}\left|\sum_{i=1}^{n} \varepsilon_i f(X_i)\right| \leq \frac{C}{\sqrt{n}}\left(\sum_{i=1}^{\infty}\min\{cr,\lambda_i\}\right)^{1/2},$$

where C and c are constants which depend only on $\|K\|_\infty$.

Proof. Using Lemma 1.5, up to an absolute multiplicative constant, the Gaussian averages upper bound the Rademacher ones. Hence, it suffices to estimate the localized Gaussian averages of the loss class, which by Theorem 3.3, are bounded by

$$C\mathbb{E}\sup_{\{h\in 2H_K:\mathbb{E}h^2\leq cr\}}\left|\sum_{i=1}^{n} g_i h(x_i)\right| = (*),$$

where C and c are constants depending only on the range of functions in the kernel class, and thus, only on $\|K\|_\infty$. Since $2H_K$ is an ellipsoid whose axes are $(2\lambda_i)_{i=1}^{\infty}$, and since the L_1 norm is upper bounded by the L_2 norm,

$$(*) \leq C\left(\mathbb{E}\sup_{\{h\in 2H_K:\mathbb{E}h^2\leq cr\}}\left|\sum_{i=1}^{n} g_i h(x_i)\right|^2\right)^{1/2},$$

which can be estimated just as in the proof of Theorem 2.1. ■

4. Estimating the Loss

Finally, we are in a position to bound the error of the empirical minimization algorithm for a base class which is a kernel class. By Lemma 3.2, the squared loss class has Bernstein type 1 with a constant depending only $\sup_{f\in F}\|f\|_\infty$. In particular, if F is the squared loss class associated with the kernel class H_K and the target T, then the star-shaped hull of F and 0, denoted by V has Bernstein type 1 with a constant depending only on $\|K\|_\infty$.

Note that

$$V_r = \{f \in V : \mathbb{E}f = r\} = \left\{ \frac{rf}{\mathbb{E}f} : f \in F,\ \mathbb{E}f \geq r \right\}, \tag{4.1}$$

and set

$$\phi(r) = \mathbb{E} \sup_{\{f \in F : \mathbb{E}f \leq r\}} \left| \sum_{i=1}^n g_i f(X_i) \right|.$$

To estimate $\|\mu_n - \mu\|_{V_r}$ we use the notion of peeling.

Lemma 4.1 *For every $t > 1$,*

$$\mathbb{E}\|\mu_n - \mu\|_{V_r} \leq \frac{C}{n} \sum_{j=1}^m \frac{\phi(t^{j+1} r)}{t^j},$$

where m is the largest integer for which $t^j r \leq \sup_{f \in F} \mathbb{E}f$ and C is an absolute constant.

Proof. Note that by (4.1),

$$\begin{aligned}
\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i \delta_{X_i} \right\|_{V_r} &= \mathbb{E} \sup_{\{f \in F : \mathbb{E}f \geq r\}} \left| \sum_{i=1}^n \varepsilon_i \frac{r}{\mathbb{E}f} f(X_i) \right| \\
&= \sum_{j=0}^m \mathbb{E} \sup_{\{f \in F : t^j r \leq \mathbb{E}f \leq t^{j+1} r\}} \left| \sum_{i=1}^n \varepsilon_i \frac{r}{\mathbb{E}f} f(X_i) \right| \\
&\leq \sum_{j=0}^m \frac{1}{t^j} \mathbb{E} \sup_{\{f \in F : t^j r \leq \mathbb{E}f \leq t^{j+1} r\}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \leq \sum_{j=0}^m \frac{1}{t^j} \phi(t^{j+1} r).
\end{aligned}$$

and the claim follows because

$$\mathbb{E}\|\mu_n - \mu\|_{V_r} \leq \frac{2}{n} \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i \delta_{X_i} \right\|_{V_r}.$$

■

Combining Lemma 4.1 with Theorem 3.4, we obtain our main result, which is the following estimate on the localized averages of the star-shaped hull of a kernel loss class.

Theorem 4.2 *There are absolute constants c and C for which the following holds. Let K be a kernel such that $\|K\|_\infty \leq 1$, and let $(\lambda_i)_{i=1}^\infty$ be the spectrum of T_K (arranged in a non-increasing order). Set H_K to be the kernel class, let $T : \Omega \to [0,1]$ and put V to be the star-shaped hull of squared loss class. Then, for every $r \geq 1/n$,*

$$\mathbb{E}\|\mu - \mu_n\|_{V_r} \leq \frac{C}{\sqrt{n}} \psi(r),$$

where $\psi(r) = \left(\sum_{i=1}^\infty \min\{r, \lambda_i\}\right)^{1/2}$ and $V_r = \{f \in V : \mathbb{E}f = r\}$.

Proof. By Lemma 4.1 and Theorem 3.4,

$$\mathbb{E}\|\mu_n - \mu\|_{V_r} \leq \frac{C}{\sqrt{n}} \sum_{i=1}^{\infty} 4^{-j} \psi(c4^{j+1} r).$$

Observe that for every $r > 0$ and $\alpha \geq 1$, $\psi(\alpha r) \leq \sqrt{2\alpha}\psi(r)$, from which our assertion easily follows. To that end, recall that $D_r = \sum_{i=1}^{\infty} \lambda_i/(r+\lambda_i)$, and that $(rD_r)^{1/2} \leq \psi(r) \leq (2rD_r)^{1/2}$. Clearly, for $\alpha \geq 1$, $\psi(\alpha r) \leq \sqrt{2\alpha}(rD_r)^{1/2} \leq \sqrt{2\alpha}\psi(r)$, as claimed.

■

References

P.L. Bartlett, O. Bousquet, and S. Mendelson. Localized Rademacher complexities. 2003. Preprint.

P.L. Bartlett and S. Mendelson. Empirical risk minimization. 2003. Preprint.

O. Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, Ecole Polytechnique, Paris, 2002.

P. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the AMS*, 39(1): 1–49, 2002.

P. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2003.

G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Annals of Statistics*, 2003. to appear.

P. Massart. About the constants in Talagrand's concentration inequality for empirical processes. *Annals of Probability*, 28(2):863–884, 2000.

S. Mendelson and G. Schechtman. The shattering dimension of sets of linear functionals. *Annals of Probability*, 2003. to appear.

S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

V.D. Milman and G. Schechtman. *Asymptotic theory of finite dimensional normed spaces*. Lecture Notes in Mathematics 1200. Springer, 1986.

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

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

T. Zhang. Effective dimension and generalization of kernel learning. In S. Thrun S. Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 454–461. MIT Press, Cambridge, MA, 2003.