Lower Bounds for the Empirical Minimization Algorithm

Shahar Mendelson

***Abstract*—In this correspondence, we present a simple argument that proves that under mild geometric assumptions on the class F and the set of target functions $\mathcal{T}$, the empirical minimization algorithm cannot yield a uniform error rate that is faster than $1/\sqrt{k}$ in the function learning setup. This result holds for various loss functionals and the target functions from $\mathcal{T}$ that cause the slow uniform error rate are clearly exhibited.**

***Index Terms*—Empirical minimization, function learning, lower bounds, statistical learning theory.**

I. Introduction

The aim of this correspondence is to present a relatively simple proof of a lower bound on the error rate of the empirical minimization algorithm in *function learning* problems.

Let us describe the question at hand. Let F be a class of functions on the probability space (Ω, μ) and consider an unknown target function T that one wishes to approximate in the following sense. The learner is given a random sample $(X_i)_{i=1}^k$, $(T(X_i))_{i=1}^k$, where $X_1, \ldots, X_k$ are independent points selected according to (the unknown) probability measure μ. The goal of the learner is to use this data to find a function $f \in F$ that approximates T with respect to some loss function ℓ; in other words, to find $f \in F$ such that the expected loss $\mathbb{E}\ell(f(X), T(X))$ is close to the best possible in the class. A typical choice of a loss function ℓ is the squared loss $|x-y|^2$, or more generally, the p-loss $|x-y|^p$ for $1 \leq p < \infty$. There are, of course, many other choices of ℓ that are used.

The function learning problem has a more general counterpart, the *agnostic learning* problem, in which the unknown target function T is replaced by a random variable Y. The data received by the learner is an independent identically distributed (i.i.d.) sample $(X_i, Y_i)_{i=1}^k$ given according to the joint probability distribution of X and Y. Again, the goal is to find some $f \in F$ for which $\mathbb{E}\ell(f(X), Y)$ is as small as possible.

An algorithm frequently used in such prediction problems is *empirical minimization*. For every sample, the algorithm produces a function $\hat{f} \in F$ that minimizes the empirical loss $\sum_{i=1}^k \ell(f(X_i), T(X_i))$ [$\sum_{i=1}^k \ell(f(X_i), Y_i)$ in the agnostic case]. The hope is to obtain a high probability estimate on the way the risk of $\hat{f}$, defined as the conditional expectation

$$\mathbb{E}\left(\ell(\hat{f}, T)|X_1, \ldots, X_k\right) - \inf_{f \in F} \mathbb{E}\ell(f, T),$$

decreases as a function of the sample size k. The expectation of this quantity is usually called the *error rate* of the problem and measures "how far" the algorithm is from choosing the best function in the class.

Manuscript received July 05, 2006; revised April 15, 2008. This work was supported in part by the Israel Science Foundation under Grant ISF 666/06 and the Australian Research Council under Discovery Grant DP0559465.

The author is with the Centre for Mathematics and its Applications, Institute of Advanced Studies, The Australian National University, Canberra, A.C.T. 0200, Australia and also with the Department of Mathematics, Technion, Israel Institute of Technology, Technion City, Haifa 32000, Israel (e-mail: shahar.mendelson@anu.edu.au).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2008.926323

If $\mathbb{E}\ell(T, \cdot)$ has a unique minimizer in F, which we denote by f^*, one can define the excess loss $\mathcal{L}(f) = \ell(f, T) - \ell(f^*, T)$. Then, the excess loss (risk) of the empirical minimizer is

$$\mathbb{E}\left(\mathcal{L}(\hat{f})|X_1, \ldots, X_k\right).$$

There are numerous results on the performance of the empirical minimization algorithm but we will not present any sort of survey of those results here. Roughly speaking, it turns out that the "richness" of the function class F as captured by the empirical process indexed by it determines how well the empirical minimization behaves in the two learning problems we mentioned above, and, in order to obtain an error rate that tends to 0 asymptotically faster than $1/\sqrt{k}$, not only does the class have to be small, but additional information on the loss is needed; for example, a Bernstein-type condition that for every $f \in F$

$$\mathbb{E}\mathcal{L}^2(f) \leq c\left(\mathbb{E}\mathcal{L}(f)\right)^\beta \tag{1.1}$$

for some constants c and $0 < \beta \leq 1$ would do.

For more details on richness parameters of classes and the way in which those, combined with conditions similar to (1.1), govern the error rate, we refer the reader to [17], [2], and [3] and to the surveys [7], [1], [21], [5], [14], [18], and [4]. For general facts concerning empirical processes, see, for example, [22] and [9].

Our main interest is in a lower bound on the performance of the empirical minimization algorithm in function learning problems. There are many lower bounds that are independent of the algorithm used (see, for example, [12] and references therein), but usually, these lower bounds deal with specific classes and are very different in nature from what we have in mind. The starting point of our discussion is the surprising result established in [15]: a lower bound on the performance of any learning algorithm in the agnostic case.

To formulate this result, we need the following definitions. We say that A is a learning algorithm if for every integer k and any sample $s = (X_i, T(X_i))_{i=1}^k$ (resp., $(X_i, Y_i)_{i=1}^k$ in the agnostic case) it assigns a function $A_s \in F$. For a random variable Y, we denote by ν the joint probability measure endowed by (X, Y).

Theorem 1.1 [15]: Let $F \subset L_2(\mu)$ be a compact class of functions bounded by 1 and set $\ell(x, y) = (x-y)^2$. Assume that there is a random variable Y bounded by β for which $\mathbb{E}(f(X) - Y)^2$ has more than a unique minimizer in F. Then, there are constants c and k_0 depending only on F, β, and μ for which the following holds. If A is a learning algorithm and $\mathcal{Y} = \{Y : \|Y\|_\infty \leq \beta\}$, then for every $k \geq k_0$,

$$\sup_Y \left(\mathbb{E}\left(\mathbb{E}\left(\ell(\hat{A}, Y)|s_k\right)\right) - \inf_{f \in F} \mathbb{E}\ell(f, Y)\right) \geq \frac{c}{\sqrt{k}}$$

where the supremum is with respect to all random variables Y taking values in $\mathcal{Y}$, $s_k = (X_i, Y_i)_{i=1}^k$ is an i.i.d sample according to the joint distribution of (X, Y), and $\hat{A} = A_{s_k}$.

In other words, there will be a range $[-\beta, \beta]$ for which no matter what learning algorithm is chosen by the learner to approximate targets taking values in that range, the best uniform error rate it can guarantee with respect to all these targets cannot be asymptotically better than $1/\sqrt{k}$.

It is important to mention that this is not the exact formulation of the result from [15]. In the original formulation, it was assumed that F is compact and nonconvex. A part of the proof was to show that under these assumptions there is some β and a random variable Y bounded by β for which $\mathbb{E}(f(X) - Y)^2$ has multiple minimizers in F, in the following sense: there are $f_1, f_2 \in F$ such that $(f_1(X) -$

$Y)^2 \neq (f_2(X) - Y)^2$ on a set of positive measure and f_1, f_2 minimize $\mathbb{E}(f(X) - Y)^2$ in F. Unfortunately, that part of the proof is incorrect (see also [16]).

The surprising point in Theorem 1.1 is that the lower bound of $1/\sqrt{k}$ does not depend on the richness of the class. The slow rate is the best possible uniform rate regardless of how "statistically small" the class F is. Somehow, the bad geometry of F is the reason for the slow rate and understanding the geometric reasons causing the bad rates is one of our goals.

A rather delicate point in Theorem 1.1 is that it does not imply that there is a single random variable Y taking values in $\mathcal{Y}$ for which the error rate is bounded from below by $c/\sqrt{k}$. What it does say is that for any fixed algorithm and any integer $k \geq k_0$ there will be some random variable Y–depending on F, μ, and k and the chosen algorithm, on which the algorithm performs poorly after being given k data points—but even for a fixed algorithm, the "bad target" Y might change with k.

Because the proof in [15] is based on a "probabilistic method" type of construction, it is indirect. For each k, the fact that a "bad" Y_k exists is exhibited (based on the existence of a target Y with multiple minimizers), but what Y_k is and how it is related to the geometry of F is not revealed by the proof. Of course, this is the best that could be expected in such a general solution since the algorithm used is not specified and can be arbitrary. The proof also uses the fact that one is allowed to select an arbitrary random variable Y as a target rather than a noiseless target function $T(X)$. Thus, the agnostic argument from [15] does not extend to the function learning setup.

We present a simple argument that proves the same lower bound in the function learning scenario for the empirical minimization algorithm. The argument requires minor assumptions on the loss functional, rather than assuming that ℓ is the squared loss. Moreover, it enables one to pin-point a "bad" target for every sample size k. Our feeling is that this proof sheds some light on the reasons why the bad geometry of F leads to poor statistical properties.

Our starting point is similar to [15] (though the proofs take very different paths). Assume that E is a reasonable normed space of functions on (Ω, μ) with a norm that is naturally connected to the loss (for example, the p-loss is connected to the L_p norm). Assume further that $F \subset E$ is "small" in an appropriate sense and that T has more than a unique best approximation in F. Fix one such best approximation $f_* \in F$. We will show that for every $k \geq k_0$ and $\lambda \sim 1/\sqrt{k}$, the function $(1-\lambda)T + \lambda f_*$ is a "bad" target function for a typical k-sample, that is, for every $k \geq k_0$

$$\mathbb{E}_{X_1,\ldots,X_k}\left(\mathbb{E}\left(\ell(\hat{f}, T_{\lambda_k})|X_1,\ldots,X_k\right) - \inf_{f\in F}\mathbb{E}\ell(f, T_{\lambda_k})\right) \geq \frac{c}{\sqrt{k}}$$

where $\hat{f}$ is the empirical minimizer and c is a constant that depends only on F, ℓ, and properties of the space E.

A corollary of this general result is Theorem 1.2. Recall that μ-Donsker classes are sets $F \subset L_2(\mu)$ that satisfy some kind of a uniform central limit theorem (see [9] and [22] for detailed surveys on this topic).

Let E be a normed space of functions on (Ω, μ) and let $N(F, E)$ be the set of functions in E that have more than a unique best approximation in F.

Theorem 1.2: Let $2 \leq p < \infty$ and set $E = L_p(\mu)$. Assume that $F \subset E$ is a μ-Donsker class of functions bounded by 1, let $R > 0$, and assume that $\mathcal{T} \subset E \cap B_{L_\infty}(0, R)$ is convex and contains F. If ℓ is the p-loss function and $\mathcal{T} \cap N(F, E) \neq \emptyset$, then for $k \geq k_0$

$$\sup_{T\in\mathcal{T}}\left(\mathbb{E}_{X_1,\ldots,X_k}\mathbb{E}\left(\ell(\hat{f}, T)|X_1,\ldots,X_k\right) - \inf_{f\in F}\mathbb{E}\ell(f, T)\right) \geq \frac{c}{\sqrt{k}}$$

where c and k_0 depend only on p, F, and R.

Let us note that the assumption that F and $\mathcal{T}$ are bounded in L_∞ is only there to ensure that Lipschitz images of these functions satisfy some technical integrability properties we require and is not essential for the proof. Also, the convexity assumption on $\mathcal{T}$ is only there to ensure that if $T \in N(F, E)$ and $f_* \in F$ then for any $\lambda \in [0, 1]$ the convex combination $(1-\lambda)T + \lambda f_* \in \mathcal{T}$, and thus, a "legal" target function.

It turns out that the reverse direction of Theorem 1.2 is also true [19]. Indeed, one can show that if the set $\mathcal{T}$ is "far away" from $N(F, E)$ then the class F satisfies a Bernstein condition. Thus, if F is "small," the uniform error rate with respect to functions in $\mathcal{T}$ decays faster than $1/\sqrt{k}$.

To formulate this reverse direction, we need the following definition.

Definition 1.3: Let E be a Banach space, set $F \subset E$ to be compact, and assume that $T \notin \overline{N(F, E)}$. If f^* is the best approximation of T in F, let

$$\lambda^*(T) = \sup\left\{\lambda \geq 1 : \lambda f^* + (1-\lambda)T \notin \overline{N(F, E)}\right\}.$$

In other words, if $T \notin \overline{N(F, E)}$ and if one considers the ray originating in f^* that passes through T, $\lambda^*(T)$ measures how far "up" this ray one can move while still remaining at a positive distance from the set $N(F, E)$. Clearly, if $d(T, N(F, E)) > 0$ then $\lambda^*(T) > 1$.

Theorem 1.4 [19]: Let $1 < p < \infty$ and set $F \subset L_p(\mu)$ to be a compact set of functions that are bounded by 1. Let T be a function bounded by 1 for which $d(T, N(F, E)) > 0$. Then, for every $f \in F$

$$\mathbb{E}\mathcal{L}_f^2 \leq B(\mathbb{E}\mathcal{L}_f)^{\alpha_p}$$

where $\mathcal{L}_f = |f - T|^p - |f^* - T|^p$ is the p-excess loss associated with f and T, $\alpha_p = \min\{p/2, 2/p\}$, and

$$B = c(p)\frac{\lambda^*(T)}{\lambda^*(T) - 1}.$$

In particular, for $p = 2$, the combination of Theorems 1.2 and 1.4 gives an almost characterization of the uniform error rate associated with a set of targets $\mathcal{T}$. Indeed, if a target is "far away" from the set $N(F, E)$, the excess loss class satisfies a Bernstein condition, implying that the error rate of the empirical minimizer depends only on the statistical complexity of F and not on its geometry. In particular, if $d(\mathcal{T}, N(F, E)) > 0$ and F is small enough, one has very fast error rates—uniformly in $\mathcal{T}$. On the other hand, $\mathcal{T}$ is closed and convex and $d(\mathcal{T}, N(F, E)) = 0$, and there is some $T \in \mathcal{T} \cap N(F, E)$. Hence, by Theorem 1.2, the best possible error rate is $\sim 1/\sqrt{k}$.

Finally, we have a word about notation. Throughout this correspondence, all constants will be denoted by c, c_1, etc. Their vales may change from line to line. We will also emphasize when a constant is absolute (that is, a fixed positive number) and when it depends on other parameters of the problem (for example, the diameter of the set F with respect to the norm). Constants that will remain fixed throughout this correspondence will be denoted by C_1, C_2, etc.

Let $P_k g = k^{-1}\sum_{i=1}^k g(X_i)$ where $X_1, \ldots, X_k$ are independent, distributed according to μ, and set $\|P_k - P\|_G$ to be the supremum of the empirical process indexed by G, that is, $\sup_{g\in G}|k^{-1}\sum_{i=1}^k g(X_i) - \mathbb{E}g|$. If E is a normed space, let $B(x, r)$ be the closed ball centered at x and of radius r and set B_E to be the closed unit ball.

II. Preliminaries

Let $(E, \|\ \|)$ be a normed space of functions on the probability space (Ω, μ). We need to assume that E has a nice structure, namely, that it is smooth and uniformly convex (defined below). For more information on these geometric notions, we refer the reader to [8], [6], and [10].

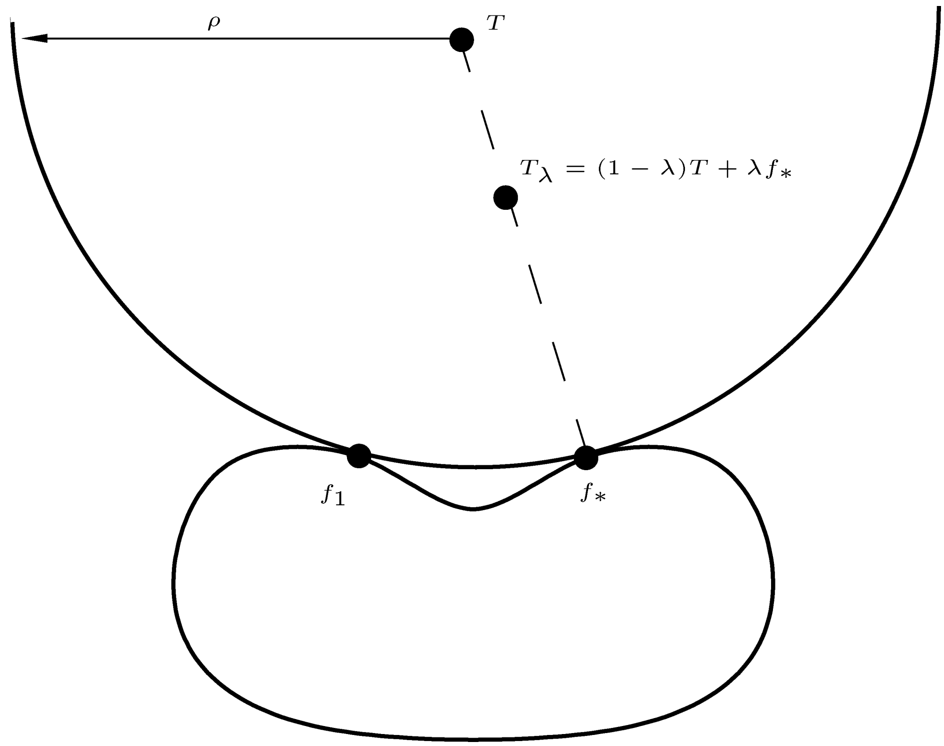

Fig. 1. Illustration of T_λ.

We say that a normed space is smooth if the norm is Gâteaux differentiable in any $x \neq 0$. There is an equivalent geometric formulation of this notion that for every x on the unit sphere of E there is a unique, norm one linear functional that supports the unit ball in x [that is, a unique functional x^*, such that $\|x^*\| = 1$ and $x^*(x) = 1$].

Uniform convexity measures how far "inside" the unit ball $(x+y)/2$ is, where x and y are distant points on the unit sphere.

Definition 2.1: E is called uniformly convex if there is a positive function $\delta_E(\varepsilon)$ satisfying that for every $0 < \varepsilon < 2$ and every $x, y \in B_E$ for which $\|x - y\| \geq \varepsilon$, $\|x + y\| \leq 2 - 2\delta(\varepsilon)$. In other words

$$\delta_E(\varepsilon) = \inf\left\{1 - \frac{1}{2}\|x+y\| : \|x\|, \|y\| \leq 1,\ \|x - y\| \geq \varepsilon\right\}.$$

The function $\delta(\varepsilon)$ is called the modulus of convexity of E.

A fact we will use later is that $\delta_E(\varepsilon)$ is an increasing function of ε and that if $0 < \varepsilon_1 \leq \varepsilon_2 \leq 2$ then $\delta_E(\varepsilon_1)/\varepsilon_1 \leq \delta_E(\varepsilon_2)/\varepsilon_2$ (see, e.g., [8, Ch. 8]).

Next, let us turn to some of the properties of the learning problem we study. Consider the sets F and $\mathcal{T}$ that are closed subsets of E. The aim of the learner is to approximate an unknown target function $T \in \mathcal{T}$ using functions from F, and the notion of approximation is via the loss functional $\ell(x, y)$. The assumptions on the loss ℓ are that it is a Lipschitz function from $\mathbb{R}^2$ to $\mathbb{R}$ and that the expected loss is compatible with the norm.

Assumption 2.1: Assume that $\ell : \mathbb{R}^2 \to \mathbb{R}$ is a nonnegative Lipschitz function with constant $\|\ell\|_{\rm lip}$. Assume further that there is some function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ that is differentiable, strictly increasing, and convex, such that for every $f, g \in E$, $\mathbb{E}\ell(f, g) = \phi(\|f - g\|)$.

This assumption is natural in the context of function learning. For example, the p loss function $\ell_p(x, y) = |x - y|^p$ satisfies Assumption 2.1 when $E = L_p(\mu)$ and $\phi(t) = t^p$.

In our construction, we consider the excess loss functional that measures how far f is from being the best in the class. To that end, set $F_{*,T} = \{f \in F : \mathbb{E}\ell(f, T) = \inf_{g \in F} \mathbb{E}\ell(g, T)\}$. It is well known (see, e.g., [15], [17], and [18]) that under various assumptions on the class and on the loss functional, $F_{*,T}$ is a singleton, which we denote by f^*. In such a case, one can define the excess loss functional

$$\mathcal{L}_T(f) = \ell(f, T) - \ell(f^*, T)$$

and the excess loss class $\mathcal{L}_T(F) = \{\mathcal{L}_T(f) : f \in F\}$. If $F_{*,T}$ contains more than one element, the excess loss class is not well defined, but we will not have to tackle this issue here.

A. An Outline of the Proof

Let $\mathcal{T}$ be a convex set containing F and put $N(F, \ell)$ to be the set of functions $T \in E$ for which $\mathbb{E}\ell(T, \cdot)$ has multiple minimizers in $\{\ell(f, T) : f \in F\}$. Assume that $\mathcal{T} \cap N(F, \ell) \neq \emptyset$ and let $T \in \mathcal{T} \cap N(F, \ell)$. Thus, there is a set $F_{*,T} \subset F$ of cardinality strictly larger than 1 with the following properties.

1) If $f_1, f_2 \in F_{*,T}$ and $f_1 \neq f_2$, then $\ell(f_1, T) \neq \ell(f_2, T)$ on a set of positive μ-probability.
2) For every $g \in F_{*,T}$, $\mathbb{E}\ell(g, T) = \min_{f \in F} \mathbb{E}\ell(f, T) \equiv R$.

Therefore, as $\mathbb{E}\ell(f, T) = \phi(\|f - T\|)$ and since ϕ is strictly increasing, then $F_{*,T} \subset F \cap B(T, \rho)$ where $\rho = \phi^{-1}(R)$. Fix any $f_* \in F_{*,T}$ and for every $\lambda \in (0, 1]$ define $T_\lambda = (1 - \lambda)T + \lambda f_*$.

Our construction has three components. First, one has to show that $T_\lambda \notin N(F, \ell)$, and that the unique minimizer of $\mathbb{E}\ell(T_\lambda, \cdot)$ is f_*. This follows from the fact that T_λ has a unique nearest point in F with respect to the norm. In particular, the excess loss functional with respect to the target T_λ (denoted by $\mathcal{L}_\lambda$) is well defined and is given by

$$\mathcal{L}_\lambda(f) = \ell(f, T_\lambda) - \ell(f_*, T_\lambda).$$

Note that T_λ is a convex combination of points in $\mathcal{T}$ and thus $T_\lambda \in \mathcal{T}$. Therefore, if the empirical minimization algorithm is to give fast rates, it must produce with high probability functions for which the conditional expectation satisfies

$$\lim_{k \to \infty} \sqrt{k}\,\mathbb{E}\left(\mathcal{L}_\lambda(\hat{f}) | X_1, \ldots, X_k\right) = 0.$$

On the other hand, take any other (fixed) $f_1 \in F_{*,T}$. It is clear that $\mathbb{E}|\ell(f_1,T) - \ell(f_*,T)| \equiv \Delta > 0$, otherwise $\ell(f,T) = \ell(f_*,T)$ almost surely, contradicting our assumption that those are distinct points in $F_{*,T}$.

Observe that $\mathrm{var}\,(\mathcal{L}_\lambda(f_1))$ tends to $\mathrm{var}\,(\ell(f_1,T) - \ell(f_*,T))$ as $\lambda \to 0$. Thus, there is a constant λ_0 such that for $\lambda \le \lambda_0$, $\mathcal{L}_\lambda(f_1)$ has a "large" variance σ^2 satisfying $\sigma \ge \Delta/2$. On the other hand, a rather simple calculation shows that the expectation of $\mathcal{L}_\lambda(f_1)$ is at most $c\lambda$.

Because typical values of $P_k\mathcal{L}_\lambda(f_1)$ are

$$\frac{1}{k}\sum_{i=1}^{k}\left(\mathcal{L}_\lambda(f_1)\right)(X_i) \sim \mathbb{E}\left(\mathcal{L}_\lambda(f_1)\right) \pm \sigma/\sqrt{k}$$

then by a quantitative version of the central limit theorem, there is an integer k_0 (that depends only on T, f_1, and ℓ) such that for every $k \ge k_0$ there is a set of samples $(X_i)_{i=1}^k$, denoted by S_k, with the following properties:

1) the measure $\mu^k(S_k) \ge 1/4$;
2) for every sample in S_k

$$P_k\mathcal{L}_\lambda(f_1) \le -c_1\frac{\Delta}{\sqrt{k}} + c_2\lambda \le -c_3\frac{\Delta}{\sqrt{k}}$$

if one takes $\lambda \le \min\{c_4(\Delta)/\sqrt{k}, \lambda_0\}$.

Hence, for such a choice of λ, if f is a potential empirical minimizer for $\mathcal{L}_\lambda$ with respect to a sample $(X_i)_{i=1}^k \in S_k$, then its empirical error must be smaller than $-c_3\Delta/\sqrt{k}$.

On the other hand, by the uniform law of large numbers, for every $r > 0$, the empirical process indexed by $\mathcal{L}_{\lambda,r}(F) = \{\mathcal{L}_\lambda(f) : \mathbb{E}\mathcal{L}_\lambda(f) \le r\}$ satisfies that with high μ^k-probability (say, at least 5/6)

$$\|P_k - P\|_{\mathcal{L}_{\lambda,r}(F)} \le c\mathbb{E}\|P_k - P\|_{\mathcal{L}_{\lambda,r}(F)} \equiv W_\lambda(r)$$

where c is an absolute constant. In other words, if $g \in F$ satisfies that $\mathcal{L}_\lambda(g) \le r$, then

$$|P_k\mathcal{L}_\lambda(g)| \le W_\lambda(r) + r.$$

It remain to show that if we select $\lambda \sim 1/\sqrt{k}$ and $r \sim 1/\sqrt{k}$, then $W_\lambda(r) + r \le c_3\Delta/(2\sqrt{k})$. Thus, if $\mathcal{L}_\lambda(f)$ has small expectation, its empirical expectation is not negative enough to defeat the empirical error generated by f_1 and the empirical minimizer $\hat{f}$ must have a relatively large risk.

We will show that one can find such choices of λ and r if the Gaussian processes indexed by random coordinate projections of F are continuous in some sense with respect to the norm on E, and this continuity assumption is satisfied, for example, if F is a μ-Donsker class and ℓ is the squared loss.

This analysis shows that for $\lambda_k \sim 1/\sqrt{k}$, the target T_{λ_k} causes the risk of the empirical minimizer to be larger than $c/\sqrt{k}$ for typical samples of size k. However, as we mentioned in the introduction, it was proved in [19] that for each $0 < \lambda < 1$, the error rate associated with the fixed target T_λ, that is, the rate at which

$$\mathbb{E}_{X_1,\ldots,X_k}\left(\mathbb{E}\left(\mathcal{L}_\lambda(\hat{f})|X_1,\ldots,X_k\right)\right)$$

tends to 0 as a function of k, will be significantly better than $1/\sqrt{k}$ if F is small enough. Thus, the slow uniform error rate is not caused by a single target function, and the guilty party truly changes with the sample size k.

III. Detailed Proof

In this section, we will present the details of the construction leading to the promised lower bound.

Consider $\mathcal{L}_\lambda(F)$ and $W_\lambda(r)$ as above and set $F_{\lambda,r} = \{f : \mathbb{E}\mathcal{L}_\lambda(f) \le r\}$. The first part of the proof will be to show that with high probability, if $\lambda \sim 1/\sqrt{k}$, then functions $\mathcal{L}_\lambda(f)$ with expectation smaller than $c/\sqrt{k}$ have an empirical error that is close to 0, and in particular, not "very negative." This, of course, requires some assumption on the richness of F, which will be captured by a variant of the notion of asymptotic equicontinuity (see, e.g., [13] and [9]).

Definition 3.1: We say that F is compatible with the norm $\|\ \|$ if the following holds. For every $u > 0$, there is some integer k_0 and $q > 0$ for which

$$\sup_{k \ge k_0}\frac{1}{\sqrt{k}}\mathbb{E}\sup_{\{f,h\in F:\|f-h\|\le q\}}\left|\sum_{i=1}^{k}g_i(f-h)(X_i)\right| \le u.$$

In other words, the oscillation of the Gaussian process indexed by a "typical" coordinate projection of F is well behaved with respect to the norm on E.

A fundamental fact due to Giné and Zinn (see, e.g., [13, Th. 14.6] and [9]) is that a μ-Donsker class is compatible with the L_2 norm, and therefore, it is compatible with any L_p norm for $2 \le p < \infty$.

Theorem 3.2: Let F be a class of functions that is compatible with the norm on E and set $F_{\lambda,r} = \{f : \mathbb{E}\mathcal{L}_\lambda(f) \le r\}$. For every $\alpha, u >$, there is an integer k_0 and $0 < \beta \le u$ such that for $k \ge k_0$, with probability $5/6$

$$\sup_{f \in F_{\alpha/\sqrt{k},\beta/\sqrt{k}}}|P_k\mathcal{L}_{\alpha/\sqrt{k}}(f)| \le \frac{2u}{\sqrt{k}}.$$

The first step in the proof of Theorem 3.2 is the following standard lemma.

Lemma 3.3: There exists an absolute constant c such that for every $r, \lambda > 0$

$$\mathbb{E}\sup_{f\in F_{\lambda,r}}\left|\frac{1}{k}\sum_{i=1}^{k}\left(\mathcal{L}_\lambda(f)\right)(X_i) - \mathbb{E}\mathcal{L}_\lambda(f)\right| \le \frac{c\|\ell\|_{\mathrm{lip}}}{k}\mathbb{E}\sup_{f,h\in F_{\lambda,r}}\left|\sum_{i=1}^{k}g_i\left(f-h\right)(X_i)\right| \quad (3.1)$$

where $(g_i)_{i=1}^k$ are independent, standard Gaussian variables.

Proof: Fix $r, \lambda > 0$. By the Giné–Zinn symmetrization argument [11] and the fact that the expectation of the supremum of a Rademacher processes is dominated by the expectation of the supremum of the Gaussian process indexed by the same set (see, for example, [13, Ch. 4])

$$\begin{aligned}\mathbb{E}\sup\nolimits_{f\in F_{\lambda,r}}&\left|\frac{1}{k}\sum_{i=1}^{k}\left(\mathcal{L}_\lambda(f)\right)(X_i) - \mathbb{E}\mathcal{L}_\lambda(f)\right|\\ &\le 2\mathbb{E}_X\mathbb{E}_\varepsilon\sup_{f\in F_{\lambda,r}}\left|\frac{1}{k}\sum_{i=1}^{k}\varepsilon_i(\mathcal{L}_\lambda(f))(X_i)\right|\\ &\le c\mathbb{E}_X\mathbb{E}_g\sup_{f\in F_{\lambda,r}}\left|\frac{1}{k}\sum_{i=1}^{k}g_i(\mathcal{L}_\lambda(f))(X_i)\right|\end{aligned}$$

where c is an absolute constant.

Clearly, for every k-sample $(X_i)_{i=1}^k$ and every $h, f \in F_{\lambda,r}$

$$\begin{aligned}\mathbb{E}_g&\left(\sum_{i=1}^{k}g_i(\mathcal{L}_\lambda(f))(X_i) - \sum_{i=1}^{k}g_i(\mathcal{L}_\lambda(h))(X_i)\right)^2\\ &\le \|\ell\|_{\mathrm{lip}}^2\sum_{i=1}^{k}\left((f-f_*)-(f_*-h)\right)^2(X_i)\\ &= \|\ell\|_{\mathrm{lip}}^2\mathbb{E}_g\left(\sum_{i=1}^{k}g_i\left(f-f_*\right) - \sum_{i=1}^{k}g_i\left(f_*-h\right)(X_i)\right)^2.\end{aligned}$$

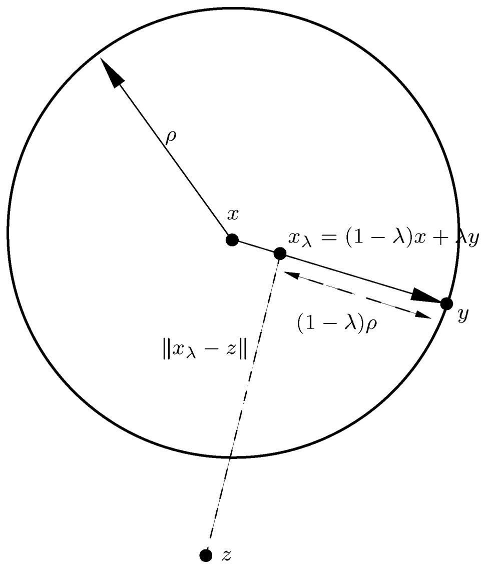

Fig. 2. Illustration of Lemma 3.4.

Therefore, by Slepian's lemma [9], for every k-sample

$$\mathbb{E}_g \sup_{f\in F_{\lambda,r}} \left| \frac{1}{k}\sum_{i=1}^{k} g_i(\mathcal{L}_\lambda(f))(X_i) \right| \le \mathbb{E}_g \sup_{f,h\in F_{\lambda,r}} \left| \frac{1}{k}\sum_{i=1}^{k} g_i(f-h)(X_i) \right|$$

and the claim follows by taking the expectation with respect to $X_1,\ldots,X_k$. □

The next step is to show that the set $F_{\lambda,r}$ is contained in a relatively small ball (with respect to the norm) around f_*. The main point in the proof of this fact is the following geometric lemma.

Lemma 3.4: Let E be a uniformly convex, smooth normed space and consider $x, y \in E$ and $\rho \in \mathbb{R}_+$ such that $\|y-x\| = \rho$. Let $0 < \lambda < 1$ and set $x_\lambda = (1-\lambda)x + \lambda y$. If z satisfies that $\|z - x\| \ge \rho$, then

$$\|x_\lambda - z\| - (1-\lambda)\rho \ge 2\lambda \|z-x\| \delta_E\left(\frac{\|z-y\|}{\|z-x\|}\right).$$

Proof: Without loss of generality, we can assume that $x = 0$. Fix $z \ne y$ and by our assumption, $\|z\| \ge \|y\|$. Define the function

$$H(\lambda) = \frac{\|x_\lambda - z\|}{\|z\|} = \frac{\|\lambda y - z\|}{\|z\|}$$

and observe that H is a convex function and $H(0) = 1$. Also, since E is smooth, H is differentiable in $\lambda = 0$. Thus, $H(\lambda) - H(0) = H(\lambda) - 1 \ge H'(0)\lambda$. Therefore

$$H(\lambda) - (1-\lambda) \ge \big(H'(0)+1\big)\lambda$$

and to complete the proof one has to bound $H'(0)$ from below.

Applying the chain rule, $H'(0) = u^*\left(\frac{y}{\|z\|}\right)$, where u^* is the unique functional of norm one supporting the unit sphere in $-z/\|z\| \equiv u$. Let $v = y/\|z\|$ and since $\|u^*\| = 1$, then

$$u^*(u-v) \le \|u-v\| \le 2 - 2\delta_E(\|u+v\|) = 2 - 2\delta_E\left(\frac{\|y-z\|}{\|z\|}\right).$$

Because $u^*(u) = 1$, $-u^*(v) \le 1 - 2\delta_E\left(\frac{\|y-z\|}{\|z\|}\right)$, implying that

$$H(\lambda) - (1-\lambda) \ge \big(H'(0)+1\big)\lambda \ge 2\lambda\delta_E\left(\frac{\|y-z\|}{\|z\|}\right).$$

Therefore

$$\begin{aligned}\|x_\lambda - z\| - (1-\lambda)\rho &\ge \|x_\lambda - z\| - (1-\lambda)\|z\| \\ &\ge 2\lambda\|z\|\delta_E\left(\frac{\|y-z\|}{\|z\|}\right).\end{aligned}$$

□

The first application of Lemma 3.4 is that $\mathbb{E}\mathcal{L}_\lambda$ has a unique minimizer in F for any $0 < \lambda \le 1$, if the underlying space E is smooth and uniformly convex and $\mathbb{E}\ell(f,g) = \phi(\|f-g\|)$, as in Assumption 2.1.

Corollary 3.5: Consider the excess loss $\mathcal{L}_\lambda(f) = \ell(f,T_\lambda) - \ell(f_*,T_\lambda)$; set $D = \sup_{f\in F}\|T-f\|$ and $\rho = \|T - f_*\|$. Then, for every $0 < \lambda \le 1$ and every $f \in F$

$$\mathbb{E}\mathcal{L}_\lambda(f) \ge \phi'((1-\lambda)\rho)\cdot 2\lambda D\delta_E\left(\frac{\|f-f_*\|}{D}\right).$$

In particular, for every $0 < \lambda \le 1$, f_* is the unique minimizer of $\mathbb{E}\ell(\cdot,T_\lambda)$ in F.

Proof: Using Assumption 2.1 and applying Lemma 3.4 for $x = T$, $y = f_*$, and $z = f$

$$\begin{aligned}\mathbb{E}\mathcal{L}_\lambda(f) &= \phi(\|T_\lambda - f\|) - \phi(\|T_\lambda - f_*\|) \\ &\ge \phi'(\|T_\lambda - f_*\|)\cdot(\|T_\lambda - f\| - \|T_\lambda - f_*\|) \\ &\ge 2\lambda\|f-T\|\delta_E\left(\frac{\|f-f_*\|}{\|f-T\|}\right)\cdot\phi'(\|T_\lambda - f_*\|) \\ &\ge 2\lambda D\delta_E\left(\frac{\|f-f_*\|}{D}\right)\cdot\phi'(\|T_\lambda - f_*\|)\end{aligned}$$

where the first inequality follows from the convexity of ϕ, the second one is Lemma 3.4, and the last one is the monotonicity property of δ_E. Indeed, let $\varepsilon_2 = \|f-f_*\|/\|f-T\| \ge \varepsilon_1 = \|f-f_*\|/D$. Thus

$$\frac{D}{\|f-f_*\|}\cdot\delta_E\left(\frac{\|f-f_*\|}{D}\right) \le \frac{\|f-T\|}{\|f-f_*\|}\cdot\delta_E\left(\frac{\|f-f_*\|}{\|f-T\|}\right)$$

as claimed. □

Let us clarify the meaning of Corollary 3.5. First, we will be interested in "small" values of λ. With this in mind, $\phi'(\|T_\lambda - f_*\|) = \phi'((1-\lambda)\rho)$ is a positive constant (when $\lambda \to 0$, it tends to the derivative of ϕ at ρ, which is a fixed, positive number) and the term $D\delta_E\left(\frac{\|f-f_*\|}{D}\right)$ also does not depend on λ, but rather on properties of E, F, and T, and the distance between f and f^*. In particular, the minimizer of $\mathbb{E}\mathcal{L}_\lambda$ in F is unique, and moreover, for λ sufficiently small

$$\mathbb{E}\mathcal{L}_\lambda(f_1) \ge c\lambda$$

where c depends only on properties of E, F, ϕ, and $\|f_1 - f_*\|$.

The second outcome is that functions with risk $\mathbb{E}\mathcal{L}_\lambda(f) \le r$ are contained in a small ball around f_*. Since this is a straightforward application of Corollary 3.5, we omit its proof.

Corollary 3.6: Assume that $\delta_E(\varepsilon) \ge \eta\varepsilon^p$ for some fixed η. Then, for every $0 < \lambda < 1$ and $r > 0$

$$F_{\lambda,r} \subset B\left(f_*, C_0\left(\frac{r}{\lambda}\right)^{1/p}\right)$$

where C_0 depends only on ρ, ϕ, D, and η.

Set $B = B\left(f_*, C_0\left(\frac{r}{\lambda}\right)^{1/p}\right)$. Combining Corollary 3.6 with (3.1) shows that there is a constant C_1 such that for every k, with μ^k-probability of at least $5/6$

$$\sup_{\{f:\mathbb{E}\mathcal{L}_\lambda(f)\le r\}} |P_k(\mathcal{L}_\lambda(f))(X_i)| \le r + \frac{C_1\|\ell\|_{\rm lip}}{k}\mathbb{E}\sup_{f,h\in F\cap B}\left|\sum_{i=1}^k g_i(f-h)(X_i)\right| \quad (3.2)$$

where the expectation is with respect to both $(g_i)_{i=1}^k$ and $(X_i)_{i=1}^k$.

Proof of Theorem 3.2: Fix $\alpha, u > 0$, let $\beta > 0$ to be named later, and set C_1 as in (3.2). Applying the compatibility assumption for $u' = u/C_1\|\ell\|_{\rm lip}$, there is some k_0 and $q > 0$ such that for $k \ge k_0$

$$\frac{C_1\|\ell\|_{\rm lip}}{k}\mathbb{E}\sup_{\{f,h\in F:\|f-h\|\le q\}}\left|\sum_{i=1}^k g_i(f-h)(X_i)\right| \le \frac{u}{\sqrt{k}}.$$

Set β to satisfy that $q = 2C_0\left(\frac{\beta}{\alpha}\right)^{1/p}$ and since q can be made smaller if we wish, we can assume that $\beta \le u$.

By Corollary 3.6, taking $\lambda = \alpha/\sqrt{k}$ and $r = \beta/\sqrt{k}$, it is evident that

$$F_{\alpha/\sqrt{k},\beta/\sqrt{k}} \subset B\left(f_*, C_0\left(\frac{\beta}{\alpha}\right)^{1/p}\right)\cap F.$$

Thus, with probability at least $5/6$

$$\begin{aligned}\sup_{f\in\mathcal{L}_{\alpha/\sqrt{k},\beta/\sqrt{k}}} &|P_kF_{\alpha/\sqrt{k}}(f)| \\ &\le \frac{\beta}{\sqrt{k}} + \frac{C_1\|\ell\|_{\rm lip}}{k}\mathbb{E}\sup_{\{f,h\in F:\|f-h\|\le q\}}\left|\sum_{i=1}^k g_i(f-h)(X_i)\right| \\ &\le \frac{2u}{\sqrt{k}}.\end{aligned}$$

□

A. Constructing "Bad" Functions

So far, we showed that with high μ^k-probability, if $\lambda \sim 1/\sqrt{k}$, then functions $\mathcal{L}_\lambda(f)$ with an expectation smaller than $\sim 1/\sqrt{k}$ have an empirical expectation that is not very negative. To complete the proof of the lower bound, one has to construct functions with a very negative empirical excess loss $P_k\mathcal{L}_\lambda(f)$.

To make things even more difficult, recall that for every $f \in F$, $\mathbb{E}\mathcal{L}_\lambda(f) \ge 0$, and by the central limit theorem $k^{-1/2}\sum_{i=1}^k(\mathcal{L}_\lambda(f) - \mathbb{E}\mathcal{L}_\lambda(f))$ converges to a centered Gaussian variable. Thus, the hope of generating some excess loss function $\mathcal{L}_\lambda(g)$ with a relatively large expectation but with a negative empirical loss on a typical sample of cardinality k is realistic only if $\mathbb{E}\mathcal{L}_\lambda(g) \le c_1\sigma/\sqrt{k}$, where σ^2 is the variance of $\mathcal{L}_\lambda(g)$, because typical values of $P_k\mathcal{L}_\lambda(g)$ are $\mathbb{E}\mathcal{L}_\lambda(g) \pm c_1\sigma/\sqrt{k}$. Thus, it seems natural to expect that the bad behavior, if indeed exists, is generated by a family of functions, each one for a different value of k.

We will show that $\mathcal{L}_{\lambda_k}(f_1)$ is such a family of functions for the choice of $\lambda_k \sim 1/\sqrt{k}$.

Our starting point is the Berry–Esséen theorem (see, e.g., [20])

Theorem 3.7: There exists an absolute constant c for which the following holds. Let $(\xi)_{i=1}^k$ be independent, identically distributed random variables, with mean 0 and variance σ^2 and denote by $F_k(x)$ the distribution function of $\frac{1}{\sigma\sqrt{k}}\sum_{i=1}^k\xi_i$. Then

$$\sup_{x\in\mathbb{R}}|F_k(x) - \Phi(x)| \le c\frac{\mathbb{E}|\xi_1|^3}{\sigma^3\sqrt{k}}$$

where Φ is the distribution function of a standard Gaussian variable.

Applying the Berry–Esséen theorem to the random variable $\xi = g(X) - \mathbb{E}g$, we obtain the following.

Corollary 3.8: Let g be a function with variance σ and a finite third moment. Then, there is some constant k_1 that depends only on σ and $\mathbb{E}|g|^3$ and an absolute constant C_2, such that for every $k \ge k_1$, with probability at least $1/4$

$$\frac{1}{k}\sum_{i=1}^k g(X_i) \le \mathbb{E}g - C_2\frac{\sigma}{\sqrt{k}}.$$

In what follows, we abuse notation and write $\mathcal{L}(f) = \ell(T,f) - \ell(T,f_*)$.

Lemma 3.9: There exist constants C_3, C_4 and λ_0, k_1 that depend only on $\rho = \|T - f_*\|$, $\|\phi\|_{\rm lip}$, $\|\ell\|_{\rm lip}$, $D_3 = \sup_{f\in F}\|T - f\|_{L_3(\mu)}$, and $\Delta = \mathbb{E}|\mathcal{L}(f_1)|$ for which the following hold. For every $0 \le \lambda \le \lambda_0$:

1) $\mathbb{E}\mathcal{L}_\lambda(f_1) \le 2\lambda\rho\|\phi\|_{\rm lip}$;
2) $\mathrm{var}(\mathcal{L}_\lambda(f_1)) \ge \Delta^2/4$;
3) for every $k \ge k_1$ and $\lambda_k \le \min\{\lambda_0, C_3/\sqrt{k}\}$, with probability at least $1/4$, $k^{-1}\sum_{i=1}^k(\mathcal{L}_\lambda(f_1))(X_i) \le -C_4/\sqrt{k}$.

Proof: For the first part, since $\|T - f_1\| = \|T - f_*\|$ and ϕ is Lipschitz, then

$$\begin{aligned}\mathbb{E}\mathcal{L}_\lambda(f_1) &= \phi(\|T_\lambda - f_1\|) - \phi(\|T_\lambda - f_*\|)\\ &= (\phi(\|T_\lambda - f_1\|) - \phi(\|T - f_1\|)) \\ &\quad + (\phi(\|T - f_*\|) - \phi(\|T_\lambda - f_*\|))\\ &\le 2\|\phi\|_{\rm lip}\|T - T_\lambda\| = 2\lambda\|\phi\|_{\rm lip}\|T - f_*\|\\ &= 2\lambda\rho\|\phi\|_{\rm lip}.\end{aligned}$$

Turning to the second part, recall that $\mathbb{E}|\mathcal{L}(f_1)| = \Delta > 0$. Since $\mathcal{L}_\lambda(f_1)$ tends to $\mathcal{L}(f_1)$ in L_2 as $\lambda \to 0$, then by the first part and standard calculations, there is some λ_0 such that for any $0 < \lambda \le \lambda_0$

$$\mathrm{var}(\mathcal{L}_\lambda(f_1)) \ge \Delta^2 - \mathbb{E}|\mathcal{L}^2(f_1) - \mathcal{L}_\lambda^2(f_1)| - (\mathbb{E}\mathcal{L}_\lambda(f_1))^2 \ge \frac{\Delta^2}{4}.$$

Finally, let k_1 as in Corollary 3.8 and set $0 < \lambda < \lambda_0$. Since $\Delta^2/4 \le \mathrm{var}(\mathcal{L}_\lambda(f_1))$ and

$$\mathbb{E}|\ell(T_\lambda, f_1) - \ell(T_\lambda, f_*)|^3 \le \|\ell\|_{\rm lip}^3\mathbb{E}|f_1 - f_*|^3$$

this choice of k_1 does not depend on λ as long as $\lambda \le \lambda_0$. Therefore, by Corollary 3.8, with probability at least $1/4$

$$\frac{1}{k}\sum_{i=1}^k(\mathcal{L}_\lambda(f_1))(X_i) \le \mathbb{E}\mathcal{L}_\lambda(f_1) - C_2\frac{\Delta}{2\sqrt{k}} \le 2\lambda\rho\|\phi\|_{\rm lip} - C_2\frac{\Delta}{2\sqrt{k}}.$$

Setting $C_3 = C_2\Delta/(8\rho\|\phi\|_{\rm lip})$, it is evident that if one takes

$$\lambda \le \min\{\lambda_0, C_3/\sqrt{k}\}$$

the claim follows. □

Now, let us formulate and prove our main result. To make the formulation simpler, we will refer to the space E, $\rho = \|T - f_*\|$, $\Delta =$

$\mathbb{E}|\mathcal{L}(f_1)|$, $D = \sup_{f \in F} \|T - f\|$, and $D_3 = \sup_{f \in F} \|T - f\|_{L_3(\mu)}$, the functions ϕ and ℓ, and the asymptotic equicontinuity properties of F (see Definition 3.1) as "the parameters of the problem." The constants in the formulation and the proof of the main result depend on these parameters.

Theorem 3.10: Assume that E is a smooth, uniformly convex normed space with a modulus of convexity $\delta_E(\varepsilon) \geq \eta\varepsilon^p$ for some $2 \leq p < \infty$ and $\eta > 0$. Let ϕ, ℓ, T, T_λ, and $\mathcal{L}_\lambda$ as above and consider $F \subset E$ that is compatible with the norm. There are constants λ' and k' and c_1, c_2 depending on the parameters of the problem for which the following holds. Let $k \geq k'$ and set $\lambda_k = \min\{\lambda', c_1/\sqrt{k}\}$. Then, with probability at least $1/12$, the empirical minimizer $\hat{f}$ of $\sum_{i=1}^k \ell(f, T_{\lambda_k})(X_i)$ satisfies

$$\mathbb{E}\left(\ell(f, T_{\lambda_k})|X_1, \ldots, X_k\right) \geq \inf_{f \in F} \mathbb{E}\ell(f, T_{\lambda_k}) + \frac{c_2}{\sqrt{k}}.$$

Proof: Let f_1 be the function constructed in Lemma 3.9 and set k_1, λ_0, C_3, C_4, and λ_k as in the formulation of the lemma. Let $\lambda' = \lambda_0$ and $k > k_1$. Clearly, by increasing k_1 if needed, one can assume $\lambda_k = C_3/\sqrt{k}$. Therefore, by Lemma 3.9, with μ^k-probability $1/4$, $P_k\mathcal{L}_\lambda(f_1) \leq -C_4/\sqrt{k}$. Hence, for each $k > k_1$ with that probability, the empirical minimizer with respect to the target $T_{C_3/\sqrt{k}}$ must have an empirical error smaller than $-C_4/\sqrt{k}$.

We now apply Theorem 3.2 for $\alpha = C_3$ and $u = C_4/4$. Let $\beta > 0$ and k_0 be as in the assertion of that theorem for those values of α and u, and put $k' = \max\{k_0, k_1\}$. Hence, if $k > k'$ and $r_k = \beta/\sqrt{k}$, then with μ^k-probability $5/6$

$$\sup_{f \in F_{\lambda_k, r_k}} |P_k\mathcal{L}_{\lambda_k}(f)| = \sup_{f \in F_{C_3/\sqrt{k}, \beta/\sqrt{k}}} \left|P_k\mathcal{L}_{C_3/\sqrt{k}}(f)\right| \leq \frac{2u}{\sqrt{k}} = \frac{C_4}{2\sqrt{k}}.$$

Therefore, with μ^k-probability of at least $1/12$, the empirical minimizer with respect to the target $T_{C_3/\sqrt{k}}$ is outside F_{λ_k, r_k} and thus its risk satisfies

$$\mathbb{E}\left(\mathcal{L}_{\lambda_k}(\hat{f})|X_1, \ldots, X_k\right) \geq \frac{\beta}{\sqrt{k}}.$$

□

As we mentioned before, it is well known that μ-Donsker classes are compatible with the $L_2(\mu)$ norm and thus are compatible with any $L_p(\mu)$ norm for $2 \leq p < \infty$. Because the L_p norms are uniformly convex (with a polynomial modulus of convexity) and smooth for $1 < p < \infty$, Theorem 1.2 follows from Theorem 3.10.

ACKNOWLEDGMENT

The author would like to thank O. Neviot for many stimulating discussions.

REFERENCES

[1] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[2] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," *Ann. Statist.*, vol. 33, no. 4, pp. 1497–1537, 2005.

[3] P. L. Bartlett and S. Mendelson, "Empirical minimization," *Probab. Theory Related Fields*, vol. 135, pp. 311–334, 2006.

[4] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: A Survey of recent advances," *ESAIM: Probab. Statist.*, vol. 9, pp. 323–375, 2005.

[5] O. Bousquet, "Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms," Ph.D. dissertation, Dept. Comput. Sci., Ecole Polytechnique, Palaiseau, France, 2002.

[6] R. Deville, G. Godefroy, and V. Zizler, *Smoothness and Renorming in Banach Spaces*. New York: Wiley, 1993.

[7] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[8] J. Diestel, "Sequences and series in Banach spaces," in *Graduate Text in Mathematics*. Berlin, Germany: Springer-Verlag, 1984, vol. 92.

[9] R. M. Dudley, *Uniform Central Limit Theorems*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[10] P. Habala, P. Hajek, and V. Zizler, *Introduction to Banach Spaces*. Prague, Czech Republic: Univ. Karlovy, Matfyzpress, 1996, vol. I and II.

[11] E. Giné and J. Zinn, "Some limit theorems for empirical processes," *Ann. Probab.*, vol. 12, no. 4, pp. 929–989, 1984.

[12] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, ser. Statistics. Berlin, Germany: Springer-Verlag, 2002.

[13] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. New York: Springer-Verlag, 1991.

[14] G. Lugosi, "Pattern classification and learning theory," in *Principles of Nonparametric Learning*, L. Györfi, Ed. New York: Springer-Verlag, 2002, pp. 1–56.

[15] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "The importance of convexity in learning with squared loss," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1974–1980, Sep. 1998.

[16] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "Correction to "The importance of convexity in learning with squared loss"," 2008, prepint.

[17] S. Mendelson, "Improving the sample complexity using global data," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1977–1991, Jul. 2002.

[18] S. Mendelson, "Geometric parameters in learning theory," in *GAFA Lecture Notes*. Berlin, Germany: Springer-Verlag, 2004, vol. 1850, LNM, pp. 193–236.

[19] S. Mendelson, "Obtaining fast error rates in nonconvex situations," *J. Complexity*, to be published.

[20] D. Stroock, *Probability Theory, an Analytic View*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[21] S. van de Geer, *Empirical Processes in M-Estimation*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[22] A. W. Van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York: Springer-Verlag, 1996.