Bernoulli **16**(3), 2010, 605–613
DOI: 10.3150/09-BEJ225

Sharper lower bounds on the performance of the empirical risk minimization algorithm

GUILLAUME LECUÉ[1] and SHAHAR MENDELSON[2]

[1]*CNRS, LATP, Marseille 13000, France. E-mail: lecue@latp.univ-mrs.fr*
[2]*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.*
E-mail: shahar.mendelson@anu.edu.au

We present an argument based on the multidimensional and the uniform central limit theorems, proving that, under some geometrical assumptions between the target function T and the learning class F, the excess risk of the empirical risk minimization algorithm is lower bounded by

$$\frac{\mathbb{E}\sup_{q\in Q} G_q}{\sqrt{n}}\delta,$$

where $(G_q)_{q\in Q}$ is a canonical Gaussian process associated with Q (a well chosen subset of F) and δ is a parameter governing the oscillations of the empirical excess risk function over a small ball in F.

Keywords: empirical risk minimization; learning theory; lower bound; multidimensional central limit theorem; uniform central limit theorem

1. Introduction

In this note, we study lower bounds on the empirical minimization algorithm. To explain the basic setup of this algorithm, let (Ω, μ) be a probability space and set X to be a random variable taking values in Ω, distributed according to μ. We are interested in the *function learning* (noiseless) problem, in which one observes n independent random variables $X_1, \dots, X_n$, distributed according to μ, and the values $T(X_1), \dots, T(X_n)$ of an unknown target function T.

The goal is to construct a procedure that uses the data $D = (X_i, T(X_i))_{1\le i\le n}$ with a *risk* as close as possible to the best one in F. That is, we want to construct a statistic $\hat{f}_n$ such that for every n, with high μ^n-probability,

$$R(\hat{f}|D) \le \inf_{f\in F} R(f) + r_n(F), \tag{1.1}$$

where the risk of f is defined by $R(f) = \mathbb{E}\ell(f(X), T(X))$ and $\ell : \mathbb{R}^2 \to \mathbb{R}$ is the loss function that measures the pointwise error between T and f. The residue $r_n(F)$ somehow captures the complexity or richness of the class F and the risk of a statistic $\hat{f}$ is the conditional expectation $R(\hat{f}|D) = \mathbb{E}(\ell(\hat{f}(X), T(X))|D)$.

It is well known (see, e.g., [10]) that if the class F is not too large, for example, if it satisfies some kind of uniform central limit theorem, T is bounded by 1 and ℓ is reasonable, then there are

upper bounds on $r_n(F)$ that are of the form $\sqrt{\mathrm{Comp}(F)/n}$, where $\mathrm{Comp}(F)$ is a complexity term that is independent of n. The algorithm that is used to produce the function $\hat{f}$ is the empirical risk minimization algorithm, in which one chooses a function in F that minimizes the empirical risk function $f \longmapsto \sum_{i=1}^n \ell(f,T)(X_i)$ in F.

There is a well developed theory concerning ways in which the complexity term may be controlled, using various parameters associated with the geometry of the class (cf. [2,8–10] and references therein). It turns out that this type of error rate, $\sim 1/\sqrt{n}$, is very pessimistic in many cases. In fact, if the class is small enough, then, under some structural assumptions (see, e.g., [1]), $r_n(F)$ can be much smaller – of the order of $\mathrm{Comp}(F)/n$.

In this note, we are going to focus on "small classes" in which empirical minimization performs poorly, despite the size of the class. Recently, it has been shown (cf. [7]) that under mild assumptions on ℓ and F, if there is more than a single function in

$$V := \Big\{ \ell(f,T)\colon \mathbb{E}\ell(f,T) = \inf_{f\in F} \mathbb{E}\ell(f,T) \Big\},$$

then the following holds: for every n large enough, there will be a perturbation T_n of T (with respect to the L_∞-norm) for which $\mathbb{E}\ell(\cdot,T_n)$ has a unique minimizer in F, but where the empirical minimization algorithm performs poorly trying to predict T_n on samples of cardinality n. To be more exact, relative to the target T_n, with μ^n-probability at least $1/12$,

$$R(\hat{f}|D) \geq \inf_{f\in F} R(f) + \frac{c}{\sqrt{n}}, \tag{1.2}$$

where c is a constant depending only on F.

Although it is reasonable to expect that the larger the set V is, the more likely it is that the empirical minimization algorithm will perform poorly, this does not follow from the analysis in [7]. Therefore, our goal here is to provide a bound on the constant c in (1.2) that does take into account the complexity of the set of minimizers V.

Just as in [7], our method of analysis can be applied to a wide variety of losses. However, for the sake of simplicity, we will only present here what is arguably the most important case – that in which the risk is measured relative to the squared loss, $\ell(x,y) = (x-y)^2$.

To explain our result, we need several definitions from empirical processes theory. Other standard notions we require from the theory of Gaussian processes can be found in [2].

For every set $F \subset L_2(\Omega,\mu)$, let $\{G_f\colon f \in F\}$ be the canonical Gaussian process indexed by F (i.e., with the covariance structure $\mathbb{E}G_tG_s = \langle s,t\rangle$) and set $H(F) = \mathbb{E}\sup_{f\in F} G_f$ – the expectation of the supremum of the Gaussian process indexed by F. Also, for every integer n and δ, let

$$\mathrm{osc}_n(F,\delta) := \frac{1}{\sqrt{n}}\mathbb{E} \sup_{\{f,h\in F:\|f-h\|\leq\delta\}} \left| \sum_{i=1}^n g_i(f-h)(X_i) \right|,$$

where $(g_i)_{i=1}^n$ are standard, independent Gaussian random variables and $(X_i)_{i=1}^n$ are independent, distributed according to μ. It is well known that if F is a class consisting of uniformly bounded functions, then it is a μ-Donsker class if and only if for every $\delta > 0$, $\mathrm{osc}_n(F,\delta)$ tends

to 0 as n tends to infinity (cf. [2], page 301). For any $f \in F$, let

$$\mathrm{osc}_n(F, f, \delta) := \frac{1}{\sqrt{n}}\mathbb{E} \sup_{\{h\in F:\|f-h\|\le\delta\}} \left|\sum_{i=1}^n g_i(f-h)(X_i)\right|,$$

that is, the oscillation in a ball around f. The quantity $\mathrm{osc}_n(F, f^*, \delta)$ is a natural upper bound for some intrinsic quantity of the problem we study here (cf. Lemma 2.3).

Let V be as above – the set of loss functions $\ell(f, T)$ that minimize the risk in F – select $f^* \in F$ for which $\ell(f^*, T) \in V$ and consider the following subset of excess loss functions:

$$Q := \{\ell(f, T) - \ell(f^*, T):\ \ell(f, T) \in V\}.$$

It turns out that the desired constant in (1.2) can be bounded from below by two parameters: the expectation of the supremum of the canonical Gaussian process indexed by Q and the oscillation around f^*. In particular, if Q is a rich set and one of the minimizers of $f \to \mathbb{E}\ell(f, T)$ is isolated, then for any n large enough, the error of the empirical minimizer with respect to a wisely selected target (denoted by T_{λ_n} in what follows) which is a perturbation of T will be at least $\sim H(Q)/\sqrt{n}$. The core idea of this work is that a small, wisely chosen perturbation of a target function T with multiple oracles (functions achieving $\min_{f\in F} \mathbb{E}\ell(t, T)$) is badly estimated by the empirical risk minimization procedure (for further discussion of this fact, we refer the reader to [7]).

Although the general philosophy of the proof presented here is similar to the proof from [7], it is much simpler. And, in fact, it seems that the method used in the proof from [7] cannot be directly extended to obtain the sharper estimate on the constant as we do here. Naturally, this result recovers the previous estimates on lower bounds for the empirical risk minimization algorithm from [3–6]

Next, a word about notation. Throughout, all absolute constants will be denoted by c, c_1 and C, C_1, etcetera. Their values may change from line to line.

If $\mathbb{E}\ell(\cdot, T)$ has a unique minimizer in F, then we denote it by f^*. If the minimizer is not unique, then we will fix one function in the set of minimizers and denote it by f^*. For every $f \in F$, let $\mathcal{L}(f) = \ell(f, T) - \ell(f^*, T)$ be the excess loss function associated with the target T. For every $0 < \lambda \le 1$, set $T_\lambda = (1-\lambda)T + \lambda f^*$ and define $\mathcal{L}_\lambda(f) = \ell(f, T_\lambda) - \ell(f^*, T_\lambda)$. It is standard to verify (cf. [7] or Theorem 2.1 in what follows) that f^* is a minimizer of $\mathbb{E}\ell(\cdot, T_\lambda)$ and that under mild convexity assumptions on ℓ that clearly hold if ℓ is the squared loss, it is the unique minimizer in F of $f \to \mathbb{E}\ell(f, T_\lambda)$.

If $X_1, \ldots, X_n$ is an independent sample selected according to μ, we set $P_n f = n^{-1} \times \sum_{i=1}^n f(X_i)$ and let $Pf = \mathbb{E}f$. Thus, $\mathbb{E}\sup_{f\in F} |(P_n - P)(f)|$ is the expectation of the supremum of the empirical process indexed by F. Finally, when the target function is T_λ, we will denote the function produced by the empirical risk minimization algorithm by $\hat{f}_\lambda$ – which is one element of the set $\mathrm{Arg}\min_{f\in F} P_n\ell(f, T_\lambda)$.

Finally, if E is a normed space, we denote its unit ball by $B(E)$, the inner product of $L_2(\mu)$ will be denoted by $\langle\cdot,\cdot\rangle$ and the corresponding norm by $\|\cdot\|$.

Let us now formulate our main result.

Theorem 1.1. *Let $F \subset L_2(\mu) \cap B(L_\infty)$, which is μ-pre-Gaussian (cf. [2]), and assume that $T \in B(L_\infty)$. Set ℓ to be the squared loss and put $Q = \{\mathcal{L}(f):\ f \in F, \mathbb{E}\mathcal{L}(f) = 0\}$.*

There exist some absolute constants C_1 and C_2 and an integer $N(F)$ for which the following holds. For every $n \geq N(F)$, with μ^n-probability at least C_1,

$$\mathbb{E}\mathcal{L}_{\lambda_n}(\hat{f}_{\lambda_n}) \geq C_2 \frac{H(Q)}{\sqrt{n}} \delta^2 \|T - f^*\|,$$

where δ is such that for every integer $n \geq N(F)$, $\mathrm{osc}_n(F, f^*, \delta) \leq C_2 H(Q)/\sqrt{n}$ and $\lambda_n = C_2 H(Q)/\sqrt{n}$.

Thus, two parameters control the behavior of the constant in (1.2): the complexity of the set of excess loss functions of the oracles of T and the parameter δ. When one of the oracles f^* of T is isolated, one can take δ as an absolute constant. This leads to a lower bound of the order of $H(Q)/\sqrt{n}$, which is optimal in the sense that an upper bound can be obtained of the order of $H(Q_0)/\sqrt{n}$ for some set Q_0 such that $Q \subset Q_0 \subset \mathcal{L}_F$ (see, e.g., [1] or [3]). In other settings, the lower bound obtained in Theorem 1.1 may fail to match exactly with an upper bound. For instance, in settings where the oscillation function $\mathrm{osc}_n(F, f^*, \cdot)$ of all the oracles f^* of T decreases to zero very slowly and at the same convergence rate, the factor δ^2 should break down the lower bound, whereas it seems that it should not appear in the lower bound. From a technical point of view, this comes from the fact that we did not take into account the complexity "around" the points in Q' (cf. Theorem 2.2 and equation (2.2) in what follows).

Finally, the noiseless model considered here is the worst case scenario to prove the lower bound. Indeed, adding some noise to the target function would increase the lower bound.

2. The lower bound

The core of the proof is to find a set that can "compete" with a set $B_r = \{f \in F\colon \mathbb{E}\mathcal{L}_\lambda(f) \leq r\}$ that contains f^*, in the sense that the empirical excess risk function

$$\mathcal{E}_n : f \in F \longmapsto \frac{1}{n}\sum_{i=1}^n \mathcal{L}_\lambda(f)(X_i)$$

will be more negative on the set than it can possibly be on B_r. Once this task is achieved, it is obvious that the empirical risk minimization algorithm will produce a function $\hat{f}_\lambda$ which is outside B_r and, thus, with a certain probability,

$$\mathbb{E}[\mathcal{L}_\lambda(\hat{f}_\lambda)|D] > r.$$

Hence, the proof consists of two parts. First, we will show that the empirical excess risk function $\mathcal{E}_n$ is likely to be very negative on Q and we will then find some r on which the oscillations in B_r are small.

The first result we need is the following lower estimate on the expectation of the excess loss relative to the target $T_\lambda = (1-\lambda)T + \lambda f^*$, according to the distance of f from f^*. This proposition is based on the fact that the functional $(f, g) \longmapsto \mathbb{E}\ell(f, g)$ inherits a strong convex structure from the norm and was proven in [7] in a far more general situation.

Theorem 2.1. *Let $D = \sup_{f\in F} \|T - f\|$ and $\rho = \|T - f^*\|$. There exists an absolute constant c such that for any function $f \in F$, if $0 \le \lambda \le 1/2$, $r > 0$ and*

$$\frac{r}{\lambda} \le c\frac{\rho}{D}\|f - f^*\|^2,$$

then

$$r \le \mathbb{E}\mathcal{L}_\lambda(f).$$

Recall that $Q = \{\mathcal{L}(f)\colon \mathbb{E}\mathcal{L}(f) = 0,\ f \in F\}$ is the set of excess loss functions associated with the true minimizers of $f \to \mathbb{E}\ell(f, T)$ in F. We will show that if $Q' \subset Q$ is a finite set, then for n large enough, with a non-trivial μ^n-probability there will be some $\mathcal{L}(f) \in Q'$ for which the empirical error $P_n\mathcal{L}_{\lambda_n}(f)$ is very negative (for a well chosen λ_n).

Theorem 2.2. *There exist constants c_1, c_2 and c_3, depending only on the $L_\infty(\mu)$-diameter of $F \cup \{T\}$, for which the following holds. If Q' is a finite subset of Q that contains 0, then there exists an integer $n_0 = n_0(Q')$ such that for every integer $n \ge n_0$, with μ^n-probability at least c_1,*

$$\inf_{\mathcal{L}(f)\in Q'} \frac{1}{n}\sum_{i=1}^n (\mathcal{L}_{\lambda_n}(f))(X_i) \le -c_2\frac{H(Q')}{\sqrt{n}},$$

where $\lambda_n = c_3 H(Q')/\sqrt{n}$ and $H(Q') = \mathbb{E}\sup_{q\in Q'} G_q$ is the expectation of the canonical Gaussian process associated with Q'.

Proof. Let $M = |Q'|$ and recall that each $q \in Q' = \{q_1, \dots, q_M\}$ has mean zero. Consider the random vector $U = (q_1(X), \dots, q_M(X)) \in \mathbb{R}^M$ and let $(U_i)_{i=1}^\infty$ be independent copies of U (i.e., $U_i = (q_1(X_i), \dots, q_M(X_i))$). By the vector-valued central limit theorem (see, e.g., [2]), $n^{-1/2}\sum_{i=1}^n U_i$ converges weakly to the canonical Gaussian process indexed by Q', which we denote by G. Fix $t \le 0$ and $0 < c < 1$, to be given later, for which

$$A_t = \{x \in \mathbb{R}^M\colon \forall 1 \le j \le M, x_j > t\}$$

is such that $p := \Pr(G \in A_t) \le c$. Set $n_0 = n_0(t, c)$ to be such that for $n \ge n_0$,

$$\left|\Pr(G \in A_t) - \Pr\left(n^{-1/2}\sum_{i=1}^n U_i \in A_t\right)\right| \le \frac{1-p}{2},$$

which clearly exists by weak convergence. Since

$$\begin{aligned}\Pr\left(\exists 1 \le j \le M\colon n^{-1/2}\sum_{i=1}^n \langle U_i, e_j\rangle \le t\right) &= 1 - \Pr\left(n^{-1/2}\sum_{i=1}^n U_i \in A_t\right)\\ &\ge \frac{1-p}{2} \ge \frac{1-c}{2} =: c_1 > 0,\end{aligned}$$

it follows that, with probability at least c_1,

$$\inf_{q\in Q'}\frac{1}{n}\sum_{i=1}^n q(X_i)\le\frac{t}{\sqrt{n}}.$$

It remains to show that one may take $t=-(\mathbb{E}\sup_{q\in Q'}G_q)/4$. Indeed, by the symmetry of the Gaussian process, it follows that (for this choice of t)

$$p=\Pr(G\in A_t)=\Pr\Big(\sup_{q\in Q'}G_q<\Big(\mathbb{E}\sup_{q\in Q'}G_q\Big)/4\Big).$$

Let $Z=\sup_{q\in Q'}G_q$ and $\sigma^2=\sup_{q\in Q'}\mathbb{E}G_q^2$. Since $0\in Q'$, it follows that if $\mathbb{E}Z=0$, then it is clear that $p=1/2$. Otherwise, using the concentration property of Z around its mean (see, e.g., [9]) and since $\sigma\le c_0\mathbb{E}Z$ (where c_0 is an absolute constant), there exists an absolute constant $A>0$ such that

$$\mathbb{E}\big[Z\mathbb{1}_{[Z\ge\mathbb{E}Z+A\sigma]}\big]\le(\mathbb{E}Z)/4.$$

Therefore,

$$\begin{aligned}\mathbb{E}Z&=\mathbb{E}\big(Z\big(\mathbb{1}_{[Z\le(\mathbb{E}Z)/4]}+\mathbb{1}_{[(\mathbb{E}Z)/4\le Z\le\mathbb{E}Z+A\sigma]}+\mathbb{1}_{[Z\ge\mathbb{E}Z+A\sigma]}\big)\big)\\&\le(\mathbb{E}Z)/2+(\mathbb{E}Z)(1+c_0A)\Pr\big((\mathbb{E}Z)/4\le Z\big).\end{aligned}$$

Thus, $\Pr((\mathbb{E}Z)/4\le Z)\ge[2(1+c_0A)]^{-1}$ and so $p\le 1-[2(1+c_0A)]^{-1}:=c$ (which is an absolute constant), implying that, with probability greater than c_1,

$$\inf_{\mathcal{L}(f)\in Q'}\frac{1}{n}\sum_{i=1}^n(\mathcal{L}(f))(X_i)\le-c_2\frac{\mathbb{E}\sup_{q\in Q'}G_q}{\sqrt{n}}.$$

Next, observe that for small values of λ (as we will have in our construction), $\mathcal{L}(f)$ is a good approximation of $\mathcal{L}_\lambda(f)$ with respect to the $L_\infty(\mu)$-norm. Indeed, $\mathcal{L}_\lambda(f)=\ell(f,T_\lambda)-\ell(f^*,T_\lambda)$ and $\mathcal{L}(f)=\ell(f,T)-\ell(f^*,T)$; hence, for every $f\in F$,

$$\begin{aligned}\|\mathcal{L}_\lambda(f)-\mathcal{L}(f)\|_\infty&\le\|\ell(f,T_\lambda)-\ell(f,T)\|_\infty+\|\ell(f^*,T_\lambda)-\ell(f^*,T)\|_\infty\\&\le 2\|\ell\|_{\mathrm{lip}}\|T-T_\lambda\|_\infty=2\lambda\|\ell\|_{\mathrm{lip}}\|T-f^*\|_\infty\le c_3\lambda.\end{aligned}$$

Thus, if one selects $\lambda_n=(c_2/(2c_3))n^{-1/2}\mathbb{E}\sup_{q\in Q'}G_q$, then, with probability greater than c_1,

$$\inf_{\mathcal{L}(f)\in Q'}P_n\mathcal{L}_{\lambda_n}(f)\le-c_2\frac{\mathbb{E}\sup_{q\in Q'}G_q}{2\sqrt{n}}.$$

□

Fix a finite set $Q'\subset Q$ for which $H(Q')\ge H(Q)/2$ and $0\in Q'$. Clearly, such a set exists because Q is a pre-Gaussian as a subset of the pre-Gaussian class $\{\mathcal{L}(f)\colon\ f\in F\}$. Let $V'=\{f\in F\colon\ \mathcal{L}(f)\in Q'\}$.

Recall that a bounded class of functions F is *μ-Donsker* if and only if for every $u>0$, there exist $\delta>0$ and an integer n_0 such that for every $n\ge n_0$, $\mathrm{osc}_n(F,\delta)\le u$. Also, note that

$\operatorname{osc}_n(F, f^*, \delta) \le \operatorname{osc}_n(F, \delta)$. Let $u = \eta H(Q')$, where η is an absolute constant, to be fixed later, and set δ and n_1 to be such that for $n \ge n_1$,

$$\operatorname{osc}_n(F, f^*, \delta) \le \eta H(Q') \tag{2.1}$$

(such δ and n_1 necessarily exist because F is μ-Donsker).

The next lemma is standard and follows from a symmetrization argument combined with Slepian's lemma. Its proof may be found in, for example, [7].

Lemma 2.3. *There exists an absolute constant c for which the following holds. For any $F' \subset F$ such that $f^* \in F'$ and any $0 \le \lambda \le 1$,*

$$\mathbb{E} \sup_{f \in F'} |(P - P_n)(\mathcal{L}_\lambda(f))| \le c\mathbb{E} \sup_{f \in F'} \left| \frac{1}{n} \sum_{i=1}^n g_i (f - f^*)(X_i) \right|,$$

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

We are now ready to control the oscillation of the empirical excess risk function in the set $B_r = \{f \in F\colon \mathbb{E}\mathcal{L}_\lambda \le r\}$.

Theorem 2.4. *Let c_1, c_2 and λ_n be defined as in Theorem 2.2, and let δ and n_1 be as above. There exists an absolute constant c_3 such that for any integer $n \ge n_1$, with μ^n-probability at least $1 - c_1/2$,*

$$\inf_{\{f \in F : \mathbb{E}\mathcal{L}_{\lambda_n}(f) \le r_n\}} P_n \mathcal{L}_{\lambda_n}(f) \ge -\frac{c_2 H(Q')}{2\sqrt{n}},$$

where

$$r_n = c_3 \frac{H(Q')}{\sqrt{n}} \delta^2 \|T - f^*\|^2.$$

Proof. By Theorem 2.1, for any $r, \lambda > 0$, if $f \in F$ is such that $\mathbb{E}\mathcal{L}_\lambda(f) < r$, then

$$\frac{r}{\lambda} > c \frac{\rho}{D} \|f - f^*\|^2,$$

where D and ρ were defined in Theorem 2.1. Thus,

$$\{f \in F\colon \mathbb{E}\mathcal{L}_\lambda(f) < r\} \subset \left\{ f \in F\colon \|f - f^*\| < c_4 \sqrt{r/\lambda} \right\},$$

where $c_4 = c_4(\rho, D)$. Hence, by Lemma 2.3, for $n \ge n_1$,

$$\begin{aligned} \mathbb{E} \sup_{\{f \in F : \mathbb{E}\mathcal{L}_\lambda(f) < r\}} -P_n \mathcal{L}_\lambda(f) &\le c_5 \mathbb{E} \sup_{\{f \in F : \|f - f^*\| \le c_4 \sqrt{r/\lambda}\}} \left| \frac{1}{n} \sum_{i=1}^n g_i (f - f^*)(X_i) \right| \\ &\le \frac{c_5}{\sqrt{n}} \operatorname{osc}_n\left(F, f^*, c_4 \sqrt{r/\lambda}\right) \le \frac{c_5}{\sqrt{n}} \eta H(Q'), \end{aligned}$$

provided that $c_4\sqrt{r/\lambda} \leq \delta$. Thus, for an appropriate choice of η (e.g., $\eta = c_1c_2/(4c_5)$ would do) and setting $r_n := (c_3/(2c_4^2))n^{-1/2}H(Q')\delta^2$ (which is smaller than $\delta^2\lambda_n/c_4^2$), it is evident that

$$\mathbb{E} \sup_{\{f\in F:\mathbb{E}\mathcal{L}_{\lambda_n}(f)<r_n\}} -P_n\mathcal{L}_{\lambda_n}(f) \leq \frac{c_1c_2}{4\sqrt{n}}H(Q').$$

Therefore, with μ^n-probability at least $1 - c_1/2$,

$$\sup_{\{f\in F:\mathbb{E}\mathcal{L}_{\lambda_n}(f)<r_n\}} -P_n\mathcal{L}_{\lambda_n}(f) \leq \frac{c_2H(Q')}{2\sqrt{n}},$$

as claimed. □

We can now prove our main result.

Proof of Theorem 1.1. By Theorem 2.2 applied to the set Q', there exists some integer $n_0 = n_0(Q')$ such that for every $n \geq n_0$, with μ^n-probability at least c_1,

$$\inf_{\mathcal{L}(f)\in Q'} P_n\mathcal{L}_{\lambda_n}(f) \leq -c_2\frac{H(Q')}{\sqrt{n}}, \tag{2.2}$$

where c_1 and c_2 are two absolute constants.

By Theorem 2.4, for any integer $n \geq n_1$, with μ^n-probability at least $1 - c_1/2$,

$$\inf_{\{f\in F:\mathbb{E}\mathcal{L}_{\lambda_n}(f)<r_n\}} P_n\mathcal{L}_{\lambda_n}(f) \geq -\frac{c_2H(Q')}{2\sqrt{n}}. \tag{2.3}$$

Hence, combining equations (2.2) and (2.3), with μ^n-probability at least $c_1/2$, the excess risk of $\hat{f}_{\lambda_n}$ is such that $\mathbb{E}[\mathcal{L}_{\lambda_n}(\hat{f}_{\lambda_n})|D] \leq -c_2H(Q')/(\sqrt{n})$, while for every function $f \in F$ with $\mathbb{E}\mathcal{L}_{\lambda_n}(f) < r_n$, the empirical excess risk satisfies $P_n\mathcal{L}_{\lambda_n}(f) \geq -c_2H(Q')/(2\sqrt{n})$. Therefore, the empirical risk minimization algorithm has an excess risk (conditionally on the data D) larger than r_n, with probability greater than $c_1/2$, as claimed. □

Acknowledgements

This research was supported in part by Australian Research Council Discovery Grant DP0559465 and by Israel Science Foundation Grant 666/06.

References

[1] Bartlett, P.L. and Mendelson, S. (2006). Empirical minimization. *Probab. Theory Related Fields* **135** 311–334. MR2240689

[2] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge: Cambridge Univ. Press. MR1720712

[3] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442
[4] Lecué, G. (2007). Suboptimality of penalized empirical risk minimization in classification. In *20th Annual Conference On Learning Theory, COLT07* (G. Bshouty, ed.). *LNAI* **4539** 142–156. Berlin: Springer. MR2397584
[5] Lee, W.S., Bartlett, P.L. and Williamson, R.C. (1998). The importance of convexity in learning with squared loss. *IEEE Trans. Inform. Theory* **44** 1974–1980. MR1664079
[6] Massart P. and Nédélec, É. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366. MR2291502
[7] Mendelson, S. (2008). Lower bounds for the empirical minimization algorithm. *IEEE Trans. Inform. Theory*. **54** 3797–3803. MR2451042
[8] Talagrand, M. (2005). *The Generic Chaining*. *Springer Monographs in Mathematics*. Berlin: Springer-Verlag. MR2133757
[9] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. New York: Springer-Verlag. MR1385671
[10] Vapnik, V.N. (1998). *Statistical Learning Theory*. *Adaptive and Learning Systems for Signal Processing, Communications, and Control*. New York: Wiley. MR1641250

Received October 2008 and revised May 2009