Sharper lower bounds on the performance of the Empirical Risk Minimization Algorithm

Guillaume Lecué[1,3,4] Shahar Mendelson[2,3,5]

October 4, 2008

1 Introduction

In this note we study lower bounds on the empirical minimization algorithm. To explain the basic set up 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 $X_1, \ldots, X_n$ distributed according to μ, and the values $T(X_1), \ldots, T(X_n)$ of an unknown target function T.

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

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

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 $R(\hat{f}|D) = \mathbb{E}\big(\ell(\hat{f}(X), T(X))|D\big)$.

It is well known (see, for example, [8]) that if the class F is not too large, e.g., if it satisfies some kind of uniform Central Limit Theorem, T is bounded by 1 and ℓ is reasonable, 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

[1]CNRS, LATP, Marseille 13000, France.

[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.

[3]Supported in part by an Australian Research Council Discovery grant DP0559465 and by an Israel Science Foundation grant 666/06.

[4]Email: lecue@latp.univ-mrs.fr

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

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 on ways in which the complexity term may be controlled, using various parameters associated with the geometry of the class (cf. [7], [8], [2], [6] 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, for example, [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. [5]) that under mild assumptions on ℓ and F, if there is more than a single function in

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

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 the empirical minimization algorithm preforms 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, it does not follow from the analysis in [5]. Therefore, our goal here is to provide a bound on the constant c in (1.2) that does take into account of the complexity of the set of minimizers V.

Just like in [5], 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 – 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 by found in [2].

For every set $F \subset L_2(\Omega,\mu)$ let $\{G_f : f \in F\}$ be the canonical Gaussian process indexed by F (that is, 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], p.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.

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 let

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

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 T_n which is a perturbation of T will be at least $\sim H(Q)/\sqrt{n}$.

Although the general philosophy of the proof presented here is similar to the proof from [5], it is much simpler. And, in fact, it seems that the method used in the proof from [5] can not 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 [4, 3].

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

If $\mathbb{E}\ell(\cdot, T)$ has a unique minimizer in F we denote it by f^*. If the minimizer is not unique, 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 $T_\lambda = \lambda T + (1 - \lambda) f^*$ and denote $\mathcal{L}_\lambda(f) = \ell(f, T_\lambda) - \ell(f^*, T_\lambda)$. It is standard to verify (cf. [5] 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 $X_1, ..., X_n$ is an independent sample selected according to μ, set $P_n f = n^{-1} \sum_{i=1}^n f(X_i)$ 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, will be 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 formulate our main result:

Theorem 1.1 *Let $F \subset L_2(\mu) \cap B(L_\infty)$ which is μ-pregaussian (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 satisfies that for every integer $n \geq N(F)$, $\operatorname{osc}_n(F, f^*, \delta) \leq C_2 H(Q)/\sqrt{n}$ and $\lambda_n = C_2 H(Q)/\sqrt{n}$.

2 The lower bound

The core of the proof is to find a set that can "compete" with a set $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda(f) \leq r\}$ 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 on 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,

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

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 then we will 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 $T_\lambda = \lambda T + (1-\lambda) f^*$, 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 proved in [5] 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 $0 \leq \lambda \leq 1/2$, $r > 0$ and*

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

then

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

Recall that $Q = \{\mathcal{L}(f) : \mathbb{E}\mathcal{L}(f) = 0, f \in F\}$ is the set of the excess loss functions associated with the true minimizers of $f \to \mathbb{E}\ell(f,T)$. We will show that if $Q' \subset Q$ is a finite set, then for n large enough, with a nontrivial 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 exists 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 is an integer $n_0 = n_0(Q')$ such that for every integer $n \geq n_0$, with μ^n-probability at least c_1,*

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

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 $q \in Q' = \{q_1, \ldots, q_M\}$ has mean zero. Consider the random vector $U = (q_1(X), ..., q_M(X)) \in \mathbb{R}^M$ and let $(U_i)_{i=1}^\infty$ be independent copies of U (that is, $U_i = (q_1(X_i), \ldots, q_M(X_i))$). By the vector valued Central Limit Theorem (cf. 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 $t \leq 0$ and $0 < c < 1$ to be named later for which

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

satisfies that $p := Pr(G \in A_t) \leq c$. Set $n_0 = n_0(t,c)$ to be such that for $n \geq n_0$,

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

which clearly exists by the weak convergence. Since

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

then 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)

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

Let $Z = \sup_{q\in Q'} G_q$ and $\sigma^2 = \sup_{q\in Q'} \mathbb{E}G_q^2$. Since $0 \in Q'$ then if $\mathbb{E}Z = 0$, it is clear that $p = 1/2$. Otherwise, using the concentration property of Z around its mean (see, for example, [7]) 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, $Pr\big((\mathbb{E}Z)/4 \le Z\big) \ge [2(1+c_0A)]^{-1}$ 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 $f \in F$

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

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 $H(Q') \geq H(Q)/2$ and $0 \in Q'$. Clearly, such a set exists because Q is a pregaussian as a subset of the pregaussian class $\{\mathcal{L}(f) : f \in F\}$, and let $V' = \{f \in F : \mathcal{L}(f) \in Q'\}$.

Recall that a bounded class of functions F is μ-Donsker if and only if for every $u > 0$ there are $\delta > 0$ and an integer n_0 such that for every $n \geq n_0, \mathrm{osc}_n(F, \delta) \leq u$. Also, note that $\mathrm{osc}_n(F, f^*, \delta) \leq \mathrm{osc}_n(F, \delta)$. Let $u = \eta H(Q')$ where η is an absolute constant to be named later, and set δ and n_1 to be such that for $n \geq n_1$,

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

(such δ and n_1 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, for example, in [5].

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 $0 \leq \lambda \leq 1$,*

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

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

Now we are ready to control the oscillation of the empirical excess risk function in the set $B_r = \{f \in F : \mathbb{E}\mathcal{L}_\lambda \leq r\}$.

Theorem 2.4 *Let c_1, c_2 and λ_n be as defined 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 $n \geq n_1$, with μ^n-probability at least $1 - c_1/2$,*

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

where

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

Proof of Theorem 2.4. 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 have been defined in Theorem 2.1. Thus,

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

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

$$\mathbb{E} \sup_{\{f \in F : \mathbb{E}\mathcal{L}_\lambda(f) < r\}} -P_n \mathcal{L}_\lambda(f) \leq c_5 \mathbb{E} \sup_{\{f \in F : \|f - f^*\| \leq c_4 \sqrt{r/\lambda}\}} \Big| \frac{1}{n} \sum_{i=1}^n g_i (f - f^*)(X_i) \Big|$$
$$\leq \frac{c_5}{\sqrt{n}} \mathrm{osc}_n(F, f^*, c_4\sqrt{r/\lambda}) \leq \frac{c_5}{\sqrt{n}} \eta H(Q')$$

provided that $c_4\sqrt{r/\lambda} \leq \delta$. Thus, for an appropriate choice of η (e.g. $\eta = c_1 c_2/(4c_5)$ 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

$$\mathbb{E} \sup_{\{f \in F : \mathbb{E}\mathcal{L}_{\lambda_n}(f) < r_n\}} -P_n \mathcal{L}_{\lambda_n}(f) \leq \frac{c_1 c_2}{4\sqrt{n}} H(Q').$$

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

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

as claimed.

■

Now we can prove our main result.

Proof of Theorem 1.1. By Theorem 2.2 applied to the set Q', there is 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 $1 - c_1/2$,

$$\inf_{\{f \in F : \mathbb{E}\mathcal{L}_{\lambda_n}(f) < r_n\}} P_n \mathcal{L}_{\lambda_n}(f) \geq -\frac{c_2 H(Q')}{2\sqrt{n}}. \tag{2.3}$$

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 $\mathbb{E}[\mathcal{L}_{\lambda_n}(\hat{f}_{\lambda_n}) | D] \leq -c_2 H(Q')/(\sqrt{n})$, while for every function $f \in F$ with $\mathbb{E}\mathcal{L}_{\lambda_n}(f) < r_n$, the empirical excess risk satisfies $P_n \mathcal{L}_{\lambda_n}(f) \geq -c_2 H(Q')/(2\sqrt{n})$. Therefore, the empirical risk minimization algorithm has an excess risk (conditionally to the data D) larger than r_n with probability greater than $c_1/2$, as claimed. ■

References

[1] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probab. Theory Related Fields*, 135(3):311–334, 2006.

[2] R. M. Dudley. *Uniform central limit theorems*, volume 63 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999.

[3] Guillaume Lecué. Suboptimality of penalized empirical risk minimization in classification. In Gentile Bshouty, editor, *20th Annual Conference On Learning Theory, COLT07*, volume LNAI 4539, pages 142–156. Springer, 2007.

[4] Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. *IEEE Trans. Inform. Theory*, 44(5):1974–1980, 1998.

[5] S. Mendelson. Lower bounds for the empirical minimization algorithm. To appear in IEEE Transactions on Information Theory, 2008.

[6] Michel Talagrand. *The generic chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[7] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[8] Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. A Wiley-Interscience Publication.