Exponential Convergence Rates in Classification

Vladimir Koltchinskii* and Olexandra Beznosova

Department of Mathematics and Statistics
The University of New Mexico
Albuquerque, NM 87131-1141, U.S.A.
vlad@math.unm.edu,beznosik@math.unm.edu

Abstract. Let (X, Y) be a random couple, X being an observable instance and $Y \in \{-1, 1\}$ being a binary label to be predicted based on an observation of the instance. Let $(X_i, Y_i),\ i = 1, \dots, n$ be training data consisting of n independent copies of (X, Y). Consider a real valued classifier $\hat{f}_n$ that minimizes the following penalized empirical risk

$$\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i f(X_i)) + \lambda \|f\|^2 \to \min,\ f \in \mathcal{H}$$

over a Hilbert space $\mathcal{H}$ of functions with norm $\|\cdot\|$, ℓ being a convex loss function and $\lambda > 0$ being a regularization parameter. In particular, $\mathcal{H}$ might be a Sobolev space or a reproducing kernel Hilbert space. We provide some conditions under which the generalization error of the corresponding binary classifier $\mathrm{sign}(\hat{f}_n)$ converges to the Bayes risk exponentially fast.

1 Introduction

Let (S, d) be a metric space and (X, Y) be a random couple taking values in $S \times \{-1, 1\}$ with joint distribution P. The distribution of X (which is a measure on the Borel σ-algebra in S) will be denoted by Π. Let $(X_i, Y_i),\ i \geq 1$ be a sequence of independent copies of (X, Y). Here and in what follows all random variables are defined on some probability space $(\Omega, \Sigma, \mathbb{P})$. Let $\mathcal{H}$ be a Hilbert space of functions on S such that $\mathcal{H}$ is dense in the space $C(S)$ of all continuous functions on S and, in addition,

$$\forall x, y \in S\ |f(x)| \leq \|f\| \text{ and } |f(x) - f(y)| \leq \|f\| d(x, y). \tag{1}$$

Here $\|\cdot\| = \|\cdot\|_{\mathcal{H}}$ is the norm of $\mathcal{H}$ and $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\mathcal{H}}$ is its inner product.

We have in mind two main examples. In the first one, S is a compact domain in $\mathbb{R}^d$ with smooth boundary. For any $s \geq 1$, one can define the following inner product in the space $C^{\infty}(S)$ of all infinitely differentiable functions in S :

$$\langle f, g \rangle_s := \sum_{|\alpha| \leq s} \int_S D^{\alpha} f D^{\alpha} g dx.$$

* Partially supported by NSF grant DMS-0304861

Here $\alpha = (\alpha_1, \ldots, \alpha_d)$, $\alpha_j = 0, 1, \ldots$, $|\alpha| := \sum_{i=1}^d \alpha_i$ and

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}.$$

The Sobolev space $\mathcal{H}^s(S)$ is the completion of $\left(C^\infty(S), \langle \cdot, \cdot \rangle_s\right)$. There is also a version of the definition for any real $s > 0$ that utilizes Fourier transforms. If $s > d/2 + 1$, then it follows from Sobolev's embedding theorems that conditions (1) hold with metric d being the Euclidean distance (possibly, after a proper "rescaling" of the inner product or of the metric d to make constants equal to 1).

In the second example, S is a metric compact and $\mathcal{H} = \mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS) generated by a Mercer kernel K. This means that K is a continuous symmetric nonnegatively definite kernel and $\mathcal{H}_K$ is defined as the completion of the linear span of functions $\{K_x : x \in S\}$, $K_x(y) := K(x, y)$, with respect to the following inner product:

$$\left\langle \sum_i \alpha_i K_{x_i}, \sum_j \beta_j K_{y_j} \right\rangle_K := \sum_{i,j} \alpha_i \beta_j K(x_i, y_j).$$

It is well known that $\mathcal{H}_K$ can be identified with a subset of $C(S)$ and

$$\forall f \in \mathcal{H}_K \ f(x) = \langle f, K_x \rangle_K,$$

implying that

$$|f(x)| \le \|f\|_K \sup_{x \in S} \|K_x\|_K \text{ and } |f(x) - f(y)| \le \|f\|_K \|K_x - K_y\|_K,$$

so again conditions (1) hold with $d(x, y) := \|K_x - K_y\|_K$ (as before, a simple rescaling is needed to ensure that the constants are equal to 1).

In binary classification problems, it is common to look for a real valued classifier $\hat{f}_n$ that solves the following penalized empirical risk minimization problem

$$\frac{1}{n} \sum_{i=1}^n \ell(Y_i f(X_i)) + \lambda \|f\|^2 \to \min, \ f \in \mathcal{H}, \tag{2}$$

where ℓ is a nonnegative decreasing convex loss function such that $\ell \ge I_{(-\infty, 0]}$ and $\lambda > 0$ is a regularization parameter. For instance, if ℓ is a "hinge loss", i.e. $\ell(u) = (1 - u) \vee 0$, and $\|\cdot\|$ is a RKHS-norm, this is a standard approach in kernel machines classification.

Given a real valued classifier $f : S \mapsto \mathbb{R}$, the corresponding binary classifier is typically defined as $x \mapsto \text{sign}(f(x))$, where $\text{sign}(u) = +1$ for $u \ge 0$ and -1 otherwise. The generalization error or risk of f is then

$$R_P(f) := P\{(x, y) : y \ne \text{sign}(f(x))\}.$$

It is well known that the minimium of $R_P(f)$ over all measurable functions f is attained at the regression function η defined as

$$\eta(x) := \mathbb{E}(Y|X = x).$$

The corresponding binary classifier $\mathrm{sign}(\eta(x))$ is called the Bayes classifier, the quantity $R_* := R_P(\eta)$ is called the Bayes risk and, finally, the quantity $R_P(f) - R_*$ is often referred to as the excess risk of a classifier f.

Our goal in this note is to show that under some (naturally restrictive) assumptions the expectation of the excess risk of $\hat{f}_n$ converges to 0 *exponentially fast* as $n \to \infty$. Recently, Audibert and Tsybakov [1] observed a similar phenomenon in the case of plug-in classifiers and our analysis here continues this line of work.

Denote

$$\delta(P) := \sup\{\delta > 0 : \Pi\{x : |\eta(x)| \le \delta\} = 0\}.$$

We will assume that

(a) η is a Lipschitz function with constant $L > 0$ (which, for the sake of simplicity of notations, will be assumed to be 1 in what follows):

$$|\eta(x) - \eta(y)| \le L d(x, y).$$

(b) $\delta(P) > 0$.

These will be two main conditions that guarantee the possibilty of exponentially fast convergence rates of the generalization error to the Bayes risk. Note that condition (b), which is an extreme case of Tsybakov's low noise assumption, means that there exists $\delta > 0$ such that Π-a.e. either $\eta(x) \ge \delta$, or $\eta(x) \le -\delta$. The function η (as a conditional expectation) is defined up to Π-a.e. Condition (a) means that there exists a smooth (Lipschitz) version of this conditional expectation. Since smooth functions can not jump immediately from the value $-\delta$ to value δ, the combination of conditions (a) and (b) essentially means that there should be a wide enough "corridor" between the regions $\{\eta \ge \delta\}$ and $\{\eta \le -\delta\}$, but the probability of getting into this corridor is zero. The fact that in such situations it is possible to construct classifiers that converge to Bayes exponentially fast is essentially rather simple, it reduces to a large deviation type phenomenon, and it is even surprising that, up to our best knowledge, the possibility of such superfast convergence rates in classification has not been observed before Audibert and Tsybakov [1] (we apologize if someone, in fact, did it earlier).

Subtle results on convergence rates of the generalization error of large margin classifiers to the Bayes risk have been obtained relatively recently, see papers by Bartlett, Jordan and McAuliffe [3] and by Blanchard, Lugosi and Vayatis [5] on boosting, and papers by Blanchard, Bousquet and Massart [4] and by Scovel and Steinwart [7] on SVM. These papers rely heavily on general exponential inequalities in abstract empirical risk minimization in spirit of papers by Bartlett, Bousquet and Mendelson [2] or Koltchinskii [6] (or even earlier work by Birgé and Massart in the 90s). The rates of convergence in classification based on this

general approach are at best of the order $O(n^{-1})$. In classification problems, there are many relevant probabilistic, analytic and geometric parameters to play with when one studies the convergence rates. For instance, both papers [4] and [7] deal with SVM classifiers (so, essentially, with problem (2) in the case when $\mathcal{H}$ is RKHS). In [4], the convergence rates are studied under the assumption (b) above and under some conditions on the eigenvalues of the kernel. In [7], the authors determine the convergence rates under the assumption on the entropy of the unit ball in RKHS of the same type as our assumption (3) below, under Tsybakov's low noise assumption and some additional conditions of geomeric nature. The fact that under somewhat more restrictive assumptions imposed in this paper even exponential convergence rates are possible indicates that, probably, we have not understood to the end rather subtle interplay between various parameters that influence the behaviour of this type of classifiers.

2 Main Result

We now turn to precise formulation of the results. Our goal will be to explain the main ideas rather than to give the results in the full generality, so, we will make below several simplifying assumptions.

First, we need some conditions on the loss function ℓ and to get this out of the way, we will just assume that ℓ is the so called logit loss,

$$\ell(u) = \log_2(1 + e^{-u}),\ u \in \mathbb{R}$$

(other loss functions of the same type that are decreasing, strictly convex, satisfy the assumption $\ell \geq I_{(-\infty,0]}$ and grow slower than u^2 as $u \to \infty$ will also do). We denote

$$(\ell \bullet f)(x, y) := \ell(yf(x)).$$

For a function g on $S \times \{-1, 1\}$, we write

$$Pg = \int_{S\times\{-1,1\}} g dP = \mathbb{E}\, g(X, Y).$$

Let P_n be the empirical measure based on the training data $(X_i, Y_i),\ i = 1, \dots, n$. We will write

$$P_n g = \int_{S\times\{-1,1\}} g dP_n = n^{-1} \sum_{i=1}^{n} g(X_i, Y_i).$$

We use similar notations for functions defined on S. A simple and well known computation shows that the function $f \mapsto P(\ell \bullet f)$ attains its minimum at f_* defined by

$$f_*(x) = \log \frac{1 + \eta(x)}{1 - \eta(x)}.$$

We will assume in what follows that $f_* \in \mathcal{H}$. This assumption is rather restrictive. Since functions in $\mathcal{H}$ are uniformly bounded (see (1)) it means, in particular, that

η is bounded away from both $+1$ and -1. Although, there is a version of the main result below without this assumption, we are not discussing it in this note.

Next we need an assumption on so called uniform L_2-entropy of the unit ball in $\mathcal{H}$,

$$B_{\mathcal{H}} := \{f \in \mathcal{H} : \|f\| \le 1\}.$$

Given a probability measure Q on S, let $N\Big(B_{\mathcal{H}}; L_2(Q); \varepsilon\Big)$ denote the minimal number of $L_2(Q)$-balls needed to cover $B_{\mathcal{H}}$. Suppose that for some $\rho \in (0,2)$ and for some constant $A > 0$

$$\forall Q \ \forall \varepsilon > 0 : \ \log N\Big(B_{\mathcal{H}}; L_2(Q); \varepsilon\Big) \le \left(\frac{A}{\varepsilon}\right)^{\rho}. \tag{3}$$

Denote $B(x,\delta)$ the open ball in (S,d) with center x and radius δ. Also, let $\mathcal{H}(x,\delta)$ be the set of all functions $h \in \mathcal{H}$ satisfying the following conditions:

$$\begin{aligned}
&(i) \ \ \forall y \in S \ 0 \le h(y) \le 2\delta \\
&(ii) \ \ h \ge \delta \text{ on } B(x;\delta/2) \\
&(iii) \int_{B(x;\delta)^c} h d\Pi \le \delta \int_S h d\Pi
\end{aligned}$$

It follows from $(i) - (iii)$ that

$$\delta \Pi(B(x;\delta/2)) \le \mathbb{E}\, h(X) \le \frac{2\delta}{1-\delta} \Pi(B(x;\delta)).$$

Since there exists a continuous function h such that $0 \le h \le \frac{3}{2}\delta$, $h \ge \frac{4}{3}\delta$ on $B(x,\delta/2)$ and $h = 0$ on $B(x,\delta)^c$, and, on the other hand, $\mathcal{H}$ is dense in $C(S)$, it is easy to see that $\mathcal{H}(x,\delta) \ne \emptyset$. Denote

$$q(x,\delta) := \inf_{h \in \mathcal{H}(x,\delta)} \|h\|.$$

The quantity $q(x,\delta)$ is, often, bounded from above uniformly in $x \in S$ by a decreasing function of δ, say by $\bar{q}(\delta)$, and this will be assumed in what follows. Often, $\bar{q}(\delta)$ grows as $\delta^{-\gamma}, \ \delta \to 0$ for some $\gamma > 0$.

Example. For instance, if $\mathcal{H} = \mathcal{H}^s(S)$ is a Sobolev space of functions in a compact domain $S \subset \mathbb{R}^d$, $s > d/2 + 1$, define

$$h(y) := \delta \varphi\left(\frac{x-y}{\delta}\right),$$

where $\varphi \in C^\infty(\mathbb{R}^d)$, $0 \le \varphi \le 2$, $\varphi(x) \ge 1$ if $|x| \le 1/2$ and $\varphi(x) = 0$ if $|x| \ge 1$. Then h satisfies conditions (i)–(iii) (moreover, $h = 0$ on $B(x,\delta)^c$). A straightforward computation of Sobolev's norm of h shows that

$$\|h\|_{\mathcal{H}^s(S)} \le C\delta^{1+d/2-s},$$

implying that $q(x,\delta)$ is uniformly bounded from above by $\bar{q}(\delta) = C\delta^{-\gamma}$ with $\gamma = s - \frac{d}{2} - 1$. Similar results are also true in the case of RKHS for some kernels.

Let

$$p(x,\delta) := \delta^2 \Pi(B(x,\delta/2)).$$

In what follows, $K, C > 0$ will denote sufficiently large numerical constants (whose precise values might change from place to place). Recall our assumption that $\delta(P) > 0$. In this case it is also natural to assume that for all $\delta \le \frac{\delta(P)}{K}$ and for all x such that $|\eta(x)| \ge \delta(P)$

$$p(x,\delta) \ge \bar{p}(\delta) > 0$$

for some fixed function $\bar{p}$. This would be true, for instance, if S is a domain in $\mathbb{R}^d$ and Π has density uniformly bounded away from 0 on the set $\{x : |\eta(x)| \ge \delta(P)\}$. In this case we have for all x from this set

$$p(x,\delta) \ge c\delta^{d+2} =: \bar{p}(\delta).$$

Define now

$$r(x,\delta) := \frac{p(x,\delta)}{q(x,\delta)}.$$

Then on the set $\{x : |\eta(x)| \ge \delta(P)\}$

$$r(x,\delta) \ge \frac{\bar{p}(\delta)}{\bar{q}(\delta)}.$$

We set $U := K\Big(\|f_*\| \vee L \vee 1\Big)$ (here and in what follows $\vee$ stands for the maximum and $\wedge$ for the minimum) and define

$$\lambda^+ = \lambda^+(P) := \frac{1}{4U} \inf\left\{ r\left(x; \frac{\delta(P)}{U}\right) : |\eta(x)| \ge \delta(P) \right\}$$

and, for a fixed $\varepsilon > K\frac{\log\log n}{n}$,

$$\lambda^- := \frac{A^{2\rho/(2+\rho)}}{n^{2/(2+\rho)}} \bigvee \varepsilon.$$

Clearly,

$$\lambda^+ \ge \frac{1}{4U} \frac{\bar{p}(\delta(P)/U)}{\bar{q}(\delta(P)/U)} > 0,$$

so, λ^+ is a positive constant. Then if n is large enough and ε is not too large, we have $\lambda^- \le \lambda^+$.

Now, we are ready to formulate the main result.

Theorem 1. *Let $\lambda \in [\lambda^-, \lambda^+]$. Then there exists $\beta = \beta(\mathcal{H}, P) > 0$ such that*

$$\mathbb{E}\left(R_P(\hat{f}_n) - R_*\right) \le \exp\{-\beta n\}.$$

In fact, with sufficiently large $K, C > 0$, β is equal to $C^{-1}\left(\bar{p}\left(\frac{\delta(P)}{U}\right) \bigwedge \varepsilon\right)$, which is positive and does not depend on n, establishing the exponential convergence rate.

3 Proof

We use a well known representation of the excess risk

$$R_P(f) - R_* = \int_{\{\mathrm{sign}(f)\neq\mathrm{sign}(\eta)\}} |\eta| d\Pi$$

to get the following bound:

$$\begin{aligned}
&\mathbb{E}\,(R_P(\hat{f}_n) - R_*) \leq \\
&\mathbb{E}\int_{\{\hat{f}_n(x)\eta(x)\leq 0\}} |\eta(x)|\Pi(dx) = \mathbb{E}\int |\eta(x)| I_{\{\hat{f}_n(x)\eta(x)\leq 0\}}\Pi(dx) = \\
&\int |\eta(x)|\,\mathbb{E}\, I_{\{\hat{f}_n(x)\eta(x)\leq 0\}}\Pi(dx) = \int |\eta(x)|\,\mathbb{P}\,\{\hat{f}_n(x)\eta(x)\leq 0\}\Pi(dx) \qquad (4)
\end{aligned}$$

Our goal now is to bound, for a given x, $\mathbb{P}\,\{\hat{f}_n(x)\eta(x)\leq 0\}$. Let us assume that $\eta(x) = \delta > 0$ (the other case, when $\eta(x) < 0$, is similar). We have

$$\begin{aligned}
&\mathbb{P}\,\{\hat{f}_n(x)\eta(x)\leq 0\} = \mathbb{P}\,\{\hat{f}_n(x)\leq 0\} \leq \\
&\mathbb{P}\,\{\hat{f}_n(x)\leq 0,\ \|\hat{f}_n\|\leq U\} + \mathbb{P}\,\{\|\hat{f}_n\| > U\}. \qquad (5)
\end{aligned}$$

We start with bounding the first term. For $\delta_0 > 0$ (to be chosen later), let $h \in \mathcal{H}(x, \delta_0)$. Define

$$L_n(\alpha) := P_n(\ell \bullet (\hat{f}_n + \alpha h)) + \lambda\|\hat{f}_n + \alpha h\|^2.$$

Since $\hat{f}_n$ minimizes the functional

$$\mathcal{H} \ni f \mapsto P_n(\ell \bullet f) + \lambda\|f\|^2,$$

the function $\alpha \mapsto L_n(\alpha)$ attains its minimum at $\alpha = 0$. This function is differentiable, implying that

$$\frac{dL_n}{d\alpha}\Big(0\Big) = \frac{1}{n}\sum_{j=1}^{n} \ell'(Y_j\hat{f}_n(X_j))Y_j h(X_j) + 2\lambda\langle\hat{f}_n, h\rangle = 0.$$

Assuming that $\eta(x) = \delta > 0$, $\|\hat{f}_n\| \leq U$ and $\hat{f}_n(x) \leq 0$, we need to bound from above

$$\frac{1}{n}\sum_{j=1}^{n} \ell'(Y_j\hat{f}_n(X_j))Y_j h(X_j) + 2\lambda\langle\hat{f}_n, h\rangle,$$

trying to show that everywhere except the event of small probability the last expression is strictly negative. This would contradict the fact that it is equal to 0, implying a bound on the probability of the event $\{\hat{f}_n(x)\leq 0, \|\hat{f}_n\|\leq U\}$.

First note that

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n} \ell'(Y_j\hat{f}_n(X_j))Y_j h(X_j) \\
&= \frac{1}{n}\sum_{j:Y_j=+1} \ell'(\hat{f}_n(X_j))h(X_j) - \frac{1}{n}\sum_{j:Y_j=-1} \ell'(-\hat{f}_n(X_j))h(X_j).
\end{aligned}$$

Note also that function ℓ' is negative and increasing, h is nonnegative and $\hat{f}_n$ is a Lipschitz function with Lipschitz norm bounded by $\|\hat{f}_n\|$. The last observation and the assumption that $\hat{f}_n(x) \leq 0$ imply that, for all $y \in B(x, \delta_0)$,

$$\hat{f}_n(y) \leq \|\hat{f}_n\|\delta_0 \leq U\delta_0$$

and, as a result,

$$\ell'(\hat{f}_n(y)) \leq \ell'(U\delta_0),\ \ell'(-\hat{f}_n(y)) \geq \ell'(-U\delta_0).$$

Also, for all $y \in S$, $|\hat{f}_n(y)| \leq \|\hat{f}_n\| \leq U$, implying that

$$|\ell'(\hat{f}_n(y))| \leq |\ell'(-U)|,\ |\ell'(-\hat{f}_n(y))| \leq |\ell'(-U)|.$$

This leads to the following upper bound:

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n} \ell'(Y_j\hat{f}_n(X_j))Y_j h(X_j) \leq \\
&\frac{\ell'(U\delta_0)}{n} \sum_{j:X_j\in B(x,\delta_0),Y_j=+1} h(X_j) - \frac{\ell'(-U\delta_0)}{n} \sum_{j:X_j\in B(x,\delta_0),Y_j=-1} h(X_j) + \\
&\frac{|\ell'(-U)|}{n} \sum_{j:X_j\in B(x,\delta_0)^c} h(X_j) = \\
&\frac{\ell'(U\delta_0)}{n} \sum_{j:X_j\in B(x,\delta_0)} \frac{1+Y_j}{2} h(X_j) - \frac{\ell'(-U\delta_0)}{n} \sum_{j:X_j\in B(x,\delta_0)} \frac{1-Y_j}{2} h(X_j) + \\
&\frac{|\ell'(-U)|}{n} \sum_{j:X_j\in B(x,\delta_0)^c} h(X_j) = \\
&\frac{\ell'(U\delta_0) - \ell'(-U\delta_0)}{2n} \sum_{j=1}^{n} h(X_j) I_{B(x,\delta_0)}(X_j) + \\
&\frac{\ell'(U\delta_0) + \ell'(-U\delta_0)}{2n} \sum_{j=1}^{n} Y_j h(X_j) I_{B(x,\delta_0)}(X_j) + \\
&\frac{|\ell'(-U)|}{n} \sum_{j=1}^{n} h(X_j) I_{B(x,\delta_0)^c}(X_j).
\end{aligned}$$

Using the fact that for logit loss ℓ'' has its maximum at 0, we get

$$\begin{aligned}
&\left|\frac{\ell'(U\delta_0) + \ell'(-U\delta_0)}{2} - \ell'(0)\right| \leq \\
&\frac{|\ell'(U\delta_0) - \ell'(0)|}{2} + \frac{|\ell'(-U\delta_0) - \ell'(0)|}{2} \leq \ell''(0)U\delta_0
\end{aligned}$$

and

$$\left|\frac{\ell'(U\delta_0) - \ell'(-U\delta_0)}{2}\right| \leq \ell''(0)U\delta_0.$$

Therefore,

$$\frac{1}{n}\sum_{j=1}^{n}\ell'(Y_j\hat{f}_n(X_j))Y_jh(X_j)\le$$
$$\ell'(0)\frac{1}{n}\sum_{j=1}^{n}Y_jh(X_j)I_{B(x;\delta_0)}(X_j)+2\ell''(0)U\delta_0\frac{1}{n}\sum_{j=1}^{n}h(X_j)I_{B(x;\delta_0)}(X_j)+$$
$$\frac{|l'(-U)|}{n}\sum_{j=1}^{n}h(X_j)I_{B(x;\delta_0)^c}(X_j)=$$
$$\frac{1}{n}\sum_{j=1}^{n}\xi_j, \tag{6}$$

where $\xi,\xi_j,j\ge 1$ are i.i.d.

$$\xi:=\ell'(0)Yh(X)I_{B(x,\delta_0)}(X)+$$
$$2\ell''(0)U\delta_0h(X)I_{B(x,\delta_0)}(X)+|\ell'(-U)|h(X)I_{B(x;\delta_0)^c}(X).$$

To bound the sum of ξ_js, we will use Bernstein inequality. To this end, we first bound the expectation and the variance of ξ. We have

$$\mathbb{E}\,\xi=\ell'(0)\,\mathbb{E}\,Yh(X)I_{B(x;\delta_0)}(X)+2\ell''(0)U\delta_0\,\mathbb{E}\,h(X)I_{B(x;\delta_0)}(X)$$
$$+|\ell'(-U)|\,\mathbb{E}\,h(X)I_{B(x;\delta_0)^c}(X).$$

Since η is Lipschitz with the Lipschitz constant L and $\eta(x)=\delta$,

$$\eta(y)\ge\delta-L\delta_0$$

for all $y\in B(x;\delta_0)$. Since also $h\in\mathcal{H}(x,\delta_0)$, we have:

$$\mathbb{E}\,Yh(X)I_{B(x;\delta_0)}(X)=\mathbb{E}\,\eta(X)h(X)I_{B(x;\delta_0)}(X)$$
$$\ge(\delta-L\delta_0)\,\mathbb{E}\,h(X)I_{B(x;\delta_0)}(X)\ge(\delta-L\delta_0)(1-\delta_0)\,\mathbb{E}\,h(X),$$

$$\mathbb{E}\,h(X)I_{B(x;\delta_0)}(X)\le\mathbb{E}\,h(X),$$

and

$$\mathbb{E}\,h(X)I_{B(x;\delta_0)^c}(X)\le\delta_0\,\mathbb{E}\,h(X)$$

Recall that $\ell'(0)<0$ and $\ell''(0)\ge 0$. So, the following bound for the expectation of ξ is immediate:

$$\mathbb{E}\,\xi\le\Big[\ell'(0)(\delta-L\delta_0)(1-\delta_0)+2\ell''(0)U\delta_0+|\ell'(-U)|\delta_0\Big]\,\mathbb{E}\,h(X).$$

We will choose δ_0 small enough to make

$$[\ell'(0)(\delta-L\delta_0)(1-\delta_0)+2\ell''(U\delta_0)U\delta_0+|\ell'(-U)|\delta_0]\le-\delta_0.$$

A simple computation shows that it is enough to take

$$\delta_0 = \frac{1}{C}\frac{\delta}{U \vee L} \le \frac{\delta}{L + 4U + 12},$$

which can be always achieved by making the numerical constant C large enough. Then the expectation satisfies the bound

$$\mathbb{E}\,\xi \le -\delta_0\,\mathbb{E}\,h(X).$$

As far as the variance of ξ is concerned, using an elementary bound $(a+b+c)^2 \le 3a^2 + 3b^2 + 3c^2$, it is easy to check that

$$\mathrm{Var}(\xi) \le C\delta_0\,\mathbb{E}\,h(X)$$

with a sufficiently large numerical constant C. Finally, it is also straightforward that with some $C > 0$ $|\xi| \le C\delta_0$.

Now Bernstein inequality easily yields with a sufficiently large numerical constant $C > 0$

$$\mathbb{P}\left\{\frac{1}{n}\sum \xi_j \ge -\frac{1}{2}\delta_0\,\mathbb{E}\,h(X)\right\} \le 2\exp\left\{-\frac{n\delta_0\,\mathbb{E}\,h(X)}{C}\right\}.$$

Then, since

$$\delta_0\,\mathbb{E}\,h(X) \ge \delta_0^2 \Pi(B(x;\delta_0/2)) = p(x,\delta_0),$$

we have with probability at least $1 - 2\exp\left\{-\frac{np(x,\delta_0)}{C}\right\}$:

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} \ell'(Y_j \hat{f}_n(X_j)) Y_j h(X_j) + 2\lambda\langle \hat{f}_n, h\rangle \le \\
\le -\frac{1}{2}\delta_0\,\mathbb{E}\,h(X) + 2\lambda\langle \hat{f}_n, h\rangle \le \\
\le -\frac{1}{2}\delta_0\,\mathbb{E}\,h(X) + 2\lambda U\|h\| \le \\
\le -\frac{1}{2}p(x,\delta_0) + 2\lambda U q(x,\delta_0) \qquad (7)
\end{aligned}$$

So, if

$$\lambda < \frac{p(x,\delta_0)}{4Uq(x,\delta_0)} = \frac{r(x,\delta_0)}{4U},$$

then

$$\frac{1}{n}\sum_{j=1}^{n} \ell'(Y_j \hat{f}_n(X_j)) Y_j h(X_j) + 2\lambda\langle \hat{f}_n, h\rangle < 0$$

with probability at least $1 - 2\exp\left\{-\frac{np(x,\delta_0)}{C}\right\}$. The conclusion is that if $\eta(x) = \delta$ and $\lambda < \frac{r(x,\delta_0)}{4U}$, then

$$\mathbb{P}\left\{\hat{f}_n(x) \le 0, \|\hat{f}_n\| \le U\right\} \le 2\exp\left\{-\frac{np(x,\delta_0)}{C}\right\}.$$

Thus, for $\lambda \le \lambda^+$, we have

$$\mathbb{P}\{\hat{f}_n(x) \le 0, \|\hat{f}_n\| \le U\} \le 2\exp\left\{-\frac{n\bar{p}(\delta_0)}{C}\right\}. \tag{8}$$

We now turn to bounding the probability $\mathbb{P}\{\|\hat{f}_n\| \ge U\}$ for a properly chosen U. This is the only part of the proof where the condition (3) on the uniform entropy of the unit ball $B_{\mathcal{H}}$ is needed. It relies heavily on recent excess risk bounds in Koltchinskii [6] as well as on some of the results in spirit of Blanchard, Lugosi and Vayatis [5] (see their Lemma 4). We formulate the bound we need in the following lemma.

Lemma 1. *Suppose that condition (3) holds and (for simplicity) that ℓ is the logit loss. Let $R \ge 1$. Then, there exists a constant $K > 0$ such that for any $t > 0$, the following event*

$$\forall f \in \mathcal{H} \text{ with } \|f\| \le R \tag{9}$$

$$P_n(\ell \bullet f) - \inf_{\|g\| \le R} P_n(\ell \bullet g) \le$$
$$2\Big(P(\ell \bullet f) - \inf_{\|g\| \le R} P(\ell \bullet g)\Big) + K\left(\frac{RA^{2\rho/(2+\rho)}}{n^{2/(2+\rho)}} + \frac{tR}{n}\right), \tag{10}$$

has probability at least $1 - e^{-t}$.

The argument that follows will provide a bound that is somewhat akin to some of the bounds in [7] and in [4].

Denote $E(R)$ the event of the lemma. Let $R \ge \|f_*\| \vee 1$. On the event $E(R)$, the condition $R/2 < \|\hat{f}_n\| \le R$ implies

$$\lambda\|\hat{f}_n\|^2 \le P_n(\ell \bullet \hat{f}_n) - \inf_{\|g\| \le R} P_n(\ell \bullet g) + \lambda\|\hat{f}_n\|^2 =$$
$$\inf_{\|f\| \le R}\left[P_n(\ell \bullet f) - \inf_{\|g\| \le R} P_n(\ell \bullet g) + \lambda\|f\|^2\right] \le$$
$$2\inf_{\|f\| \le R}\left[P(\ell \bullet f) - \inf_{\|g\| \le R} P(\ell \bullet g) + \lambda\|f\|^2 + K\left(\frac{RA^{2\rho/(2+\rho)}}{n^{2/(2+\rho)}} + \frac{tR}{n}\right)\right] \le$$
$$2\left[P(\ell \bullet f_*) - \inf_{\|g\| \le R} P(\ell \bullet g) + \lambda\|f_*\|^2\right] + 2K\left(\frac{RA^{2\rho/(2+\rho)}}{n^{2/(2+\rho)}} + \frac{tR}{n}\right) \le$$
$$2\lambda\|f_*\|^2 + 2K\left(\frac{RA^{2\rho/(2+\rho)}}{n^{2/(2+\rho)}} + \frac{tR}{n}\right),$$

which implies that

$$\frac{R^2}{4} \le \|\hat{f}_n\|^2 \le 2\|f_*\|^2 + 2K\left(\frac{RA^{2\rho/(2+\rho)}}{\lambda n^{2/(2+\rho)}} + \frac{tR}{\lambda n}\right).$$

Solving this inequality with respect to R shows that on $E(R)$ the condition $R/2 \le \|\hat{f}_n\| \le R$ implies

$$R \le K\left(\|f_*\| \bigvee 1 \bigvee \frac{A^{2\rho/(2+\rho)}}{\lambda n^{2/(2+\rho)}} \bigvee \frac{t}{\lambda n}\right).$$

If now $t = n\varepsilon$ and $\lambda \ge \lambda^-$, then it yields

$$R \le K(\|f_*\| \vee 1).$$

Note that

$$P_n(\ell \bullet \hat{f}_n) + \lambda \|\hat{f}_n\|^2 \le \ell(0)$$

(just plug in $f = 0$ in the target functional). Therefore, we have $\lambda \|\hat{f}_n\|^2 \le \ell(0)$, or

$$\|\hat{f}_n\| \le \sqrt{\frac{\ell(0)}{\lambda}} =: \bar{R}.$$

Define $R_k = 2^k,\ k = 0, 1, 2, \dots, N := \log_2 \bar{R} + 1$. Note that, for our choice of λ, we have $N \le C \log n$ with some numerical constant $C > 0$. Let $E_k := E(R_k)$. Clearly, $\mathbb{P}(E_k) \ge 1 - e^{-t}$ and, on the even E_k, the condition $R_{k-1} \le \|\hat{f}_n\| \le R_k$ implies

$$\|\hat{f}_n\| \le R_k \le K(\|f_*\| \vee 1).$$

Thus, $\|\hat{f}_n\|$ can be larger than the right hand side of the last bound only on the event $\bigcup_{k=1}^N E_k^c$, whose probabilty is smaller than $Ne^{-n\varepsilon}$. This establishes the following inequality:

$$\mathbb{P}\left\{\|\hat{f}_n\| \ge K(\|f_*\| \vee 1)\right\} \le Ne^{-n\varepsilon} \le e^{-n\varepsilon/2}, \tag{11}$$

provided that $\varepsilon \ge K\frac{\log\log n}{n}$, as it was assumed.

Combining bounds (8) and (11) and plugging the resulting bound in (5) and then in (4) easily completes the proof (subject to a minor adjustment of the constants).

Acknowledgement. The first author is very thankful to Alexandre Tsybakov for several useful and interesting conversations on the subject of the paper.

References

1. Audibert, J.–Y. and Tsybakov, A. Fast convergence rates for plug-in estimators under margin conditions. *Unpublished manuscript*, 2004.
2. Bartlett, P., Bousquet, O. and Mendelson, S. Local Rademacher Complexities. *Annals of Statistics*, 2005, to appear.
3. Bartlett, P., Jordan, M. and McAuliffe, J. Convexity, Classification and Risk Bounds. *J. American Statistical Soc.*, 2004, to appear.

4. Blanchard, G., Bousquet, O. and Massart, P. Statistical Performance of Support Vector Machines. *Preprint*, 2003, 4, 861-894.
5. Blanchard, G., Lugosi, G. and Vayatis, N. On the rates of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 2003, 4, 861-894.
6. Koltchinskii, V. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. Preprint. *Preprint*, 2003.
7. Scovel, C. and Steinwart, I. Fast Rates for Support Vector Machines. *Preprint*, 2003.
to