An inequality for uniform deviations of sample averages from their means

Peter Bartlett
Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra 0200, Australia.
email: Peter.Bartlett@anu.edu.au.

Gábor Lugosi†
Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain.
email: lugosi@upf.es.

February 28, 1998

Key words: Vapnik-Chervonenkis inequality, uniform laws of large numbers, empirical risk minimization.

Classification: C13.

---

*This work was supported by the Australian Research Council. The work of the second author was supported by DGES grant PB96-0300.
Abstract

We derive a new inequality for uniform deviations of averages from their means. The inequality is a common generalization of previous results of Vapnik and Chervonenkis (1974) and Pollard (1986). Using the new inequality we obtain tight bounds for empirical loss minimization learning.
1 Introduction

Let $X_1^n = (X_1, \ldots, X_n)$ be a sequence of independent, identically distributed random variables taking their values from some set $\mathcal{X}$, and consider a class $\mathcal{F}$ of uniformly bounded functions $f : \mathcal{X} \to [0,1]$. We are interested in the maximal difference between the sample average

$$P_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

and the mean $P(f) = \mathbb{E}f(X_1)$ over $f \in \mathcal{F}$. Several upper bounds have been established for quantities of the above type, all of them involve some quantity to measure the size of the class $\mathcal{F}$. We work with covering numbers defined as follows: let $x_1, \ldots, x_n \in \mathcal{X}$, and consider the distance $d_\infty$ between two functions

$$d_\infty(f,g) = \max_{i \leq n} |f(x_i) - g(x_i)|.$$

A finite set of functions $\{g_1, \ldots, g_N\}$ is called an $\epsilon$-cover of $\mathcal{F}$ (with respect to the distance $d_\infty$) if for every $f \in \mathcal{F}$ there exists a $g_i$, $i \leq N$, such that $d_\infty(f,g) < \epsilon$. Let $N_\infty(\mathcal{F}, x^n_1, \epsilon)$ denote the smallest integer $N$ for which such a covering exists. The covering number $N_1(\mathcal{F}, x^n_1, \epsilon)$ is defined similarly but with the distance $d_1(f,g) = (1/n) \sum_{i \leq n} |f(x_i) - g(x_i)|$ replacing $d_\infty$.

The first, now classical, work of Vapnik and Chervonenkis (1971) concentrated on the special case when all functions in $\mathcal{F}$ take only two values: zero and one. For that case they showed that

$$\mathbb{P} \left[ \sup_{f \in \mathcal{F}} |P(f) - P_n(f)| > \epsilon \right] \leq 4E \log N_1 \left( \mathcal{F}, X^{2n}_1, \frac{1}{2n} \right) e^{-\epsilon^2/8}. $$

This inequality fails to capture the phenomenon that for those $f \in \mathcal{F}$ for which $P(f)$ is small, the deviation $|P(f) - P_n(f)|$ is also small with large probability. Still for the case of binary-valued functions, Vapnik and Chervonenkis (1974) improved the previous inequality to

$$\mathbb{P} \left[ \sup_{f \in \mathcal{F}} \frac{|P(f) - P_n(f)| \sqrt{P(f)}}{\sqrt{P(f)}} > \epsilon \right] \leq 4E \log N_1 \left( \mathcal{F}, X^{2n}_1, \frac{1}{2n} \right) e^{-\epsilon^2/4}. \quad (1)$$

(The present constants were achieved by Anthony and Shawe-Taylor (1993).)

Several inequalities have been proved also for the general case of uniformly bounded classes of functions. Haussler (1992), improving on an earlier result by Pollard (1986), proved the following useful inequality: if $\nu > 0$ and $\beta \in (0,1)$, then

$$\mathbb{P} \left[ \sup_{f \in \mathcal{F}} \frac{|P(f) - P_n(f)|}{P(f) + P_n(f) + \nu} > \beta \right] \leq 4E \log N_1 \left( \mathcal{F}, X^{2n}_1, \frac{\beta \nu}{8} \right) e^{-\nu \beta^2 / 8}. \quad (2)$$
Even though this inequality is useful to bound the probabilities of large relative uniform deviations, when specialized to binary-valued functions, it is somewhat weaker than (1). In this paper we prove an inequality which is a common generalization of both (1) and (2). In Section 3 we apply the new inequality for a general learning problem. For more inequalities on probabilities of uniform deviations, see, for example, Alexander (1984), Devroye (1982), Pollard (1984), Talagrand (1994), and Vapnik (1982).

2 The inequality

**Theorem 1** If \( n \geq 1/\gamma^2 \), then

\[
\Pr \left[ \sup_{f \in \mathcal{F}} \frac{P(f) - P_n(f) - \gamma}{\sqrt{P(f)}} > \epsilon \right] \leq 2E N_\infty (\mathcal{F}, X_1^{2n}, \gamma \frac{\gamma}{4}) e^{-n \epsilon^2 / 4}
\]

and

\[
\Pr \left[ \sup_{f \in \mathcal{F}} \frac{P_n(f) - P(f) - \gamma}{\sqrt{P_n(f)}} > \epsilon \right] \leq 2E N_\infty (\mathcal{F}, X_1^{2n}, \gamma \frac{\gamma}{4}) e^{-n \epsilon^2 / 4}.
\]

**Proof.** We start with the first inequality. Our proof uses some ideas from the beautiful short proof of (1) by Anthony and Shawe-Taylor (1993).

**Step 1.** Let \( X_1', \ldots, X_n' \) be auxiliary i.i.d. random variables, having the same distribution at that of the \( X_i \) and independent of them. Denote \( P_n'(f) = (1/n) \sum_{i=1}^n f(X_i') \). If \( n \geq 1/\gamma^2 \), then

\[
\Pr \left[ \sup_{f \in \mathcal{F}} \frac{P(f) - P_n(f) - \gamma}{\sqrt{P(f)}} > \epsilon \right] \leq 2 \Pr \left[ \sup_{f \in \mathcal{F}} \frac{P_n'(f) - P_n(f) - \gamma}{\sqrt{1/n (P_n'(f) + P_n(f))}} > \epsilon \right].
\]

**PROOF:** Let \( f \) satisfy \( P(f) - P_n(f) - \gamma > \epsilon \sqrt{P(f)} \). If \( P_n'(f) \geq P(f) - \gamma/2 \), then since \((x - a)/\sqrt{x + a}\) is a monotone increasing function in \( x > 0 \) (when \( a \geq 0 \)), we have that

\[
\frac{P_n'(f) - P_n(f) - \gamma}{\sqrt{1/n (P_n'(f) + P_n(f))}} \geq \frac{P(f) - P_n(f) - \gamma}{\sqrt{1/n (P(f) + P_n(f) - \gamma/4)}} \geq \frac{P(f) - P_n(f) - \gamma}{\sqrt{P(f)}}.
\]

But by the Chebyshev-Cantelli inequality (see, e.g., Chow and Teicher, 1978), since \( \operatorname{var}(f(X)) \leq 1/4 \), if \( n \geq 1/\gamma^2 \),

\[
\Pr \left[ P_n'(f) \leq P(f) - \gamma/2 \right] \leq \frac{1}{n} \frac{\operatorname{var}(f(X))}{\sqrt{1/n (P_n'(f) + P_n(f) - \gamma/4)}} \leq \frac{1}{2},
\]

which completes the proof of the first step.
STEP 2. Let $\delta > 0$, and define $F_\delta$ to be a minimal $\delta$-cover of $F$ with respect to the metric $d_\infty$ defined on the points $X_1, \ldots, X_n, X'_1, \ldots, X'_n$. Define

$$F_\delta = \{ g = f - \delta : f \in F_\delta \}.$$ 

For $\delta = \gamma/4$,

$$P \left[ \sup_{f \in F_\delta} \frac{P'_n(f) - P_n(f) - \frac{\gamma}{2}}{\sqrt{\frac{1}{2} (P'_n(f) + P_n(f))}} > \epsilon \right] \leq P \left[ \max_{g \in F_\delta} \frac{P'_n(g) - P_n(g)}{\sqrt{\frac{1}{2} (P'_n(g) + P_n(g))}} > \frac{\epsilon}{\sqrt{2}} \right].$$

PROOF: Assume that $f$ satisfies

$$\frac{P'_n(f) - P_n(f) - \frac{\gamma}{2}}{\sqrt{\frac{1}{2} (P'_n(f) + P_n(f))}} > \epsilon.$$

For each $f \in F_\delta$, there exists a $g \in F_\delta$ such that $g(x) \leq f(x) \leq g(x) + 2\delta$ for every $x = X_i$ and $x = X'_i$. Thus,

$$\frac{P'_n(f) - P_n(f) - \frac{\gamma}{2}}{\sqrt{\frac{1}{2} (P'_n(f) + P_n(f))}} \leq \frac{P'_n(g) + 2\delta - P_n(g) - \frac{\gamma}{2}}{\sqrt{\frac{1}{2} (P'_n(g) + P_n(g))}} = \frac{P'_n(g) - P_n(g)}{\sqrt{\frac{1}{2} (P'_n(g) + P_n(g))}}.$$

This completes the proof of Step 2.

STEP 3. Let $\sigma_1, \ldots, \sigma_n$ be a sequence of i.i.d. random variables with $P[\sigma_1 = -1] = P[\sigma_1 = 1] = 1/2$. Clearly,

$$P \left[ \max_{g \in F_\delta} \frac{P'_n(g) - P_n(g)}{\sqrt{P'_n(g) + P_n(g)}} > \frac{\epsilon}{\sqrt{2}} \right] = P \left[ \max_{g \in F_\delta} \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i (g(X_i) - g(X'_i))}{\sqrt{P'_n(g) + P_n(g)}} > \frac{\epsilon}{\sqrt{2}} \right]$$

$$= E \left( P \left[ \max_{g \in F_\delta} \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i (g(X_i) - g(X'_i))}{\sqrt{P'_n(g) + P_n(g)}} > \frac{\epsilon}{\sqrt{2}} \right| X_1^{2n} \right).$$

Now, we may bound the conditional probability above by the union bound and Hoeffding’s inequality (Hoeffding, 1963):

$$P \left[ \max_{g \in F_\delta} \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i (g(X_i) - g(X'_i))}{\sqrt{P'_n(g) + P_n(g)}} > \frac{\epsilon}{\sqrt{2}} \right| X_1^{2n} \right]$$

$$\leq |F_\delta| \max_{g \in F_\delta} P \left[ \frac{\frac{1}{n} \sum_{i=1}^n \sigma_i (g(X_i) - g(X'_i))}{\sqrt{P'_n(g) + P_n(g)}} > \frac{\epsilon}{\sqrt{2}} \right| X_1^{2n} \right]$$

$$\leq |F_\delta| e^{-n\epsilon^2/4}. $$
This completes the proof of the first inequality. To prove the second inequality, only the first step has to be modified:

$$P \left[ \sup_{f \in \mathcal{F}} \frac{P_n(f) - P(f) - \gamma}{\sqrt{P(f)}} > \epsilon \right] \leq 2P \left[ \sup_{f \in \mathcal{F}} \frac{P_n(f) - P_n'(f) - \frac{\gamma}{2}}{\sqrt{\frac{1}{2}(P_n'(f) + P_n(f))}} > \epsilon \right].$$

**Proof:** Assume that $P_n(f) - P(f) - \gamma > \epsilon \sqrt{P_n(f)}$. Then obviously $P_n(f) > P(f) + \gamma/2$. If $P_n'(f) \leq P(f) + \gamma/2$, then since $(a-x)/\sqrt{a+x}$ is a monotone decreasing function in $x > 0$ (when $a \geq 0$), we have that

$$\frac{P_n(f) - P_n'(f) - \frac{\gamma}{2}}{\sqrt{\frac{1}{2}(P_n'(f) + P_n(f))}} \geq \frac{P_n(f) - \left(P(f) + \frac{\gamma}{2}\right) - \frac{\gamma}{2}}{\sqrt{\frac{1}{2}(P(f) + P(f) + \frac{\gamma}{2})}} \geq \frac{P_n(f) - P(f) - \gamma}{\sqrt{P_n(f)}}.$$

Again, by the Chebyshev-Cantelli inequality,

$$P \left[ P_n(f) > P(f) + \gamma/2 \right] \leq \frac{\frac{1}{n} \text{var}(f(X))}{\frac{1}{n} \text{var}(f(X)) + \frac{\gamma}{4}} \leq \frac{1}{2^4}.$$

The rest of the proof of the second inequality is identical to that of the first one. \(\square\)

**Remark.** The condition $n \geq 1/\gamma^2$ may be relaxed somewhat. By a trivial modification of the proof we obtain, for any $\alpha > 0$, that if $n \geq 4\alpha/\gamma^2$, then

$$P \left[ \exists f \in \mathcal{F} : \frac{P(f) - P_n(f) - \gamma}{\sqrt{P(f)}} > \epsilon \right. \text{ and } \frac{\text{var}(f(X))}{\frac{1}{n} \text{var}(f(X)) + \frac{\gamma}{4}} \leq \alpha \left] \leq 2\mathbb{E} N_\infty \left( \mathcal{F}, \mathcal{X}_1^n, \frac{\gamma}{4} \right) \right] e^{-n\alpha^2/4},$$

and a similar version of the second inequality of Theorem 1 also remains valid. \(\square\)

**Remark.** Theorem 1 uses $d_\infty$ covering numbers, where (1) and (2) use $d_1$ covering numbers. It is easy to see that $N_1(\mathcal{F}, x_1^n, \epsilon) \leq N_\infty(\mathcal{F}, x_1^n, \epsilon)$. While these covering numbers can be very different for a particular probability distribution on $\mathcal{X}$, if we consider worst case distributions, they are closely related. To see this, we need to introduce another type of covering number. For a probability distribution $P$ on the set $\mathcal{X}$, let $d_P(f,g) = P|f-g|$, and let $N(\mathcal{F}, d_P, \epsilon)$ denote the size of the smallest $\epsilon$-cover of $\mathcal{F}$ with respect to $d_P$. Then there are constants $b_1, b_2, c_1, c_2$ such that, for $n \geq (b_1/\epsilon^2) \log N(\mathcal{F}, d_P, c_1\epsilon)$,

$$\log \max_P \mathbb{E}N_1(\mathcal{F}, X_1^n, \epsilon) \leq \log \max_{x_1^n} \mathbb{E}N_\infty(\mathcal{F}, x_1^n, \epsilon) \leq \log \max_{x_1^n} N_\infty(\mathcal{F}, x_1^n, \epsilon) \leq b_1 \log \max_P N(\mathcal{F}, d_P, c_1\epsilon) \log^2(n/\epsilon) \leq b_2 \log \max_P \mathbb{E}N_1(\mathcal{F}, X_1^n, c_2\epsilon) \log^2(n/\epsilon).$$
(The first two inequalities are trivial, the third follows from Lemma 1 and Theorem 1 of Bartlett, Kulkarni, and Posner (1997), and the last follows from an argument due to Haussler (1992) in the proof of his Lemma 4.) This shows that when we apply this result in the next section, the use of the $d_\infty$ (instead of $d_1$) covering numbers introduces no more than log factors into the bounds on the sample size for empirical loss minimization learning. □

Next we point out that using Theorem 1, we may recover an inequality like (2).

**Corollary 1** If $\beta \in (0, 1)$, $\nu > 0$, and $n \geq 4(1 + \beta)^2/\beta^2 \nu^2$, then

$$
P \left[ \sup_{f \in \mathcal{F}} \frac{|P(f) - P_n(f)|}{P(f) + P_n(f) + \nu} > \beta \right] \leq 2EN_\infty \left( \mathcal{F}, X_1^{2n}, \frac{\beta \nu}{8(1 + \beta)} \right) e^{-\nu \beta^2/(1 - \beta^2)}.
$$

**Proof.** We show that the first inequality of Theorem 1 implies

$$
P \left[ \sup_{f \in \mathcal{F}} \frac{P(f) - P_n(f)}{P(f) + P_n(f) + \nu} > \beta \right] \leq 2EN_\infty \left( \mathcal{F}, X_1^{2n}, \frac{\beta \nu}{8(1 + \beta)} \right) e^{-\nu \beta^2/(1 - \beta^2)}.
$$

The other side of the inequality follows similarly from the second inequality of Theorem 1.

If $f \in \mathcal{F}$ is such that $P(f) - P_n(f) - \gamma \leq \epsilon \sqrt{P(f)}$, then for any $\alpha > 0$ we have two cases:

1. If $P(f) < (1 + 1/\alpha)^2 \epsilon^2$ then $P(f) < P_n(f) + \gamma + (1 + \alpha) \epsilon^2/\alpha$.

2. If $P(f) \geq (1 + 1/\alpha)^2 \epsilon^2$, then $P(f) \leq P_n(f) + \gamma + \alpha P(f)/(1 + \alpha)$, and so $P(f) \leq (1 + \alpha) P_n(f) + (1 + \alpha) \gamma$.

In either case,

$$
P(f) \leq (1 + \alpha) P_n(f) + (1 + \alpha) (\gamma + \epsilon^2/\alpha).
$$

Therefore, the first inequality of Theorem 1 implies that

$$
P \left[ \exists f \in \mathcal{F} : P(f) > (1 + \alpha) P_n(f) + (1 + \alpha) (\gamma + \epsilon^2/\alpha) \right] \leq 2EN_\infty \left( \mathcal{F}, X_1^{2n}, \frac{\gamma}{4} \right) e^{-\nu^2/4}.
$$

By choosing $\alpha = 2 \beta/(1 - \beta)$, $\gamma = \nu \beta^2/(1 + \beta)$, and $\epsilon^2 = \nu \beta^2/(1 - \beta^2)$ we obtain

$$
P \left[ \exists f \in \mathcal{F} : P(f) > (1 + \alpha) P_n(f) + (1 + \alpha) (\gamma + \epsilon^2/\alpha) \right] = P \left[ \exists f \in \mathcal{F} : P(f) > \frac{1 + \beta}{1 - \beta} P_n(f) + \frac{\beta \nu}{1 - \beta} \right] = P \left[ \sup_{f \in \mathcal{F}} \frac{P(f) - P_n(f)}{P(f) + P_n(f) + \nu} > \beta \right],
$$

which proves the corollary. □
Even though the exponent in the upper bound of Corollary 1 is slightly better in most cases than that of (2), Corollary 1 is weaker than Pollard’s result. First, Corollary 1 has $d_{\infty}$ covering numbers instead of the smaller $d_1$ covering numbers, though this is a minor difference for worst case probability distributions, as we have pointed out above. However, more importantly, the condition $n \geq 4(1 + \beta)^2/\beta^2 \nu^2$ may be quite restrictive in some applications (though it might be relaxed somewhat according to the remark following the proof of Theorem 1 above). Nevertheless, in the next section we show that in a typical situation where (2) has been used, the new inequalities provide significantly stronger results.

3 Application for learning

In this section we apply Theorem 1 to obtain tighter upper bounds for the loss of a decision selected by empirical loss minimization from a class of decisions.

Let $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ be i.i.d. training data, and let $l(\cdot, \cdot)$ be a loss function taking its values in $[0, 1]$. Denote

$$L = \{l(f(\cdot), \cdot) : f \in \mathcal{F}\}.$$ 

The loss of any $f \in \mathcal{F}$ is

$$L(f) = \mathbb{E}[l(f(X), Y)].$$

Define the optimal loss in the class by $L^* = \inf_{f \in \mathcal{F}} L(f)$. Let $f_n$ be any function in $\mathcal{F}$ which minimizes the empirical risk

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} l(f(X_i), Y_i).$$

**Theorem 2** For every $\epsilon > 0$,

$$\mathbb{P}[L(f_n) - L^* > 2\epsilon] \leq 2\mathbb{E}N_{\infty}(L, D_{2n}, \frac{\epsilon}{\delta}) e^{-\frac{\epsilon^2}{8}} + e^{-\frac{\epsilon^2}{8\delta^2}}.$$

In particular, for every $\epsilon > 0$ and $\delta > 0$ we have $\mathbb{P}[L(f_n) - L^* > \epsilon] \leq \delta$ if

$$n \geq \max \left( \frac{16L^*}{\epsilon^2} \left( \log N \left( \frac{\epsilon}{16} \right) + \log \frac{4}{\delta} \right), \frac{16}{\epsilon} \left( \log N \left( \frac{\epsilon}{16} \right) + \log \frac{4}{\delta} \right) \right),$$

where $N(\epsilon) = \mathbb{E}N_{\infty}(L, D_{2n}, \epsilon)$.

**Proof.** If

$$\sup_{f \in \mathcal{F}} \frac{L(f) - L_n(f) - \frac{\epsilon}{7}}{\sqrt{L(f)}} \leq \frac{\epsilon}{\sqrt{L^* + 2\epsilon}},$$

then...
then for each \( f \in \mathcal{F} \)
\[
L_n(f) \geq L(f) - e \sqrt{\frac{L(f)}{L^* + 2 \epsilon}} - \frac{\epsilon}{2}.
\]

If, in addition, \( f \) is such that \( L(f) > L^* + 2 \epsilon \), then by the monotonicity of the function \( x - c \sqrt{x} \) (for \( c > 0 \) and \( x > c^2/4 \)),
\[
L_n(f) \geq L^* + 2 \epsilon - e \sqrt{\frac{L^* + 2 \epsilon}{L^* + 2 \epsilon}} - \frac{\epsilon}{2} = L^* + \frac{\epsilon}{2}.
\]

Therefore,
\[
P \left[ \inf_{f: L(f) > L^* + 2 \epsilon} L_n(f) < L^* + \frac{\epsilon}{2} \right] \leq P \left[ \sup_{f \in \mathcal{F}} \frac{L(f) - L_n(f) - \frac{\epsilon}{2}}{\sqrt{L(f)}} > \frac{\epsilon}{\sqrt{L^* + 2 \epsilon}} \right].
\]

But if \( L(f_n) - L^* > 2 \epsilon \), then there exists an \( f \in \mathcal{F} \) such that \( L(f) > L^* + 2 \epsilon \) and \( L_n(f) < L_n(f^*) \). Thus,
\[
P[L(f_n) - L^* > 2 \epsilon] \leq P \left[ \inf_{f: L(f) > L^* + 2 \epsilon} L_n(f) < L_n(f^*) \right] \\
\leq P \left[ \inf_{f: L(f) > L^* + 2 \epsilon} L_n(f) < L^* + \frac{\epsilon}{2} \right] + P \left[ L_n(f^*) > L^* + \frac{\epsilon}{2} \right] \\
\leq P \left[ \sup_{f \in \mathcal{F}} \frac{L(f) - L_n(f) - \frac{\epsilon}{2}}{\sqrt{L(f)}} > \frac{\epsilon}{\sqrt{L^* + 2 \epsilon}} \right] + P \left[ L_n(f^*) - L^* > \frac{\epsilon}{2} \right].
\]

Straightforward application of Theorem 1 and Bennett’s version (Bennett, 1962) of Bernstein’s inequality finishes the proof of the first inequality. The second statement is a straightforward consequence. \( \square \)

**Remark.** Results for \( \{0, 1\} \)-valued functions \( f \) with the discrete loss function \( l(y, y') = 0 \) if \( y = y' \) and 1 otherwise) show that, for some probability distributions, the convergence rate of Theorem 2 cannot be improved (see Devroye and Lugosi, 1995, Ehrenfeucht et al, 1989). \( \square \)

**References**


