The Glivenko Cantelli Theorem and its Generalizations

Thomas Kahle

August 21, 2006

In this note we will study upper bounds of random variables of the type

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)|,$$

where $\mathcal{A}$ is a class of sets that needs to fulfill certain assumptions. These bounds are important tools in the analysis of learning processes and probabilistic theories of pattern recognition. The presentation given here is based on [DGL96].

1 Hoeffdings inequality

Lemma 1 (Chebyshev inequality)**.** *Let $\epsilon > 0$ and $X \in \mathcal{L}^2$ then*

$$\mathbb{P}\{|X - \mathbb{E}X| \geq \epsilon\} \leq \frac{\mathbb{V}X}{\epsilon^2}.$$

Theorem 2 (Hoeffdings inequality)**.** *Let $X_1, \ldots, X_n$, be independent bounded random variables such that X_i fall in the interval $[a_i, b_i]$ with probability one. Denote their sum by $S_n = \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$ we have*

$$\mathbb{P}\{S_n - \mathbb{E}S_n \geq \epsilon\} \leq \mathrm{e}^{-2\epsilon^2/\sum_{i=1}^n (b_i-a_i)^2}$$

and

$$\mathbb{P}\{S_n - \mathbb{E}S_n \leq -\epsilon\} \leq \mathrm{e}^{-2\epsilon^2/\sum_{i=1}^n (b_i-a_i)^2}$$

Proof. Use convexity of the exponential function to prove that for a random variable X with $\mathbb{E}X = 0$ and $a \leq X \leq b$ for any $s > 0$ we have

$$\mathbb{E}\left\{\mathrm{e}^{sX}\right\} \leq \mathrm{e}^{s^2(b-a)^2/8}.$$

The proof is now based on *Chernoff's bounding method*: Use the Markov inequality to see that for a nonnegative random variable X and $\epsilon > 0$ we have

$$\mathbb{P}\{X \geq \epsilon\} \leq \frac{\mathbb{E}X}{\epsilon}$$

Therefore if $s > 0$ and X an arbitrary random variable

$$\mathbb{P}\{X \geq \epsilon\} = \mathbb{P}\left\{\mathrm{e}^{sX} \geq \mathrm{e}^{s\epsilon}\right\} \leq \frac{\mathbb{E}\mathrm{e}^{sX}}{\mathrm{e}^{s\epsilon}}.$$

Chernoff's method now is to find an $s > 0$ that minimizes the upper bound or at least makes it small. In our case we have

$$\begin{aligned}
\mathbb{P}\{S_n - \mathbb{E}S_n \geq \epsilon\} &\leq \mathrm{e}^{-s\epsilon}\mathbb{E}\left\{\exp\left(s\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right)\right\} \\
&= \mathrm{e}^{-s\epsilon}\prod_{i=1}^{n}\mathbb{E}\left\{\mathrm{e}^{s(X_i-\mathbb{E}X_i)}\right\} \quad \text{by independence} \\
&\leq \mathrm{e}^{-s\epsilon}\prod_{i=1}^{n}\mathrm{e}^{s^2(b_i-a_i)^2/8} \quad \text{by first line of proof} \\
&= \mathrm{e}^{-s\epsilon}\mathrm{e}^{s^2\sum_{i=1}^{n}(b_i-a_i)^2/8} \\
&= \mathrm{e}^{-2\epsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2} \quad \text{choose } s = 4\epsilon/\sum_{i=1}^{n}(b_i-a_i)^2
\end{aligned}$$

The second inequality is proved analogously. □

2 The Glivenko Cantelli Theorem

As a first step we study an alternative proof of the well known theorem

Theorem 3 (Glivenko-Cantelli). *Let $Z_1, \ldots, Z_n$ be i.i.d. real valued random variables with distribution function $F(z) = \mathbb{P}(Z_1 \leq z)$. Let*

$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{Z_i \leq z\}}$$

be the standard empirical distribution function. Then

$$\mathbb{P}\left\{\sup_{z\in\mathbb{R}}|F(z) - F_n(z)| > \epsilon\right\} \leq 8\,(n+1)\,\mathrm{e}^{-n\epsilon^2/32},$$

and in particular by the Borel-Cantelli lemma

$$\lim_{n\to\infty}\sup_{z\in\mathbb{R}}|F(z) - F_n(z)| = 0 \qquad \textit{with probability one.}$$

The proof we will give is not the simplest one possible, but it contains the ideas of the generalization we will consider later. The argument given here is due to symmetrization ideas of Dudley [Dud78] and Pollard [Pol84].

Proof. Assume $n\epsilon^2 > 2$, otherwise the bound is trivial. Introduce the following notation $\nu(A) = \mathbb{P}\{Z_1 \in A\}$ and $\nu_n(A) = \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}_{\{Z_j\in A\}}$ where $A \subseteq \mathbb{R}$ is a measurable set. Denote by $\mathcal{A}$ the class of sets of the form $(-\infty, z], z \in \mathbb{R}$. Then we have

$$\sup_{z\in\mathbb{R}}|F(z) - F_n(z)| = \sup_{A\in\mathcal{A}}|\nu_n(A) - \nu(A)|\,.$$

Step 1. First symmetrization by a ghost sample
Define random variables $Z_1', \ldots, Z_n' \in \mathbb{R}$ such that $Z_1, \ldots, Z_n, Z_1', \ldots, Z_n'$ are all i.i.d. Denote by ν_n' the empirical measure of the primed variables. For $n\epsilon^2 \geq 2$ we have

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A) - \nu(A)| > \epsilon\right\} \leq 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A) - \nu_n'(A)| \geq \frac{\epsilon}{2}\right\}$$

To see this let A^* be a set for which $|\nu_n(A^*) - \nu(A^*)| > \epsilon$ if existent or a fixed set in $\mathcal{A}$ otherwise. The measurability of such a choice needs to be checked, using the technique of step 3 of the proof, showing that its actually a choice from finitely many sets only. Then

$$\begin{aligned}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu_n'(A)|>\frac{\epsilon}{2}\right\} &\geq \mathbb{P}\left\{|\nu_n(A^*)-\nu_n'(A^*)|>\frac{\epsilon}{2}\right\}\\ &\geq \mathbb{P}\left\{|\nu_n(A^*)-\nu(A^*)|>\epsilon, |\nu_n'(A^*)-\nu(A^*)|<\frac{\epsilon}{2}\right\}\\ &= \mathbb{E}\left\{\mathbb{1}_{\{|\nu_n(A^*)-\nu(A^*)|>\epsilon\}}\mathbb{P}\left\{|\nu_n'(A^*)-\nu(A^*)|<\frac{\epsilon}{2}|Z_1,\ldots,Z_n\right\}\right\}.\end{aligned}$$

The conditional probability inside the expectation may be bounded by Chebyshev's inequality as follows:

$$\begin{aligned}\mathbb{P}\left\{|\nu_n'(A^*)-\nu(A^*)|<\frac{\epsilon}{2}|Z_1,\ldots,Z_n\right\} &\geq 1-\frac{\nu(A^*)\left(1-\nu(A^*)\right)}{n\epsilon^2/4}\\ &\geq 1-\frac{1}{n\epsilon^2}\geq\frac{1}{2},\end{aligned}$$

because the variance of the random variable $\nu_n(A)-\nu(A)$ equals $\frac{1}{n}\nu(A)\left(1-\nu(A)\right)$. Altogether we find

$$\begin{aligned}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu_n'(A)|>\frac{\epsilon}{2}\right\} &\geq \frac{1}{2}\mathbb{P}\left\{|\nu_n(A^*)-\nu(A^*)|>\epsilon\right\}\\ &\geq \frac{1}{2}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu(A)|>\epsilon\right\}.\end{aligned}$$

Step 2. Second Symmetrization by random signs
Let $\sigma_1,\ldots,\sigma_n$ be i.i.d. sign variables that are also independent of all Z_i, Z_i' and satisfy $\mathbb{P}\{\sigma=1\}=\mathbb{P}\{\sigma=-1\}=\frac{1}{2}$. Then using step 1

$$\begin{aligned}\mathbb{P}\left\{\sup_{A\in\mathcal{A}}|\nu_n(A)-\nu(A)|>\epsilon\right\} &\leq 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n\left(\mathbb{1}_{\{A\}}(Z_i)-\mathbb{1}_{\{A\}}(Z_i')\right)\right|>\frac{\epsilon}{2}\right\}\\ &= 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i\left(\mathbb{1}_{\{A\}}(Z_i)-\mathbb{1}_{\{A\}}(Z_i')\right)\right|>\frac{\epsilon}{2}\right\}.\end{aligned}$$

We now apply the union bound to remove the auxiliary random variables $Z_1',\ldots,Z_n'$.

$$\begin{aligned}&\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i\left(\mathbb{1}_{\{A\}}(Z_i)-\mathbb{1}_{\{A\}}(Z_i')\right)\right|>\frac{\epsilon}{2}\right\}\\ &\leq \mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i\mathbb{1}_{\{A\}}(Z_i)\right|>\frac{\epsilon}{4}\right\}+\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i\mathbb{1}_{\{A\}}(Z_i')\right|>\frac{\epsilon}{4}\right\}\\ &= 2\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i\mathbb{1}_{\{A\}}(Z_i)\right|>\frac{\epsilon}{4}\right\}.\end{aligned}$$

Step 3. Conditioning To find a bound on the probability

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i\mathbb{1}_{\{A\}}(Z_i)\right|>\frac{\epsilon}{4}\right\}=\mathbb{P}\left\{\sup_{z\in\mathbb{R}}\frac{1}{n}\left|\sum_{i=1}^n\sigma_i\mathbb{1}_{\{Z_i\leq z\}}\right|>\frac{\epsilon}{4}\right\}$$

we will condition on $Z_1,\ldots,Z_n$. The conditional probability can be bounded and afterwards we will remove the conditioning by just taking the expectation value.

We observe that, having $z_1, \dots, z_n \in \mathbb{R}$ fixed, as z ranges over $\mathbb{R}$ the number of different vectors $\left(\mathbb{1}_{\{z_1 \leq z\}}, \dots, \mathbb{1}_{\{z_n \leq z\}}\right)$ is at most $n+1$. Therefore conditioned on $Z_1, \dots, Z_n$ the supremum in the probability above is just a maximum taken over the at most $n+1$ random variables. We apply the union bound to find

$$\begin{aligned}
&\mathbb{P}\left\{ \sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{A\}}(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\} \\
&\quad \leq (n+1) \sup_{A \in \mathcal{A}} \mathbb{P}\left\{ \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{A\}}(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\}.
\end{aligned}$$

Having the supremum outside the probability we are left with finding an exponential bound on the probability

$$\mathbb{P}\left\{ \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{A\}}(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\}.$$

Step 4. Hoeffdings inequality As we have conditioned on the values of the $Z_1, \dots, Z_n$ we can regard $z_1, \dots, z_n$ as fixed. Then $\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{A\}}(z_i)$ is the sum of n independent zero mean random variables between -1 and 1. We can therefore apply Hoeffdings inequality:

$$\mathbb{P}\left\{ \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{A\}}(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\} \leq 2\mathrm{e}^{-n\epsilon^2/32}.$$

Thus we find

$$\mathbb{P}\left\{ \sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{A\}}(Z_i) \right| > \frac{\epsilon}{4} \mid Z_1, \dots, Z_n \right\} \leq 2(n+1)\mathrm{e}^{-n\epsilon^2/32}.$$

Take the expectation value on both sides to find

$$\mathbb{P}\left\{ \sup_{A \in \mathcal{A}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{A\}}(Z_i) \right| > \frac{\epsilon}{4} \right\} \leq 2(n+1)\mathrm{e}^{-n\epsilon^2/32}.$$

Altogether we have

$$\mathbb{P}\left\{ \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| > \epsilon \right\} \leq 8(n+1)\mathrm{e}^{-n\epsilon^2/32}.$$

which finishes the proof.

□

3 Generalizations

Next we want to prove the Vapnik Chervonenkis inequality, which is a mighty generalization of the Glivenko Cantelli Theorem. The proof we just studied is after a slight adjustment already proof of the stronger theorem.
The aim of the generalization is to give the statement for arbitrary classes of measurable sets $\mathcal{A}$. We then need to refine the argument in the proof were the sup is identified as a maximum.

Definition 4. Let $\mathcal{A}$ be a collection of measurable sets in $\mathbb{R}^d$. For $z_1, \dots, z_n \in \mathbb{R}^d$ we call

$$N_{\mathcal{A}}(z_1, \dots, z_n) := |\{\{z_1, \dots, z_n\} \cap A : A \in \mathcal{A}\}|$$

the index of the points $z_1, \ldots, z_n$. Further we define

$$s(\mathcal{A}, n) := \max_{z_1,\ldots,z_n \in \mathbb{R}^d} N_{\mathcal{A}}(z_1, \ldots, z_n)$$

to be the n-th shatter coefficient of the class $\mathcal{A}$.

The index is the number of different subsets of given n points that can be identified by intersecting with sets in $\mathcal{A}$. Then obviously the shatter coefficient is the maximal number of different subsets that can be picked out. It measures the diversity of the class. Clearly $s(\mathcal{A}, n) \leq 2^n$ as there are only 2^n subsets of n points. The largest integer h for which $s(\mathcal{A}, h) = 2^h$ is called the VC-dimension of the class $\mathcal{A}$. It plays a crucial role in the statistical learning theory (see [Vap00]).

Theorem 5 (Vapnik-Chervonenkis)**.** *For any probability measure ν and class of sets $\mathcal{A}$, and for any n and $\epsilon > 0$,*

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} |\nu(A) - \nu_n(A)| > \epsilon\right\} \leq 8 s(\mathcal{A}, n) \mathrm{e}^{-n\epsilon^2/32},$$

Proof. We modify the proof of Theorem 3: In step 3 we now argue that the supremum is actually a maximum over at most $s(\mathcal{A}, n)$ sets. □

Remark (Measurability). The supremum in the theorem is not always measurable. In fact this must be checked for every family $\mathcal{A}$.

Remark (Optimal Exponent). For the sake of brevity we followed Pollards([Pol84]) proof instead of the original one by Vapnik ([VC71]). In particular the exponent $-n\epsilon^2/32$ is worse than the $-n\epsilon^2/8$ in the original paper. The best known exponent for the Glivenko Cantelli Theorem is $-2n\epsilon^2$ ([Mas90]). For the Vapnik Chervonenkis inequality several refinements exist that play with the prefactor and the exponent. See for instance [Dev82].

In fact it should be clear that the statement could also be given in a probabilistic manner:

Theorem 6.

$$\mathbb{P}\left\{\sup_{A\in\mathcal{A}} |\nu(A) - \nu_n(A)| > \epsilon\right\} \leq 8\mathbb{E}\left\{N_{\mathcal{A}}(Z_1, \ldots, Z_n)\right\} \mathrm{e}^{-n\epsilon^2/32},$$

For real world applications this statement is more difficult to handle, but it is of great theoretical interest. We say that a uniform law of large numbers holds if

$$\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| \to 0 \quad \text{in probability}$$

It follows from the theorem that this is the case if

$$\frac{\log \mathbb{E}\left\{(N_{\mathcal{A}}(Z_1, \ldots, Z_n))\right\}}{n} \to 0$$

Vapnik and Chervonenkis showed is [VC71, Vap98] that this condition is also necessary for the uniform law of large numbers. Another characterization is given by Talagrand [Tal87], who showed that the uniform law of large numbers holds if and only if there does not exist a set $A \subseteq \mathbb{R}^d$ with $\nu(A) > 0$ such that, with probability one, the set $\{Z_1, \ldots, Z_n\} \cap A$ is shattered by $\mathcal{A}$.

References

[Dev82] L. Devroye, *Bounds for the uniform deviation of empirical measures*, Journal of Multivariate Analysis **12** (1982), 72–79.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi, *A probabilistic theory of pattern recognition (stochastic modelling and applied probability)*, Springer, 1996.

[Dud78] R. Dudley, *Central limit theorems for empirical measures*, Annals of Probability **6** (1978), 899–929.

[Mas90] P. Massart, *The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequaltiy*, Annals of Probability **18** (1990), 1269–1283.

[Pol84] D. Pollard, *Convegence of stochastic processes*, Springer-Verlag New York, 1984.

[Tal87] M. Talagrand, *The Glivenko-Cantelli problem*, Annals of Probability **15** (1987), 837–870.

[Vap98] N. Vladimir Vapnik, *Statistical learning theory*, Wiley Interscience, 1998.

[Vap00] Vladimir N. Vapnik, *The nature of statistical learning theory*, 2nd ed., Springer, 2000.

[VC71] V.N. Vapnik and A. Ya. Chervonenkis, *On the uniform convergence of relative frequencies of events to their probabilities*, Theory of Probability and its Applications **16** (1971), no. 2, 264–280.