The Hoeffding Bound for Tail Probabilities

Jerry Alan Veeh

January 2, 2017

Contents

1 Introduction

This note provides a development of the Hoeffding bound for the tail probability of certain types of random variables and explores some of its generalizations and applications. The bound due to Hoeffding is found in his paper *Probability Inequalities for Sums of Bounded Random Variables*, Journal of the American Statistical Association 58(301), 1963, 13–30.

The overall objective is to obtain exponential order bounds on tail probabilities, at least initially, for a sum of independent random variables. The method introduces a free parameter which is ultimately used to tighten the bound obtained by Chebyshev type methods.

2 The Bernoulli Case

To begin the discussion, consider the case of a sum of independent Bernoulli random variables. Suppose $B_1, \ldots, B_n$ are independent Bernoulli variables and denote by S the sum of the B's. Introducing the free parameter $t > 0$, and proceding as in the proof of Chebyshev's Inequality yields

$$P[S \geq x] = P[e^{tS} \geq e^{tx}] = P[e^{tS}/e^{tx} \geq 1] \leq E[e^{tS}/e^{tx}].$$

Now $E[e^{tS}]$ is the moment generating function of S, which by independence is the product of the moment generating functions of the B's. Thus

$$E[e^{tS}] = \prod_{i=1}^{n} E[e^{tB_i}].$$

The objective now is to obtain an exponential order bound on this product, as a function of t, and this will be done by bounding each of the factors.

Since an exponential order bound on each factor is desired, the logarithm of each factor will be bounded. To this end, denote by B a generic Bernoulli variable, set

$$L(t) = \ln E[e^{tB}]$$

and note that $L(0) = 0$. Now

$$L'(t) = E[Be^{tB}]/E[e^{tB}]$$

so that $L'(0) = E[B]$ and

$$\begin{aligned} L''(t) &= \frac{E[e^{tB}]E[B^2e^{tB}] - \left(E[Be^{tB}]\right)^2}{E[e^{tB}]^2} \\ &= \frac{E[B^2e^{tB}]}{E[e^{tB}]} - \left(\frac{E[Be^{tB}]}{E[e^{tB}]}\right)^2 \\ &= \frac{E[Be^{tB}]}{E[e^{tB}]}\left(1 - \frac{E[Be^{tB}]}{E[e^{tB}]}\right) \end{aligned}$$

since $B^2 = B$. This product, being of the form $u(1-u)$, does not exceed $1/4$. Applying the Fundamental Theorem of Calculus twice gives

$$L(t) = L(0) + \int_0^t L'(s)\,ds$$

$$
\begin{aligned}
&= \int_0^t \left(L'(0) + \int_0^s L''(r)\, dr \right) ds \\
&\leq E[B]t + \int_0^t \int_0^s 1/4\, dr\, ds \\
&= E[B]t + t^2/8
\end{aligned}
$$

for $t > 0$. Combining this with the previous inequality leads to the conclusion that for $t > 0$

$$P[S \geq x] \leq \exp\{tE[S] + nt^2/8 - tx\}.$$

Since this is valid for any $t > 0$, a reasonably tight bound can be obtained by minimizing the exponent over positive t. This minimum occurs at $t = 4(x - E[S])/n$, when $x \geq E[S]$, and using this value of t leads finally to the bound

$$P[S \geq x] \leq \exp\{-2(x - E[S])^2/n\}$$

for $x \geq E[S]$.

An inequality for the left hand tail can be obtained by interchanging success and failure, since $P[S \leq x] = P[n - S \geq n - x]$. Now $(n - S)$ is the sum of the independent Bernoulli variables $1 - B_i$ so by the preceding inequality,

$$P[S \leq x] \leq \exp\{-2(n - x - (n - E[S]))^2/n\} = \exp\{-2(x - E[S])^2/n\}$$

for $x \leq E[S]$.

3 Sums of Independent Bounded Variables

An examination of the preceding argument shows that the fact that the summands were Bernoulli was only used near the end of the computation of L''. In an online document, Maxim Raginsky observed that the expression

$$\frac{E[B^2 e^{tB}]}{E[e^{tB}]} - \left(\frac{E[Be^{tB}]}{E[e^{tB}]} \right)^2$$

is the variance of B computed relative to the probability measure

$$P'(A) = E[1_A\, e^{tB}]/E[e^{tB}]$$

where the expectations on the right of the equality are computed relative to the original probability measure P. Now if X is any bounded random variable on any

probability space, $E[(X-c)^2]$ is minimized, as a function of c, when $c = E[X]$. Thus $E[(X-E[X])^2] \leq E[(X-c)^2]$ for any real number c. If $a \leq X \leq b$ for real numbers a and b, then taking $c = (a+b)/2$ shows that the variance of X does not exceed $(b-a)^2/4$. Thus for such a bounded random variable X, $L(t) = \ln E[e^{tX}] \leq E[X]t + t^2(b-a)^2/8$, by using this modified estimate for L'' in the last part of the preceding argument.

Making use of this observation leads to the following result. Suppose S is the sum of n independent random variables $X_1, \ldots, X_n$ where $a_i \leq X_i \leq b_i$ for $1 \leq i \leq n$. Then for any $t > 0$,

$$P[S \geq x] \leq \exp\left\{ t\,E[S] + t^2 \sum_{i=1}^{n} (b_i - a_i)^2/8 - tx \right\}.$$

As before, the minimum of the right hand expression over positive t occurs at $t = 4(x - E[S])/\sum(b_i - a_i)^2$ provided $x \geq E[S]$. So

$$P[S \geq x] \leq \exp\left\{ -2(x - E[S])^2 / \sum_{i=1}^{n} (b_i - a_i)^2 \right\}.$$

for $x \geq E[S]$. Since this bound applies to any such X's, replacing each X_i by $-X_i$ gives

$$P[S \leq x] \leq \exp\left\{ -2(x - E[S])^2 / \sum_{i=1}^{n} (b_i - a_i)^2 \right\}.$$

for $x \leq E[S]$.

4 From Independence to Dependence

A review of the arguments given thus far show that independence of the summands was really only used to write the moment generating function of the sum as a product of the moment generating functions of the summands. This suggests that independence may not be required. In fact, the random variable of interest is not even required to be a sum.

Given a bounded random variable S on some probability space and an increasing sequence of sub sigma algebras $\mathscr{F}_0 \subset \mathscr{F}_1 \subset \cdots \subset \mathscr{F}_n$, S can be written as a telescoping sum

$$S = E[S|\mathscr{F}_0] + \sum_{i=0}^{n-1} E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i]$$

provided only that S is $\mathscr{F}_n$ measurable. Suppose also that $\mathscr{F}_0$ is the trivial sigma algebra so that $E[S|\mathscr{F}_0] = E[S]$. Then

$$E[e^{tS}] = e^{tE[S]} E\left[\prod_{i=0}^{n-1} \exp\{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])\}\right].$$

To continue, consider bounding the conditional expectation of a single factor in this expression given the smaller sigma algebra. Set

$$L(t) = \ln E[\exp\{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])\}|\mathscr{F}_i]$$

and note that $L(0) = 0$. Now

$$L'(t) = \frac{E[(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i]) \exp\{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])\}|\mathscr{F}_i]}{E[\exp\{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])\}|\mathscr{F}_i]}$$

so that $L'(0) = E[(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])|\mathscr{F}_i] = 0$ and as before

$$\begin{aligned} L''(t) = & \frac{E[(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])^2 e^{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])}|\mathscr{F}_i]}{E[e^{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])}|\mathscr{F}_i]} \\ & - \left(\frac{E[(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i]) e^{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])}|\mathscr{F}_i]}{E[e^{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])}|\mathscr{F}_i]}\right)^2, \end{aligned}$$

which is the variance of $E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i]$ computed using the probability measure defined by

$$P'(A) = \frac{E[1_A\, e^{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])}|\mathscr{F}_i]}{E[e^{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])}|\mathscr{F}_i]}.$$

Notice that expectations computed relative to P' are expectations relative to P conditionally give $\mathscr{F}_i$. So the argument given earlier about bounding the variance with respect to P' shows that if there are $\mathscr{F}_i$ measurable functions a_i and b_i with $a_i \le E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i] \le b_i$, then $L''(t) \le (b_i - a_i)^2/4$. In the cases considered here the difference $b_i - a_i$ will be bounded by an absolute constant. Under this proviso, using the Taylor expansion argument given earlier shows that $L(t) \le t^2(b_i - a_i)^2/8$. This leads recursively to the bound

$$E[e^{tS}] = e^{tE[S]} E\left[\prod_{i=0}^{n-1} \exp\{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])\}\right]$$

$$
\begin{aligned}
&= e^{tE[S]} E\left[E\left[\prod_{i=0}^{n-1} \exp\{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])\} \middle| \mathscr{F}_{n-1}\right]\right] \\
&= e^{tE[S]} E\left[\prod_{i=0}^{n-2} e^{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])} E\left[e^{t(E[S|\mathscr{F}_n] - E[S|\mathscr{F}_{n-1}])} \middle| \mathscr{F}_{n-1}\right]\right] \\
&\leq e^{tE[S]} E\left[\prod_{i=0}^{n-2} e^{t(E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i])}\right] e^{t^2(b_{n-1}-a_{n-1})^2/8} \\
&\vdots \\
&\leq e^{tE[S]} \prod_{i=0}^{n-1} \exp\{t^2(b_i - a_i)^2/8\}
\end{aligned}
$$

subject to the condition that $a_i \leq E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i] \leq b_i$ for $\mathscr{F}_i$ measurable functions with $b_i - a_i$ constant. Applying these estimates and the Chebyshev bound shows that for $t > 0$

$$
P[S \geq x] \leq E\left[\exp\left\{tE[S] + t^2 \sum_{i=0}^{n-1} (b_i - a_i)^2/8 - tx\right\}\right]
$$

from which the tail estimates

$$
P[S \geq x] \leq \exp\left\{-2(x - E[S])^2 / \sum_{i=0}^{n-1} (b_i - a_i)^2\right\}
$$

for $x \geq E[S]$ and

$$
P[S \leq x] \leq \exp\left\{-2(x - E[S])^2 / \sum_{i=0}^{n-1} (b_i - a_i)^2\right\}
$$

for $x \leq E[S]$ follow by minimization on t as before. This type of reasoning was mentioned by Hoeffding in his paper cited above. The inequality obtained in this way is often referred to as McDiarmid's Inequality.

As a simple illustration, return to the sum of independent Bernoulli variables considered at the start. Taking $\mathscr{F}_i$ to be the sigma algebra generated by the first i summands gives $E[S|\mathscr{F}_{i+1}] - E[S|\mathscr{F}_i] = B_{i+1} - E[B_{i+1}]$, so that $a_i = -E[B_{i+1}]$ and $b_i = 1 - E[B_{i+1}]$ with $b_i - a_i = 1$. This leads to the same tail bound as derived earlier.

5 Hypergeometric Tails

A simple but more interesting application of this technique is estimating the tail probabilities of a hypergeometric random variable. Suppose S_n is the number of successes in n draws without replacement from an urn containing s balls labeled 'success' and f ball labeled 'failure.' Then S_n has a hypergeometric distribution. Suppose also that $0 < n < s+f$. Let $\mathscr{F}_0$ be the trivial sigma algebra and for each $1 \le i \le n$ denote by $\mathscr{F}_i$ the sigma algebra generated by $S_1, \ldots, S_i$. Given $\mathscr{F}_i$, there are $s - S_i$ success balls in the urn out of a total of $s+f-i$ balls. Thus $E[S_n|\mathscr{F}_i] = S_i + (n-i)(s-S_i)/(s+f-i)$ since each draw after the ith is a Bernoulli trial with success probability $(s-S_i)/(s+f-i)$. Simplification gives $E[S_n|\mathscr{F}_i] = \frac{s(n-i)}{s+f-i} + \frac{s+f-n}{s+f-i} S_i$. Since $S_i \le S_{i+1} \le 1 + S_i$, $E[S_n|\mathscr{F}_{i+1}] - E[S_n|\mathscr{F}_i]$ is easily seen to be bounded between two linear functions a_i and b_i of S_i with $b_i - a_i = \frac{s+f-n}{s+f-i-1} \le 1$, for $0 \le i \le n-1$. Using the upper bound of 1 in the estimate of the previous section shows that the moment generating function of S_n obeys the same bound as obtained for the sum of independent Bernoulli trials, and so the tail estimate for S_n is also the same. This is not surprising, since a hypergeometric random variable can always be written as the sum of independent (but not identically distributed) Bernoulli variables.

However, using the more precise formula $b_i - a_i = \frac{s+f-n}{s+f-i-1}$ will yield a tighter bound, since this difference will be substantially less than 1 in most instances. An estimate given by Serfling in the paper cited below indicates the amount of improvement. Simple estimation gives

$$\frac{1}{i^2} \le \frac{1}{(i-\frac{1}{2})(i+\frac{1}{2})} = \int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \frac{1}{x^2}\, dx.$$

Thus

$$\begin{aligned}
\sum_{i=0}^{n-1} (b_i - a_i)^2 &= (s+f-n)^2 \sum_{i=1}^{n} \frac{1}{(s+f-i)^2} \\
&= (s+f-n)^2 \sum_{i=s+f-n}^{s+f-1} \frac{1}{i^2} \\
&= (s+f-n)^2 \left(\frac{1}{(s+f-n)^2} + \sum_{i=s+f-n+1}^{s+f-1} \frac{1}{i^2} \right) \\
&\le (s+f-n)^2 \left(\frac{1}{(s+f-n)^2} + \int_{s+f-n+\frac{1}{2}}^{s+f-\frac{1}{2}} \frac{1}{x^2}\, dx \right)
\end{aligned}$$

$$
\begin{aligned}
&= (s+f-n)^2\left(\frac{1}{(s+f-n)^2}+\frac{n-1}{(s+f-n+\frac{1}{2})(s+f-\frac{1}{2})}\right)\\
&\le (s+f-n)^2\left(\frac{1}{(s+f-n)^2}+\frac{n-1}{(s+f-n)(s+f)}\right)\\
&= \frac{s+f+(s+f-n)(n-1)}{s+f}\\
&= n\left(1-\frac{n-1}{s+f}\right).
\end{aligned}
$$

Using this in place of n to complete the tail bound gives the estimate found by Serfling in the general case of sampling from a finite population. (See R.J. Serfling *Probability Inequalities for the Sum in Sampling Without Replacement*, The Annals of Statistics, volume 2, number 1, 39–48, 1974).

To complete the details, the foregoing estimate leads to the moment generating function bound

$$
E[\exp\{tS_n\}] \le \exp\left\{t\,E[S_n]+t^2 n\left(1-\frac{n-1}{s+f}\right)/8\right\}
$$

from which

$$
P[S_n \ge x] \le \exp\left\{-2(x-E[S_n])^2/n\left(1-\frac{n-1}{s+f}\right)\right\}
$$

for $x \ge E[S]$ with the same estimate for the left hand tail.

6 The Multivariate Glivenko–Cantelli Theorem

As a final application, a proof of the multivariate Glivenko–Cantelli Theorem will be given.

Suppose $X_1, X_2, \ldots$ are independent identically distributed random vectors on $\mathbf{R}^d$ and define the empirical distribution function F_n by

$$
F_n(v) = \frac{1}{n}\sum_{i=1}^{n} 1_v(X_i)
$$

where $1_v(x) = 1$ exactly when $x \le v$ in the coordinatewise ordering, and is zero otherwise. Denote by F the common multivariate distribution function of the X's and set

$$
D_n = \sup_v |F_n(v) - F(v)|.
$$

The Glivenko–Cantelli Theorem is the assertion that D_n converges to zero almost surely as $n \to \infty$.

The proof given here uses some ideas from the paper by Vapnik and Chervonenkis (*On the Uniform Convergence of Relative Frequencies of Events to their Probabilities*, Theory of Probability and Its Applications, Volume 16, number 2, 264–280,1971). They appear to be the first to use the technique below to bound the supremum. The idea of replacing the actual distribution function with an empirical distribution function based on a larger sample is taken from Devroye *Bounds for the Uniform Deviation of Empirical Measures*, Journal of Multivariate Analysis, Volume 12, 72–79, 1982.

The overall idea is to bound the tail probability $P[D_n \geq x]$ with a bound that is summable on n. An application of the Borel–Cantelli Lemma will complete the proof of almost sure convergence. The tail probability bound will be obtained using a modification of the now familiar methods.

The first step will of course be the Chebyshev bound

$$P[D_n > x] \leq E[e^{tD_n}]\, e^{-tx}.$$

The bulk of the work is to obtain a bound on the moment generating function of D_n. The first step in this enterprise is to eliminate the actual distribution function F by replacing it with an independent copy of the empirical distribution function. This independent copy is given by

$$G_m(v) = \frac{1}{m} \sum_{i=n+1}^{n+m} 1_v(X_i)$$

where $m \geq 1$ will be selected later. Let $\mathscr{F}$ denote the sigma algebra generated by $X_1, \ldots, X_n$ and note that by independence of the X's, $E[G_m(v)|\mathscr{F}] = F(v)$, while $E[F_n(v)|\mathscr{F}] = F_n(v)$ since F_n is $\mathscr{F}$ measurable. Then for $t > 0$

$$\begin{aligned}
E[e^{tD_n}] &= E\left[\exp\left\{t \sup_v |F_n(v) - F(v)|\right\}\right] \\
&= E\left[\exp\left\{t \sup_v \left|F_n(v) - E\left[G_m(v)|\mathscr{F}\right]\right|\right\}\right] \\
&= E\left[\exp\left\{t \sup_v \left|E\left[F_n(v) - G_m(v)|\mathscr{F}\right]\right|\right\}\right] \\
&\leq E\left[\exp\left\{t \sup_v E\left[|F_n(v) - G_m(v)|\,|\mathscr{F}\right]\right\}\right]
\end{aligned}$$

$$
\begin{aligned}
&\leq E\left[\exp\left\{E\left[t\sup_v |F_n(v)-G_m(v)|\,\middle|\,\mathscr{F}\right]\right\}\right]\\
&\leq E\left[E\left[\exp\left\{t\sup_v |F_n(v)-G_m(v)|\right\}\,\middle|\,\mathscr{F}\right]\right]\\
&= E\left[\exp\left\{t\sup_v |F_n(v)-G_m(v)|\right\}\right]
\end{aligned}
$$

where the conditional form of Jensen's Inequality was used for the final inequality. Now the X's are independent and identically distributed, so if σ is a random permutation of the set $\{1,\ldots,n+m\}$ that is independent of the X's, replacing each X_i by $X_{\sigma(i)}$ does not change this last expectation. Denote now by $\mathscr{G}$ the sigma algebra generated by $X_1,\ldots,X_{n+m}$. Using this notation and making the foregoing substitution gives

$$
\begin{aligned}
&E\left[\exp\left\{t\sup_v |F_n(v)-G_m(v)|\right\}\right]\\
&= E\left[\exp\left\{t\sup_v \left|\frac{1}{n}\sum_{i=1}^{n}1_v(X_{\sigma(i)})-\frac{1}{m}\sum_{i=n+1}^{n+m}1_v(X_{\sigma(i)})\right|\right\}\right]\\
&= E\left[E\left[\exp\left\{t\sup_v \left|\frac{1}{n}\sum_{i=1}^{n}1_v(X_{\sigma(i)})-\frac{1}{m}\sum_{i=n+1}^{n+m}1_v(X_{\sigma(i)})\right|\right\}\,\middle|\,\mathscr{G}\right]\right]\\
&= E\left[E\left[\sup_v \exp\left\{t\left|\frac{1}{n}\sum_{i=1}^{n}1_v(X_{\sigma(i)})-\frac{1}{m}\sum_{i=n+1}^{n+m}1_v(X_{\sigma(i)})\right|\right\}\,\middle|\,\mathscr{G}\right]\right].
\end{aligned}
$$

There are now three critical observations about the conditional expectation.

First, since the X's are fixed the supremum is actually a maximum over a set of at most $(n+m+1)^d$ distinct v's. To see this, consider the one dimensional case with the $n+m$ points given by the X's fixed on the line. Using sets of the form $(-\infty,v]$ at most $n+m+1$ subsets of the points can be selected, and only these subsets give rise to (potentially) different values for the sums. Since the coordinatewise ordering is used, this reasoning extends to d dimensions with the indicated bound.

Second, since $|a| = \max\{a,-a\}$ the absolute values in the exponent can be removed if the maximum is extended to include the sum as written and its negative for each of the finite number of v's.

Third, for fixed v given the X's, there are $s = \sum_{i=1}^{n+m} 1_v(X_i)$ indicators taking the value 1 in the sum, and $f = n+m-s$ indicators which are zero. If $H = \sum_{i=1}^{n} 1_v(X_{\sigma(i)})$, then H has a hypergeometric distribution based on n draws without replacement from an urn containing s 'successes' and f 'failures.' Thus

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} 1_v(X_{\sigma(i)}) - \frac{1}{m}\sum_{i=n+1}^{n+m} 1_v(X_{\sigma(i)}) &= \frac{1}{n}H - \frac{1}{m}(s-H) \\ &= \frac{n+m}{nm}\left(H - \frac{n}{n+m}s\right).\end{aligned}$$

By the earlier discussion of the hypergeometric case

$$\begin{aligned}E\left[\exp\left\{t\frac{n+m}{nm}\left(H-\frac{n}{n+m}s\right)\right\}\right] &\le \exp\left\{t^2 n\left(1-\frac{n-1}{n+m}\right)\left(\frac{n+m}{nm}\right)^2/8\right\} \\ &= \exp\left\{t^2\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2/8n\right\}\end{aligned}$$

since $E[H] = ns/n+m$. Note that this estimate does not depend on the X's or on v. The same estimate holds if $\frac{n+m}{nm}\left(H - \frac{n}{n+m}s\right)$ is replaced by its negative.

Using these three observations, the conditional expectation above is bounded by the sum of $2(n+m+1)^d$ terms, each of which is bounded by this last exponential. Thus

$$E[e^{tD_n}] \le 2(n+m+1)^d \exp\left\{t^2\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2/8n\right\}$$

so that

$$P[D_n > x] \le 2(n+m+1)^d \exp\left\{-tx + t^2\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2/8n\right\}.$$

Minimizing on $t > 0$ leads to

$$P[D_n > x] \le 2(n+m+1)^d \exp\left\{-2nx^2/\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2\right\}.$$

Now for $n \ge 2$ take $m = n^2 - n$ so that $\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2 = \frac{n^2-n+1}{n^2-2n+1}$. Then

$$n/\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2 = \frac{n(n^2-2n+1)}{n^2-n+1} = n-1+\frac{1-n}{n^2-n+1}$$

and the last fraction lies between $-1/3$ and zero for $n \geq 2$. Thus

$$-2nx^2/\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2 = -2x^2\left(n-1+\frac{1-n}{n^2-n+1}\right) \leq -2nx^2+8x^2/3$$

as long as $n \geq 2$. Finally this gives

$$P[D_n > x] \leq 2\,e^{8x^2/3}\,(n^2+1)^d\,e^{-2nx^2}$$

for $n \geq 2$. Since the term on the right is summable on n for each fixed $x > 0$, the Borel–Cantelli Lemma shows that $D_n \to 0$ almost surely as $n \to \infty$, completing the proof of the Glivenko–Cantelli Theorem.

Most of the work in the foregoing proof was to establish the bound

$$E[\exp\{tD_n\}] \leq 2(n+m+1)^d \exp\left\{t^2\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2/8n\right\}.$$

An application of Jensen's Inequality gives $\exp\{tE[D_n]\} \leq E[\exp\{tD_n\}]$ for $t > 0$, so the above argument also provides the estimate

$$\begin{aligned} E[D_n] &\leq \frac{1}{t}\ln\left(2(n+m+1)^d \exp\left\{t^2\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2/8n\right\}\right) \\ &= \ln(2(n+m+1)^d)/t + t\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2/8n \end{aligned}$$

Minimizing this expression over $t > 0$ and simplifying gives

$$\begin{aligned} E[D_n] &\leq 2\sqrt{\ln(2(n+m+1)^d)\left(1-\frac{n-1}{n+m}\right)\left(1+\frac{n}{m}\right)^2/8n} \\ &\leq 2\sqrt{\ln(2(n^2+1)^d)/2n}. \end{aligned}$$

where the last inequality follows by simple estimation after choosing $m = n^2 - n$ for $n \geq 2$. Thus $E[D_n]$ is of order $\sqrt{\ln n/n}$ in n. While of interest in its own right, a second proof of the Glivenko–Cantelli Theorem can now be given along the following lines. First, apply McDiarmid's Inequality with $b_i - a_i = 1/n$ to show that $P[|D_n - E[D_n]| > x] \leq 2\exp\{-2nx^2\}$. The Borel–Cantelli Lemma then implies that $D_n - E[D_n] \to 0$ almost surely as $n \to \infty$. The estimate on $E[D_n]$ shows

that $E[D_n] \to 0$ too, so $D_n \to 0$ almost surely. A side benefit of this approach is that the rate of convergence to zero has also been approximated.

Almost none of the above argument used the structure of the sets $x \leq v$ for fixed v in $\mathbf{R}^d$. The key issue was to bound the number of subsets this collection of infinite rectangles could pick out from a set of $n+m$ fixed points. With this in mind, the same argument can be used to compare the behavior of the empirical measure with the acutal underlying probability measure over any class of sets for which a similar, sub-exponential growth estimate on the number of picked out subsets can be obtained. Moreover, since the estimate of the number of picked out subsets occured inside the expectation, a random bound depending on the X's could also be used in a similar way. These sorts of issues are studied in connection with Vapnik–Chervonenkis classes, which will not be examined here.