Basic Statistical Bounds: From Markov to Hoeffding

Nathan Ratliff

Jan 30, 2015

Abstract

This document explores some statistical bounds that have found a number of applications in Intelligent Systems, especially in Machine Learning. We start off discussing the the basic notion of an empirical mean estimator as a random variable and derive Markov's inequality and Chebyshev's inequality as stepping stones toward proving the Weak Law of Large Numbers. We then build on these results to prove Hoeffding's inequality, which relates the probability that an empirical mean estimator deviates from the true mean to characteristics of the domain and the number of samples. The final sections explore how to interpret and manipulate Hoeffding's inequality to answer important practical questions such as how many samples we'd need to guaranteed a certain confidence that the mean estimator falls close to the true mean.

1 The Law of Large Numbers

Intuitively, empirical averages of samples should converge on the true mean of a distribution in the long run. But what happens when you have a huge value that comes up only ever so often? Every time it appears, it changes the mean significantly. Doesn't it reason, then, that maybe these random events could make the empirical mean fluctuate so drastically it moves off and converges somewhere else, or not at all, over time?

When gambling on slot machines, for instance, you might wait out a drought for that one big Jackpot so you can leave a winner and beat the system. Perhaps you really can win in gambling. But really, in rationalizing that logic you inevitably ignore the lavish decorations and displays adorning the casino's entrance, the tale-tale-sign that they've pulled millions of dollars from the pockets of similarly hopeful patrons. Casinos bank on the Law of Large Numbers, a rigorous statement that it is indeed true that empirical averages converge to the true mean over time independent of the particulars of the underlying distribution. You might leave a winner today, but the Powers that Be will make sure that it all evens out over time, and the savvy casino owners ensure that if you keep it up they'll be digging into your pocket before your stint's done.

1.1 Formalizing the problem

To start, we need to understand formally what each of these vague statements mean. Lets start with the empirical expectation. To understand the general long term behavior of an empirical expectation, we must acknowledge first that the empirical expectation, itself, is a random variable. Let X be any random variable with finite expected value $\mathrm{E}[X] = \mu$ and variance $\mathrm{Var}[X] = \sigma^2$. If we sample n values $\{x_i\}_{i=1}^n$ from X and calculate their empirical expectation $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$, we get a particular value. But if we sample a new set of n values from X, most likely the resulting empirical expectation will be slightly different. Repeating this process multiple times will create a collection of empirical estimates of the expectation, which themselves form a distribution. To characterize this distribution, we construct a new random variable

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \tag{1}$$

from a collection of n other random variables $\{X_i\}_{i=1}^n$, each assumed to be distributed identically to X and independent of one another (specifically, the covariance between any pair of them $\mathrm{Cov}(X_i, X_j)$ is 0 when $i \neq j$ and σ^2 when $i = j$). Random variables of this sort are often called independent and identically distributed, or i.i.d. for short.

Since this empirical mean estimator $\overline{X}_n$ is, itself, a random variable, we can ask what is its mean and variance. It's mean is easy to calculate since the expectation operator is linear

$$\begin{aligned}\mathrm{E}\left[\overline{X}_n\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[X_i] \\ &= \frac{1}{n} n\mu = \mu. \end{aligned} \tag{2}$$

Independent of n, the mean of the empirical mean estimator $\overline{X}_n$ is always the same as the mean of the underlying variable X, itself. Estimators with this property are known as *unbiased estimators*.

Similarly, the variance is easy to calculate as well:

$$\begin{aligned}\mathrm{Var}\left[\overline{X}_n\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^n X_i\right] \\ &= \frac{1}{n^2} n\mathrm{Var}[X_i] = \frac{\sigma^2}{n}. \end{aligned} \tag{3}$$

We can push the variance operator inside the sum to get to the second line, in this case, since each X_i is independent of all others.

The following statement formalizes what we mean when we say $\overline{X}_n$ converges to the expected value $\mu = \mathrm{E}[X]$ as n increases:

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0, \tag{4}$$

for any $\epsilon > 0$. This convergences is also denoted as $\overline{X}_n \xrightarrow{P} \mu$. A random variable is said to *converge in probability* if such a statement holds true. This type of probabilistic convergence is known as "weak" convergence, and accordingly the statement is known as the **Weak Law of Large Numbers**. A "strong" form of the result can be proven as well, although it's proof is more complicated. This **Strong Law of Large Numbers** is formally stated as

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1. \tag{5}$$

The shorthand notation for this type of convergence is $\overline{X}_n \xrightarrow{a.s.} \mu$. Random variables that converge in this way are said to converge *almost surely*.

This document presents Markov's Inequality and Chebyshev's Inequality as building blocks of a straightforward proof of the Weak Law of Large Numbers.

1.2 Markov's Inequality

Markov's inequality holds for any non-negative random variable X with finite expected value $\mathrm{E}[X]$. It states that for any $\eta > 0$,

$$P(X > \eta) \leq \frac{\mathrm{E}[X]}{\eta}. \tag{6}$$

It's true for any positive η, but is really useful only when $\eta > \mathrm{E}[X]$. It's a simple and somewhat loose statement that the probability of a sample being larger than a given threshold scales proportionally to its expected value and inversely with the threshold. Effectively, taking the magnitude of the expectation as the units, and the upper bound on probabilities diminishes in the same way for all distributions when expressed in those units.

Proving this statement amounts to a simple manipulation of the integral

$$\mathrm{E}[X] = \int_0^\infty x\, p(x)dx = \underbrace{\int_0^\eta x\, p(x)dx}_{\geq 0} + \int_\eta^\infty x\, p(x)dx \geq \int_\eta^\infty x\, p(x)dx, \tag{7}$$

where the inequality results simply because we're removing the first term which is guaranteed to be positive. The final term in that expression is lower bounded by

$$\int_\eta^\infty x\, p(x)dx \geq \int_\eta^\infty \eta\, p(x)dx$$

since $x\, p(x) \geq \eta\, p(x)$ for $x \geq \eta$. Plugging that results in we get

$$\mathrm{E}[X] \geq \eta \underbrace{\int_\eta^\infty p(x)dx}_{P(X>\eta)}.$$

Rearranging this expression gives Equation 7.

1.3 Chebyshev's Inequality

Chebyshev's inequality removes the non-negativity requirement and holds for any random variable with finite expectation and variance. It states that

$$P(|X - \mathrm{E}[X]| \geq \eta) \leq \frac{\mathrm{Var}[X]}{\eta^2} = \left(\frac{\sigma}{\eta}\right)^2, \tag{8}$$

where $\sigma = \sqrt{\mathrm{Var}[X]}$ is the standard deviation of X. The larger the variance, the larger the probability that X deviates from its mean by some threshold η.

This result is a straightforward application of Markov's inequality:

$$\begin{aligned} P(|X - \mathrm{E}[X]| \geq \eta) &= P((X - \mathrm{E}[X])^2 \geq \eta^2) \\ &\leq \frac{\mathrm{E}\left[(X - \mathrm{E}[X])^2\right]}{\eta^2} \qquad \text{(Markov's inequality)} \\ &= \frac{\mathrm{Var}[X]}{\eta^2}. \end{aligned}$$

1.4 Proof of the Weak Law of Large Numbers

This proof uses Chebyshev's inequality to express probability of $\overline{X}_n$ deviating from its mean (which is the actual mean $\mathrm{E}\left[\overline{X}_n\right] = \mathrm{E}[X] = \mu$ (see above)) in terms of its variance $\mathrm{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$. That relates the probability to n, the number of samples, which allows us to examine the limit as $n \to \infty$.

Explicitly, we get

$$\begin{aligned} P(|\overline{X}_n - \mu| \geq \epsilon) &= P(|\overline{X}_n - \mathrm{E}\left[\overline{X}_n\right]| \geq \epsilon) \\ \mathrm{E}[X] \leq \frac{\mathrm{Var}\left[\overline{X}_n\right]}{\epsilon^2} &= \frac{\sigma^2}{n\epsilon^2} \xrightarrow[n\to\infty]{} 0. \end{aligned}$$

for any finite σ and threshold ϵ. Looking back to Equation 4, we see that this result proves $\overline{X}_n \xrightarrow{P} \mu$.

2 Hoeffding's Inequality

At a high level, **Hoeffding's inequality** states that the probability of an empirical mean estimate $\overline{X}_n$ deviating from the true mean by any fixed value decreases exponentially with the number of samples. To be precise, suppose $X_1, \ldots, X_n$ are n i.i.d. random variables with $\mathrm{E}[X_i] = \mu$ and $a \leq X_i \leq b$ for all i. Then for any $\epsilon > 0$,

$$P(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}. \tag{9}$$

This inequality quantifies how the deviation of the empirical mean estimator from the true mean relates to the number of samples n, the size of the deviation from true ϵ, and the domain width $(b - a)$.

Proving Hoeffding's inequality is easy if we take the following lemma as fact. For any zero mean random variable X with bounded domain (i.e. with $\mathrm{E}[X] = 0$ and $a \leq X \leq b$), we have

$$\mathrm{E}\left[e^{tX}\right] \leq e^{t^2(b-a)^2/8}. \tag{10}$$

The proof of this lemma is tedious so we postpone it until Section 2.3.

We can always zero center our random variables by constructing a new set of variables as $Y_i = X_i - \mu$. The new mean is $\mathrm{E}[Y_i] = 0$ and the probabilistic statement $P(|\overline{Y}_n| \geq \epsilon)$ is equivalent to the original statement $P(|\overline{X}_n - \mu| \geq \epsilon)$, so we can examine just the former for simplicity.

Since the logical relations $\overline{Y}_n \geq \epsilon$ and $\overline{Y}_n \leq -\epsilon$ are both disjoint (they're never true simultaneously) and together constitute the single relation $|\overline{Y}_n| \geq \epsilon$, we can decompose the probability as

$$\begin{aligned} P(|\overline{Y}_n| \geq \epsilon) &= P(\overline{Y}_n \geq \epsilon) + P(\overline{Y}_n \leq -\epsilon) \\ &= P(\overline{Y}_n \geq \epsilon) + P(-\overline{Y}_n \geq \epsilon). \end{aligned} \tag{11}$$

We'll examine the first of these terms in depth and then see below that the second is straightforward given this analysis.

To bound the first term, we just need to coax the expression into a form suitable for applying both Markov's inequality and the above lemma given in Equation 10:

$$\begin{aligned} P(\overline{Y}_n \geq \epsilon) = P\left(\sum_{i=1}^{n} Y_i \geq n\epsilon\right) &= P\left(e^{\sum_{i=1}^{n} Y_i} \geq e^{n\epsilon}\right) = P\left(e^{t\sum_{i=1}^{n} Y_i} \geq e^{tn\epsilon}\right) \\ &\leq \frac{\mathrm{E}\left[e^{t\sum_{i=1}^{n} Y_i}\right]}{e^{tn\epsilon}} && \text{(Markov's inequality)} \\ &= e^{-tn\epsilon} \prod_{i=1}^{n} \mathrm{E}\left[e^{tY_i}\right] && \text{(Independence of } Y_i\text{)} \\ &= e^{-tn\epsilon} \left(\mathrm{E}\left[e^{tY_k}\right]\right)^n && \text{(Each } Y_i \text{ has the same mean).} \end{aligned}$$

Now, applying the lemma, we see that $\mathrm{E}\left[e^{tY_k}\right] \leq e^{t^2(b-a)^2/8}$, so we have

$$P(\overline{Y}_n \geq \epsilon) \leq e^{-tn\epsilon} e^{t^2 n(b-a)^2/8} = e^{t^2 n(b-a)^2/8 - tn\epsilon}. \tag{12}$$

This bound holds for any t, so analytically optimizing over the expression in the exponent produces the best possible bound

$$\begin{aligned} &\frac{d}{dt}\left(t^2 n(b-a)^2/8 - tn\epsilon\right) = 0 \\ \Rightarrow &\frac{n(b-a)^2}{4} t - n\epsilon = 0 \\ \Rightarrow &t = \frac{4\epsilon}{(b-a)^2}. \end{aligned}$$

This critical point for t is the global minimum because the function is quadratic and positive definite. Plugging this optimal value back into Equation 12 gives

$$P(\overline{Y}_n \geq \epsilon) \leq e^{-2n\epsilon^2/(b-a)^2}. \tag{13}$$

And finally, when we apply the same argument to the second term $P(-\overline{Y}_n \geq \epsilon)$ in Equation 11, the negative sign gets absorbed into the arbitrary scaling variable t that we ultimately optimize over in Equation 12. The result is, therefore, the same, so the right hand side of the final bound given in Equation 9 is just twice that of Equation 13.

2.1 Confidence bounds

There are a number of ways we can use Hoeffding's inequality in calculations. This section and the next explore two common manipulations.

Suppose we're handed n random samples $X_1, \ldots, X_n$ and we want to be able to state with confidence that the empirical mean of these samples lies within a given interval of the true mean. How can we use Hoeffding's inequality to make such a claim?

The probabilistic expression $1 - \delta = P(|\overline{X}_n - \mu| < \epsilon)$ states that with probability $1 - \delta$, the value $\overline{X}_n$ lies within ϵ of the true expected value μ. It's also sometimes said that we're $(1-\delta)100$ percent confident that $\overline{X}_n$ lies in that range. For instance, if $\delta = .05$ so that $(1 - \delta)100 = 95$, we'd say we're 95% confident that $\overline{X}_n$ lies within ϵ of μ.

Given that interpretation, and the relation

$$\begin{aligned} 1 - \delta = P(|\overline{X}_n - \mu| < \epsilon) &= 1 - P(|\overline{X}_n - \mu| \geq \epsilon) \\ \Rightarrow \delta &= P(|\overline{X}_n - \mu| \geq \epsilon), \end{aligned}$$

we can say, if we want the confidence probability to be $1 - \delta$, it must be that

$$\delta = P(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/c^2},$$

where $c = b - a$. Rearranging this expression gives

$$\epsilon \leq c\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}.$$

In other words, since the probabilistic statement says that we're $1 - \delta$ confident that $|\overline{X}_n - \mu| < \epsilon$, we can say with $(1 - \delta)100$ percent certainty that

$$|\overline{X}_n - \mu| < c\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}.$$

2.2 Quantifying the number of samples needed for a given accuracy

Algebraic manipulations without explicit interpretation can be misleading. Blindly setting $\delta = P(|\overline{X}_n - \mu| \geq \epsilon)$ and rearranging Hoeffding's inequality would actually work for answering the question explored in Section 2.1. But this section, examines a question for which that strategy would lead to nonsense. Specifically, we ask how many samples n would we need in order to achieve a particular confidence $1 - \alpha$ that $\overline{X}_n$ is within ϵ of μ.

Blindly setting $\alpha = P(|\overline{X}_n - \mu| \geq \epsilon)$ and solving for n in Hoeffding's inequality is wrong. Hoeffding's inequality gives us an upper bound on our confidence that $\overline{X}_n$ is outside the desired range. So solving that problem just finds a value for n for which the upper bound is larger than some given value α. That's exactly the opposite of what we're trying to do. We want to ensure that the probability that $\overline{X}_n$ is outside the desired epsilon range $P(|\overline{X}_n - \mu| \geq \epsilon)$ is small. We have an upper bound on that probability. So the way to do that is to make the upper bound small. We need to explicitly find the value of n for which the *upper bound* is smaller than our prespecified value α. If we find such an n, we're guaranteed that $P(|\overline{X}_n - \mu| \geq \epsilon)$ is smaller than α, too.

So the expression that we need to rearrange is

$$P(|\overline{X}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2} \leq \alpha.$$

Rearranging the expression gives

$$n \geq \frac{(b-a)^2}{2\epsilon^2} \log\left(\frac{2}{\alpha}\right). \tag{14}$$

To ensure a confidence of at least $1 - \alpha$ that $|\overline{X}_n - \mu| < \epsilon$, we need at least this many samples.

2.3 Proof of lemma

Lemma 1. *Let X be a random variable with $\mathrm{E}[X] = 0$ and $a \leq X \leq b$. Then*

$$\mathrm{E}\left[e^{tX}\right] \leq e^{t^2(b-a)^2/8}. \tag{15}$$

Proof. X is always between a and b, so we can represent it as a weighted average (convex combination) of the two end points $X = \alpha b + (1-\alpha)a$ where α is a random variable defined as $\alpha = (X-a)/(b-a)$. Notice that α is always between 0 and 1, so X is explicitly a weighted average of a and b.

For any convex function g the chord connecting two points on its graph always lies above the function. Algebraically, that means $g(\alpha b + (1-\alpha)a) \leq \alpha g(b) + (1-\alpha)g(a)$ for any particular sample from the random variable α. In particular, that means the expected value over the left hand side, which is an integral over all these specific values, must be smaller than the expected value

of the right hand side, which is the same integral over the larger values:

$$\begin{aligned}\mathrm{E}\left[g(\alpha b + (1-\alpha)a)\right] &= \mathrm{E}\left[g(X)\right] \\ &\leq \mathrm{E}\left[\alpha g(b) + (1-\alpha)g(a)\right] \\ &= \mathrm{E}\left[\alpha\right] g(b) + (1 - \mathrm{E}\left[\alpha\right])g(a).\end{aligned}$$

In this case, we want to prove something about $\mathrm{E}\left[e^{tX}\right]$, which is the expected value of a convex function, so replacing g this function and using

$$\begin{aligned}\mathrm{E}\left[\alpha\right] &= \frac{-a}{b-a} \\ \text{and } \ 1 - \mathrm{E}\left[\alpha\right] &= 1 - \frac{-a}{b-a} = \frac{b-a}{b-a} + \frac{a}{b-a} = \frac{b}{b-a},\end{aligned}$$

we get

$$\mathrm{E}\left[e^{tX}\right] \leq \frac{-a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta}. \tag{16}$$

Now, we want to relate $\mathrm{E}\left[e^{tX}\right]$ to a single exponential and currently we have two exponentials on the right hand side. So at this point, we need to go through some algebra to bring the right hand side into a form $e^{h(u)}$ for some $h(u)$ (where the variable u is simply a change of variable to make the expression a little nicer, because it gets complicated).

First, we can pull a factor of e^{ta} to get

$$\begin{aligned}\frac{-a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta} &= e^{ta}\left(-\frac{a}{b-a}e^{t(b-a)} + \frac{b}{b-a}\right) \\ &= e^{ta}e^{\log\left(-\frac{a}{b-a}e^{t(b-a)} + \frac{b}{b-a}\right)} \\ &= e^{\frac{a}{b-a}t(b-a) + \log\left(1 + \frac{a}{b-a} - \frac{a}{b-a}e^{t(b-a)}\right)}.\end{aligned}$$

To simplify this expression, we can make a change of variable of the form $u = t(b-a)$ and $\gamma = -a/(b-a)$. The expression in Equation 16 now becomes

$$\mathrm{E}\left[e^{tX}\right] \leq e^{-\gamma u + \log(1-\gamma+\gamma e^u)} = e^{h(u)}, \tag{17}$$

with

$$h(u) = -\gamma u + \log\left(1 - \gamma + \gamma e^u\right).$$

Now, by Taylor's theorem, there exists a point $w \in (0, u)$ such that

$$\begin{aligned}h(u) &= h(0) + h'(0)u + \frac{h''(w)}{2}u^2 \quad \text{(Taylor)} \\ &= \frac{h''(w)}{2}u^2 \qquad \text{(Because } h(0) = 0 \text{ and } h'(0) = 0\text{).}\end{aligned} \tag{18}$$

So we really just need to analyze the second derivative of h:

$$\begin{aligned} h''(u) &= \frac{d^2}{du^2}\left(-\gamma u + \log\left(1-\gamma+\gamma e^u\right)\right) = \frac{d}{du}\left(-\gamma + \frac{\gamma e^u}{1-\gamma+\gamma e^u}\right) \\ &= \frac{(1-\gamma+\gamma e^u)\gamma e^u - \gamma e^u \gamma e^u}{(1-\gamma+\gamma e^u)^2} \\ &= \frac{\gamma e^u}{1-\gamma+\gamma e^u} - \left(\frac{\gamma e^u}{1-\gamma+\gamma e^u}\right)^2 \\ &= \beta(1-\beta), \end{aligned}$$

where

$$\beta = \frac{\gamma e^u}{1-\gamma+\gamma e^u}.$$

This final expression in beta $f(\beta) = \beta(1-\beta)$ is an upside-down parabola with zeros at 0 and 1 and maximizer at $\frac{1}{2}$. It's maximal value is, therefore, $f(\frac{1}{2}) = \frac{1}{2}$.

Now, returning to Equation 18, we see that

$$h(\, t(b-a)\,) \leq \frac{1/4}{2}(\, t(b-a)\,)^2 = \frac{t^2(b-a)^2}{8}.$$

And, finally, plugging this bound into the bound in Equation 17, we get

$$\mathrm{E}\left[e^{tX}\right] \leq e^{h(\, t(b-a)\,)} \leq e^{t^2(b-a)^2/8}, \tag{19}$$

which is what we wanted to show. □