The Weak Law of Large Numbers

The WLLN is absolutely fundamental to machine learning (and really to all of probability and statistics). It basically formalizes the notion that given a series of independent samples of a random variable X, we can approximate $\mathrm{E}[X]$ by averaging the samples. The WLLN states that if $X_1, X_2, \ldots$ are independent copies of a random variable X,

$$\frac{1}{N}\sum_{n=1}^{N} X_n \to \mathrm{E}[X] \quad \text{as } N \to \infty.$$

The only condition for this convergence is that X has finite variance.

We start by stating the main result precisely. Let X be a random variable with pdf $f_X(x)$, mean $\mathrm{E}[X] = \mu$, and variance $\mathrm{var}(X) = \sigma^2 < \infty$. We observe *samples* of X labeled $X_1, X_2, \ldots, X_N$. The X_i are independent of one another, and they all have the same distribution as X. We will show that the sample mean formed from a sample of size N:

$$M_N = \frac{1}{N}(X_1 + X_2 + \cdots + X_N),$$

obeys[3]

$$\mathrm{P}(|M_N - \mu| > \epsilon) \;\leq\; \frac{\sigma^2}{N\epsilon^2},$$

where $\epsilon > 0$ is an arbitrarily small number. In the expression above, M_N is the only thing that is random; μ and σ^2 are fixed underlying properties of the distribution, N is the amount of data we see, and ϵ is something we can choose arbitrarily.

[3] This is a simple example of a *concentration bound*. It is not that tight; we will later counter inequalities of this type that are much more precise. But it is relatively simple and will serve our purpose here.

Notice that no matter how small ϵ is, the probability on the right hand side above goes to zero as $N \to \infty$. That is, for any fixed $\epsilon > 0$,

$$\lim_{N\to\infty} \mathrm{P}\left(|M_N - \mu| > \epsilon\right) = 0.$$

This result is follows from two simple but important tools known as the *Markov* and *Chebyshev* inequalities.

Markov inequality

Let X be a random variable that only takes positive values:

$$f_X(x) = 0, \quad \text{for} \;\; x < 0, \quad \text{or} \quad F_X(0) = 0.$$

Then

$$\boxed{\mathrm{P}\left(X \geq a\right) \;\leq\; \frac{\mathrm{E}\left[X\right]}{a} \quad \text{for all } a > 0.}$$

For example, the probability that X is more than 5 times its mean is 1/5, 10 times the mean is 1/10, etc. And this holds for **any distribution**.

The Markov inequality is easy to prove:

$$\begin{aligned}
\mathrm{E}\left[X\right] &= \int_0^\infty x f_X(x)\, dx \\
&\geq \int_a^\infty x f_X(x)\, dx \\
&\geq \int_a^\infty a f_X(x)\, dx \\
&= a \cdot \mathrm{P}\left(X \geq a\right)
\end{aligned}$$

and so $P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}$.

Again, this is a very general statement in that we have assumed nothing about X other than it is positive. The price for the generality is that the bound is typically very loose, and does not usually capture the behavior of $\mathrm{P}(X \geq a)$. We can, however, cleverly apply the Markov inequality to get something slightly more useful.

Chebyshev inequality

The main use of the Markov inequality turns out to be its use in deriving other, more accurate deviation inequalities. Here we will use it to derive the **Chebyshev inequality**, from which the weak law of large numbers will follow immediately.

Chebyshev inequality: If X is a random variable with mean μ and variance σ^2, then

$$\boxed{\mathrm{P}\left(|X-\mu| > c\right) \;\leq\; \frac{\sigma^2}{c^2} \quad \text{for all } c > 0.}$$

The Chebyshev inequality follows immediately from the Markov inequality in the following way. No matter what range of values X takes, the quantity $|X-\mu|^2$ is always positive. Thus

$$\mathrm{P}\left(|X-\mu|^2 > c^2\right) \;\leq\; \frac{\mathrm{E}\left[|X-\mu|^2\right]}{c^2} = \frac{\sigma^2}{c^2}.$$

Since squaring $(\cdot)^2$ is monotonic (invertible) over positive numbers,

$$\mathrm{P}\left(|X-\mu|^2 > c^2\right) = \mathrm{P}\left(|X-\mu| > c\right) \;\leq\; \frac{\sigma^2}{c^2}.$$

We now have a bound which depends on the mean and the variance of X; this leads to a more accurate approximation of the probability.

Simple proof of the weak law of large numbers

We now turn to the behavior of the the sample mean

$$M_N = \frac{X_1 + X_2 + \cdots + X_N}{N},$$

where again the X_i are iid random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{var}(X_i) = \sigma^2$. We know that

$$\mathrm{E}[M_N] = \frac{\mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_N]}{N} = \frac{N\mu}{N} = \mu,$$

and since the X_i are independent,

$$\mathrm{var}(M_N) = \frac{\mathrm{var}(X_1) + \mathrm{var}(X_2) + \cdots + \mathrm{var}(X_N)}{N^2} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}.$$

For any $\epsilon > 0$, a direct application of the Chebyshev inequality tells us that

$$\mathrm{P}(|M_N - \mu| > \epsilon) \;\leq\; \frac{\sigma^2}{N\epsilon^2}.$$

The point is that this gets arbitrarily small as $N \to \infty$ no matter what ϵ was chosen to be. We have established, in some sense, that even though $\{M_N\}_{N=1}^{\infty}$ is a sequence of random numbers, it converges to something deterministic, namely μ.

WLLN: Let $X_1, X_2, \ldots$ be iid random variables as above. For **every** $\epsilon > 0$, we have

$$\mathrm{P}\left(|M_N - \mu| > \epsilon\right) = \mathrm{P}\left(\left|\frac{X_1 + \cdots + X_N}{N} - \mu\right| > \epsilon\right) \longrightarrow 0,$$

as $N \to \infty$.

One of the philosophical consequences of the WLLN is that it tells us that probabilities can be estimated through **empirical frequencies**. Suppose I want to estimate the probability of and event A occurring related to some probabilistic experiment. We run a series of (independent) experiments, and set $X_i = 1$ if A occurred in experiment i, and $X_i = 0$ otherwise. Then given $X_1, \ldots, X_N$, we estimate the probability of A in a completely reasonable way, by computing the percentage of times it occurred:

$$p_{\text{empirical}} = \frac{X_1 + \cdots + X_N}{N}.$$

The WLLN tells us that

$$p_{\text{empirical}} \to \mathrm{P}(A) \quad \text{as} \quad N \to \infty.$$

This lends some mathematical weight to our interpretation of probabilities as *relative frequencies*.

All of the above of course applies to functions of random variables. That is, if X is a random variable, and $g(X)$ is a function of that random variable with

$$\mathrm{var}\left(g(X)\right) = \mathrm{E}\left[\left(g(X) - \mathrm{E}\left[g(X)\right]\right)^2\right] < \infty,$$

then given independent realizations $X_1, \ldots, X_N$, we have

$$\frac{1}{N} \sum_{n=1}^{N} g(X_n) \to \mathrm{E}\left[g(X)\right]$$

as $N \to \infty$.