Proof For Hoeffding's Inequality

Chieh T Wu

Northeastern University
January 22, 2021

What is the Hoeffding's Inequality. Let $X_1, ..., X_n$ be independent random variable. Assume each X_i is bounded at $a_i \leq X_i \leq b_i$. Let $\hat{\mu} = \frac{1}{n}\sum X_i$ be the sample mean. Then, the Hoeffding's Inequality states that

$$p(\hat{\mu} \geq \mathbb{E}[\hat{\mu}] + \epsilon] \leq e^{-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}. \tag{1}$$

This inequality is useful when we want to evaluate the quality of an approximation to its true value. Specifically, given a set of i.i.d random variables $X_1, ..., X_n$ with mean $\mu = \mathbb{E}[X]$, and average as $\hat{\mu} = \frac{1}{n}\sum X_i$, then Hoeffding's Inequality allow us to bound the error between them. This work aims to provide the necessary background and then, step by step, derive this inequality.

Markov's Inequality. At the most fundamental level, the Hoeffding's Inequality is really just a variation of the Markov's Inequaltiy. Given $Z \geq 0$ as a random variable, then Markov's Inequality states that

$$p(Z \geq \eta) \leq \frac{\mathbb{E}[Z]}{\eta}. \tag{2}$$

This inequality allows us to use a tail bound to bound the size of a random variable. Here, we specifically would like to bound the difference between $\hat{\mu}$ and μ. We start the proof by defining an indicator function $\eta\mathbb{I}_{Z\geq\eta}$. This is essentially a step function where

$$\eta\mathbb{I}_{Z\geq\eta} = \begin{cases} Z \geq \eta & \eta \\ Z < \eta & 0 \end{cases}. \tag{3}$$

In the 1st case where we have $Z \geq \eta$, note that

$$\eta\mathbb{I}_{Z\geq\eta} = \eta \quad \text{and} \quad Z \geq \eta. \tag{4}$$

Therefore, in the 1st case, the following statement must be true.

$$\eta\mathbb{I}_{Z\geq\eta} \leq Z. \tag{5}$$

In the 2nd case where we have $Z < \eta$, note that

$$\text{The inequality assumes that } Z \geq 0 \quad \text{and} \quad \eta\mathbb{I}_{Z\geq\eta} = 0. \tag{6}$$

Therefore, for the 2nd case, we can see that Inequality (5) is also true. Therefore, $\eta\mathbb{I}_{Z\geq\eta} \leq Z$ is a condition that is always true regardless of Z. If we next place an expectation across both terms, we get

$$\mathbb{E}[\eta\mathbb{I}_{Z\geq\eta}] \leq \mathbb{E}[Z] \tag{7}$$

$$\eta\int_0^\infty \mathbb{I}_{Z\geq\eta} \quad p(z)dz \leq \mathbb{E}[Z] \tag{8}$$

$$\eta\int_\eta^\infty p(z)dz \leq \mathbb{E}[Z] \tag{9}$$

$$\eta p(Z \geq \eta) \leq \mathbb{E}[Z] \tag{10}$$

$$p(Z \geq \eta) \leq \frac{\mathbb{E}[Z]}{\eta}. \tag{11}$$

Chebyshev's Inequality. Once we have obtained the Markov's Inequality, we can replace Z with any useful statistics. An example of this is the Chebyshev's Inequality. Specifically, we let $Z = (X - \mu)^2$ be the 2nd moment, we get

$$p(Z \geq \eta) \leq \frac{\mathbb{E}[Z]}{\eta} \tag{12}$$

$$p((X - \mu)^2 \geq \eta) \leq \frac{\mathbb{E}[(X - \mu)^2]}{\eta} \tag{13}$$

$$p(|X - \mu| \geq \sqrt{\eta}) \leq \frac{Var[X]}{\eta} \tag{14}$$

$$p(|X - \mu| \geq \epsilon) \leq \frac{Var[X]}{\epsilon^2}. \tag{15}$$

An important consequence of Chebyshev's inequality is that the average of the random variable with finite variance converges to their mean. We see this with via a simplified case where we assume that samples of Z are i.i.d, $\mathbb{E}[Z] = 0$, and $\bar{Z} = \frac{1}{n}\sum Z_i$, then

$$Var[\bar{Z}] = E\left[(\frac{1}{n}\sum Z_i)^2\right] \qquad \text{Definition of Variance with 0 mean} \tag{16}$$

$$= \frac{1}{n^2}\sum_{i,j}\mathbb{E}[Z_i Z_j] \qquad \text{Given i.i.d samples, the cross terms are 0} \tag{17}$$

$$= \frac{1}{n}\left[\frac{1}{n}\sum_{i}\mathbb{E}[Z_i^2]\right] \qquad \text{Aggregate the definition of variance again} \tag{18}$$

$$= \frac{Var[Z]}{n} \qquad \text{Now we know the variance for } \bar{Z}. \tag{19}$$

If we know the variance of $\bar{Z}$, then we can plug it back into Chebyshev's Inequality and obtain

$$p(|\bar{Z}| \geq \epsilon) \leq \frac{Var[Z]/n}{\epsilon^2} \tag{20}$$

$$p(|\bar{Z}| \geq \epsilon) \leq \frac{Var[Z]}{n\epsilon^2}. \tag{21}$$

This inequality is essentially the proof for the Weak Law of Large Numbers which given $\mu = 0$ essentially says that

$$\lim_{n\to\infty} p\left(\left|\frac{1}{n}\sum Z_i\right| \geq \epsilon\right) = 0. \tag{22}$$

Therefore, as $n \to \infty$, the average approach to the mean.

As a second point, note that a constant value bounds Markov's Inequality, i.e., $\frac{\mathbb{E}[Z]}{\eta}$. Therefore, it is a *fixed* bound that doesn't adapt to the sample size n, implying a potentially loose bound. In contrast, looking at Eq. (21), note that as n increases, the bound becomes tighter as it approaches 0. This is a much stronger condition than Markov's inequality because we become more certain of the inequality as n increases.

Here, we use Chebyshev's Inequality as an example of how we can form a tighter bound to achieve a stronger statement. In Chebyshev's case, as $n \to \infty$, the bound decrease at a rate of $1/n$. The idea of Hoeffding's Inequality also performs a change of variable for Z. However, instead of using the 2nd moment, Hoeffding came from using $Z = e^{tX}$. We will see in the next section that by setting $Z = e^{tX}$, it will allow us to achieve an even tighter bound, one that's not only decreasing linearly $(1/n)$, but one that decreases exponentially.

Moment Generating Functions. Why is setting $Z = e^{tX}$ special? When we looked at Chebyshev's inequality, notice that we used the 2nd moment as Z. The intuition is that a tighter bound can be achieved by using a higher-order moment. Instead of picking a particular higher moment, the moment generating function (MGF) is a function that allows us to pick any moment. In this section, we will briefly go over the idea of an MFG and use it to tie it back to Markov's Inequality.

For a random variable X, its moment generating function is

$$M_X(t) = \mathbb{E}[e^{tX}] = \int e^{tX} p(x) dx. \tag{23}$$

This function received its name because its kth derivative evaluated at $t = 0$ produces the kth moment. For example, we have

$$M'_X(t=0) = \frac{\partial}{\partial t}\int e^{tx}p(x)dx = \int xp(x)dx = \mathbb{E}[X], \tag{24}$$

$$M''_X(t=0) = \frac{\partial^2}{\partial t^2}\int e^{tx}p(x)dx = \int x^2p(x)dx = \mathbb{E}[X^2]. \tag{25}$$

An important property of MGF is that

- given $X_1, ..., X_n$ i.i.d random variables
- let $Y = X_1 + X_2 + ... + X_n$
- let MGF of $X_1, ..., X_n$ be $M_{X_1}(t), M_{X_2}(t), ..., M_{X_n}(t)$

then the MFG of Y becomes the product of individual MFG of X where

$$M_Y(t) = M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t). \tag{26}$$

This statement can be easily proven. If the new variable is $Y = X_1 + X_2 + ... + X_n$, then we have

Proof.

$$\begin{aligned}
M_Y(t) &= \mathbb{E}[e^{tY}] && && (27)\\
&= \mathbb{E}[e^{t(X_1+X_2+...+X_n)}] && && (28)\\
&= \int e^{t(x_1+x_2+...+x_n)}p(x_1, ..., x_n)dx_1...dx_n && \text{X is i.i.d} && (29)\\
&= \int e^{t(x_1+x_2+...+x_n)}p(x_1)...p(x_n)dx_1...dx_n && \text{Joint of i.i.d can be split} && (30)\\
&= \left[\int e^{tx_1}p(x_1)dx_1\right]...\left[\int e^{tx_n}p(x_n)dx_n\right] && && (31)\\
&= M_{X_1}(t)M_{X_2}(t)...M_{X_n}(t) = M_X(t)^n. && \text{proven} && (32)
\end{aligned}$$

□

Combining MGF and Markov's Inequality. Now that we have a basic understanding of MGF, we again return to the Markov's Inequality below.

$$p(Z \geq \eta) \leq \frac{\mathbb{E}[Z]}{\eta}. \tag{33}$$

If we let $Z = e^{tX}$ then we have

$$p\left(e^{tX} \geq \eta\right) \leq \frac{\mathbb{E}[e^{tX}]}{\eta} \tag{34}$$

$$p\left(X \geq \frac{log(\eta)}{t}\right) \leq \frac{M_X(t)}{\eta} \tag{35}$$

Let's now do a variable switch and let $\epsilon = \frac{log(\eta)}{t}$, then η can be written in terms of ϵ where

$$\epsilon = \frac{log(\eta)}{t} \tag{36}$$

$$e^{t\epsilon} = \eta. \tag{37}$$

If we put this new relationship back into Eq. (35), we have

$$p\left(X \geq \epsilon\right) \leq \frac{M_X(t)}{e^{t\epsilon}}. \tag{38}$$

This relationship is just for one single variable X. We can follow the same steps if instead we have a variable Y that is the summation of n i.i.d variables $X_1, X_2, ..., X_n$, i.e., $Y = X_1 + X_2 + ... + X_n$. We again plug this into the Markov's Inequality.

$$p(e^{t(X_1+X_2+...+X_n)} \geq \eta) \leq \frac{\mathbb{E}[e^{t(X_1+X_2+...+X_n)}]}{\eta} \tag{39}$$

$$p(X_1 + X_2 + ... + X_n \geq \frac{log(\eta)}{t}) \leq \frac{M_X(t)^n}{\eta} \tag{40}$$

Now, again we do a change of variable and set

$$n\epsilon = \frac{log(\eta)}{t} \tag{41}$$

, then we can also write η in terms of $n\epsilon$ to obtain

$$\eta = e^{t\epsilon n} = (e^{t\epsilon})^n. \tag{42}$$

Now if we plug Eq. (41) and (42) back into Eq. (40), we get

$$p(X_1 + X_2 + ... + X_n \geq n\epsilon) \leq \frac{M_X(t)^n}{(e^{t\epsilon})^n} \tag{43}$$

$$p\left(\frac{1}{n}\sum X_i \geq \epsilon\right) \leq \left(\frac{M_X(t)}{(e^{t\epsilon})}\right)^n \tag{44}$$

$$p\left(\bar{X} \geq \epsilon\right) \leq \left(\frac{M_X(t)}{(e^{t\epsilon})}\right)^n \tag{45}$$

At this point, notice that if $\mathbb{E}[X] = 0$, then the difference between $\bar{X}$ and $\mathbb{E}[X]$ is bounded by $\left(\frac{M_X(t)}{(e^{t\epsilon})}\right)^n$. Therefore, as long as we set the moment such that

$$\left(\frac{M_X(t)}{(e^{t\epsilon})}\right) < 1, \tag{46}$$

then $\left(\frac{M_X(t)}{(e^{t\epsilon})}\right)^n$ decays toward 0 at an exponential rate as $n \to \infty$. Remember that Markov's inequality was a fix bound, and the Chebyshev's Inequality was a linear bound. Therefore, the setting of $Z = e^{tX}$ in Markov's Inequality is the key to produce this tight bound.

Hoeffding's Inequality. Having already achieved exponential decay with Eq. (45), how else can we further improve upon this bound? Looking at the inequality, notice the $M_X(t)$ is a function of the random variable X. Therefore, depending on the distribution of X, $M_X(t)$ would need to be computed each time separately. Instead of requiring a different inequality for each distribution, it would be natural to ask if there is a bound that automatically work for a wide class of distribution?

It turns out that the class of sub-Gaussian distribution is ideal for this case. This realization is based on Hoeffding's Lemma. The key takeaway is that *any distribution* that has the following 2 characteristics is a sub-Gaussian.

- with a probability of 1, X is bounded between $[a, b]$
- $\mathbb{E}[X] = 0$

Hoeffding's Inequality is based on Eq. (45), except it assumes that X belongs to the sub-Gaussian distribution, resulting in a single bound that automatically applies to a wide range of distributions. The concept of sub-Gaussian distributions starts with the Gaussian distribution itself. We first compute the MGF for a Gaussian distribution of 0 means of X below.

$$\mathbb{E}[e^{tX}] = \int_{\mathcal{X}} e^{tx}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}dx \tag{47}$$

$$= \int_{\mathcal{X}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}+\frac{2\sigma^2 tx}{2\sigma^2}}dx \tag{48}$$

$$= \int_{\mathcal{X}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2-2\sigma^2 tx}{2\sigma^2}}dx \tag{49}$$

$$= \int_{\mathcal{X}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2-2\sigma^2 tx+\sigma^4t^2-\sigma^4t^2}{2\sigma^2}}dx \qquad \text{Complete the square} \tag{50}$$

$$= \int_{\mathcal{X}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\sigma^2 t)^2}{2\sigma^2}}e^{\frac{\sigma^4t^2}{2\sigma^2}}dx \tag{51}$$

$$= e^{\frac{\sigma^2t^2}{2}}\int_{\mathcal{X}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\sigma^2 t)^2}{2\sigma^2}}dx \qquad \text{The total probability is 1} \tag{52}$$

$$= e^{\frac{\sigma^2t^2}{2}} \qquad \text{The MGF for a Gaussian} \tag{53}$$

Once we have identified the MFG for a Guassian distribution, a **sub-Gaussian** is simply *any* distribution that has a MGF that is upper bounded by the MGF of a Gaussian. That is, a random variable X is a sub-Gaussian if

$$M_X(t) \leq e^{\frac{\sigma^2t^2}{2}}. \tag{54}$$

Like any other distribution, sub-Gaussian is parameterized by σ^2. Therefore, by adjusting σ^2, various sub-Gaussian can be achieved. Given a better undstanding of a sub-Gaussian, we can now tie Hoeffding's Lemma into the story.

Lemma 1. *Let X be any real-valued random variable with expected value $\mathbb{E}[X] = \mu$ such that $a \leq X \leq b$ with a probability of 1 (almost surely). Then, for all $t \in \mathbb{R}$*

$$\mathbb{E}[e^{tX}] \leq e^{t\mu + \frac{\lambda^2 (b-a)^2}{8}}. \tag{55}$$

Looking at Eq. (55), notice that if $\mathbb{E}[X] = 0$ then $\mu = 0$. Also notice that if we set $\sigma^2 = \frac{(b-a)^2}{4}$, then we get the identical bound as the sub-Gaussian.

$$e^{t\mu + \frac{t^2 (b-a)^2}{8}} \to e^{\frac{t^2 \sigma^2}{2}} \tag{56}$$

We can conclude from this observation that *any* distribution that satisfy the Hoeffding's Lemma is a sub-Gaussian, and is upper bounded by the MGF of a Gaussian distribution. We can now combine Eq. (38) with the MGF of a Gaussian distribution.

$$p\left(X \geq \epsilon\right) \leq \frac{M_X(t)}{e^{t\epsilon}} \qquad \text{let } M_X(t) = e^{\frac{t^2\sigma^2}{2}} \tag{57}$$

$$\leq e^{\frac{t^2\sigma^2}{2} - t\epsilon} \tag{58}$$

$$\leq e^{\frac{t^2\sigma^2 - 2t\epsilon}{2}}. \tag{59}$$

At this point, we need to choose a moment t such that the bound is minimized. To do this, we take the derivative of the exponent and set it to 0.

$$\frac{\partial}{\partial t} \frac{t^2\sigma^2 - 2t\epsilon}{2} = 0 \tag{60}$$

$$t\sigma^2 - \epsilon = 0 \tag{61}$$

$$t\sigma^2 = \epsilon \tag{62}$$

$$t = \epsilon/\sigma^2 \tag{63}$$

From this derivation, we see that the optimal moment t that generates the tightest bound is $t = \epsilon/\sigma^2$. We will soon combine this optimal t with the sub-Gaussian assumption. We emphasize that a random variable X is a sub-Guassian when $\sigma^2 = \frac{(b-a)^2}{4}$, or

$$X \text{ is sub-Gaussian} \implies \text{It has parameters } \sigma^2 = \frac{(b-a)^2}{4}. \tag{64}$$

But instead of X, we are more interested in the random variable $Y = \frac{1}{n}\sum X_i - E[X]$. Conveniently, it turns out to also be a sub-Gaussian if we set $\sigma^2 = \frac{1}{n^2}\sum_{i=1}^n \frac{(b_i - a_i)^2}{4}$.

$$Y \text{ is also sub-Gaussian} \implies \text{It has parameters } \sigma^2 = \frac{1}{n^2}\sum_{i=1}^n \frac{(b_i - a_i)^2}{4}. \tag{65}$$

Given this knowledge, we can now derive Hoeffding's Inequality. We start with Eq. (38).

$$p\left(X \geq \epsilon\right) \leq \frac{M_X(t)}{e^{t\epsilon}}. \tag{66}$$

The random variable we care about is not X, but $Y = \frac{1}{n}\sum X_i - E[X]$ with $\sigma^2 = \frac{1}{n^2}\sum_{i=1}^n \frac{(b_i - a_i)^2}{4}$. Therefore, we replace X with Y along with its corresponding MGF.

$$p\left(Y \geq \epsilon\right) \leq \frac{M_Y(t)}{e^{t\epsilon}} \tag{67}$$

$$p\left(\frac{1}{n}\sum X_i - E[X] \geq \epsilon\right) \leq \frac{e^{\frac{t^2\sigma^2}{2}}}{e^{t\epsilon}} \qquad \text{let } M_Y(t) = e^{\frac{t^2\sigma^2}{2}} \tag{68}$$

$$p\left(\frac{1}{n}\sum X_i - E[X] \geq \epsilon\right) \leq e^{\frac{t^2\sigma^2}{2} - \frac{2t\epsilon}{2}} \qquad \text{Replace } t \text{ with optimal } t = \epsilon/\sigma^2 \tag{69}$$

$$p\left(\frac{1}{n}\sum X_i - E[X] \geq \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\sigma^2}} \qquad \text{Replace } \sigma^2 = \frac{1}{n^2}\sum_{i=1}^n \frac{(b_i - a_i)^2}{4} \tag{70}$$

$$p\left(\frac{1}{n}\sum X_i - E[X] \geq \epsilon\right) \leq e^{-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}} \qquad \text{Hoeffding's Inequality emerges.} \tag{71}$$

$$\tag{72}$$