Tools for Bounding Probabilities

Dimitris Diochnos

University of Oklahoma
Fall 2020

Abstract

These notes were created for CS 6501 - Learning Theory at the University of Virginia during the Fall of 2015. The primary scope of the notes is the exposition of the Chernoff bounds as well as the Hoeffding bound as these are common tools used in Computational Learning Theory, and more broadly in the analysis of randomized algorithms.

1 Introduction

Proposition 1 (Union Bound). *Let $Y_1, Y_2, \ldots, Y_S$ be S events in a probability space. Then,*

$$\mathbf{Pr}\left(\bigcup_{j=1}^{S} Y_j\right) \leqslant \sum_{j=1}^{S} \mathbf{Pr}\left(Y_j\right) \ .$$

The inequality is equality for disjoint events Y_j.

Proposition 2 (Markov's Inequality). *Any non-negative random variable X satisfies*

$$\mathbf{Pr}\left(X \geqslant \alpha\right) \leqslant \frac{\mathbf{E}\left[X\right]}{\alpha}, \qquad \forall \alpha > 0\,.$$

Proposition 3 (Chebyshev's Inequality). *Let X be a random variable with expected value μ and variance σ^2. Then,*

$$\mathbf{Pr}\left(|X - \mu| \geqslant \alpha\right) \leqslant \frac{\sigma^2}{\alpha^2}, \qquad \forall \alpha > 0\,.$$

Remark 1 (Chebyshev vs. Markov). *The Chebyshev inequality tends to give better bounds than the Markov inequality, because it also uses information on the variance of X.*

Theorem 1 (Weak Law of Large Numbers). *Let $X_1, \ldots, X_N$ be a sequence of* independent identically distributed *random variables, with expected value μ. For every $\epsilon > 0$:*

$$\mathbf{Pr}\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \geqslant \epsilon\right) \to 0, \qquad \text{as } N \to \infty \tag{1}$$

Proof. Let $X_1, \ldots, X_N$ be a sequence of *independent identically distributed* random variables, with expected value μ and variance σ^2. Define the random variable $Y = \frac{1}{N}\sum_{i=1}^{N} X_i$. By linearity of expectation we get $\mathbf{E}\left[Y\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}\left[X_i\right] = \mu$. Since all the X_i are independent, the variance is $\mathbf{Var}\left[Y\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbf{Var}\left[X_i\right] = \frac{\sigma^2}{N}$. We now apply Chebyshev's inequality and obtain $\mathbf{Pr}\left(|Y - \mu| \geqslant \epsilon\right) \leqslant \frac{\sigma^2}{N\epsilon^2}$, for any $\epsilon > 0$. ■

1.1 Concentration and Tail Inequalities

In this section we examine a series of tools for estimating the concentration and bounding the probability of the tails. References for further reading are the following: for the Chernoff bounds please refer to [1] or to [5]; for the Hoeffding bound please refer to [4] or to [2].

Remark 2 (Bernoulli/Binomial Trials vs Poisson Trials). *A sequence of Bernoulli/binomial trials is a sequence of coin tosses of the same (biased) coin, while Poisson trials is a sequence of coin tosses of potentially different (biased) coins.*

Proposition 4 (Chernoff Bound for Upper Tail). *Assume $X_1, X_2, \dots, X_t$ are independent Poisson trials. Let $X = \sum_{i=1}^{t} X_i$, and $\mu = \mathbf{E}[X]$. Then, for $\gamma \in (0,1)$ it holds*

$$\mathbf{Pr}(X > (1+\gamma)\mu) \leqslant e^{-\mu\gamma^2/3} \, .$$

Proposition 5 (General Chernoff Bound for Upper Tail). *Assume $X_1, X_2, \dots, X_t$ are independent Poisson trials. Let $X = \sum_{i=1}^{t} X_i$, and $\mu = \mathbf{E}[X]$. Then, for $\gamma \geqslant 0$ it holds*

$$\mathbf{Pr}(X > (1+\gamma)\mu) \leqslant e^{-\mu\gamma^2/(2+\gamma)} \, .$$

Proposition 6 (Chernoff Bound for Lower Tail). *Assume $X_1, X_2, \dots, X_t$ are independent Poisson trials. Let $X = \sum_{i=1}^{t} X_i$, and $\mu = \mathbf{E}[X]$. Then, for $\gamma \in (0,1)$ it holds*

$$\mathbf{Pr}(X < (1-\gamma)\mu) \leqslant e^{-\mu\gamma^2/2} \, .$$

Proposition 7 (Hoeffding Bound). *Let $X_1, \dots, X_R$ be R independent random variables, each taking values in the range $\mathfrak{I} = [\alpha, \beta]$. Let $X = \frac{1}{R}\sum_{i=1}^{R} X_i$ and $\mu = \mathbf{E}[X]$ denote the mean of their expectations. Then,*

$$\mathbf{Pr}(|X - \mu| \geqslant \epsilon) \leqslant 2e^{-2R\epsilon^2/(\beta-\alpha)^2} \, .$$

2 Examples

Examples with (biased) coins are our best friends on understanding the bounds mentioned earlier. For more examples see [3, 5].

Example 1 (Fair Coin Tossing). *We toss a fair coin 100 times and 80 times we observe* $\mathtt{H}$*. What is the probability of this event?*

Solution. Let $X = \sum_{i=1}^{100} X_i$ be the number of times that we observed $\mathtt{H}$, where the X_i's are indicator random variables indicating whether we obeserved $\mathtt{H}$ or not on the i-th trial. Note that the expectation is $\mathbf{E}[X] = Np = 100 \cdot (1/2) = 50$. Also note that $\mathbf{Var}[X] = Np(1-p) = 100 \cdot (1/2) \cdot (1/2) = 25$. A direct computation for the probability p of such an event gives $p = \binom{100}{80} \cdot 2^{-100} \approx 4.2 \cdot 10^{-10}$.

- Markov's inequality yields $\mathbf{Pr}(X \geqslant 80) \leqslant 50/80 = 0.625$.
- Chebyshev yields $\mathbf{Pr}(|X - 50| \geqslant 30) \leqslant \frac{25}{30^2} = 2.7 \cdot 10^{-2}$.
- Note that setting $\gamma = 29/50 = 0.58$ we have that $(1+\gamma) \cdot \mu = 79$. Thus, the Chernoff bound for the upper tail gives $\mathbf{Pr}(X > 79) \leqslant e^{-50 \cdot (0.58)^2/3} \leqslant e^{-5.6} \leqslant 3.7 \cdot 10^{-3}$.
- Let $Y = X/100$. The Hoeffding bound gives $\mathbf{Pr}(|Y - 0.5| \geqslant 0.3) \leqslant 2e^{-200 \cdot 0.09} \leqslant 3.2 \cdot 10^{-8}$. ■

Before we jump into conclusions as to which bound is preferable, let us try the following example.

Example 2 (Biased Coin). *We toss a coin 1000 times and we observe 100 times* $\mathtt{H}$*. Give an upper bound on the true probability* p *that the coin has for bringing up* $\mathtt{H}$ *with* 95% *confidence.*

Solution. We have that empirically the probability of observing $\mathtt{H}$ is 0.1. This is the best estimate that we have for the true probability that this coin will give $\mathtt{H}$ in one particular trial.

Let p be the true probability that this coin has for bringing up $\mathtt{H}$ in one coin toss. Then, we would like to examine the probability $\mathbf{Pr}\left(X < 101\right)$. For this reason we will use the Chernoff bound for the lower tail. We have $\mu = 1000 \cdot p$ and we want $(1-\gamma)\cdot p \cdot 1000 = 101 \Rightarrow \gamma = \frac{p-0.101}{p}$. Substituting to Proposition 6 we get $\mathbf{Pr}\left(X < 101\right) \leqslant e^{-1000p\frac{(p-0.101)^2}{2p^2}}$, which we require to be upper bounded by $\delta = 0.05$ since this is our failure probability. Thus, we want $e^{-500p\frac{(p-0.101)^2}{p^2}} \leqslant 0.05 \Rightarrow \ldots \Rightarrow p^2 - 0.208p + (0.101)^2 \geqslant 0$. Hence, for $p \geqslant 0.129$ the Chernoff bound indicates that we would observe less than 101 $\mathtt{H}$ in our experiment with probability at most 5%. In other words, we can say that the coin has true probability of bringing up heads not more than 0.129 with confidence 95%.

Now, let's try to answer the same question using the Hoeffding bound. By Proposition 7 we want $\mathbf{Pr}\left(|0.1-\mu|\right) \leqslant 2e^{-2000\varepsilon^2} \leqslant 0.05 \Rightarrow \ldots \Rightarrow \varepsilon \geqslant \sqrt{\frac{\ln(40)}{2000}} \approx 0.04295$. In other words if we allow $\varepsilon \geqslant 0.043$ with probability at least 0.95 the absolute difference $|0.1-\mu|$ is at most ε. Thus $\mu \leqslant 0.143$ and since $\mu = 1000p/1000 = p$ we have that $p \leqslant 0.143$ with confidence at least 95%. ■

Remark 3 (Chernoff or Hoeffding?). *Note that in the second example the Hoeffding bound gives a worse bound. While this is true, the Hoeffding bound in the second example also states that based on the experiment that we conducted, with confidence at least* 95%*, the true probability that the coin has for bringing up* $\mathtt{H}$ *is in the interval* $[0.057, 0.143]$*. Thus, typically, when we are interested in concentration we use the Hoeffding bound, while if we are interested in only one-sided bounds, we tend to prefer the Chernoff bound.*

2.1 Towards the Double Sample Argument

Lemma 1. *Let* $p \geqslant \varepsilon > 0$*. Then,* $4p - 4\varepsilon + \varepsilon^2/p \geqslant p$*.*

Proof. Set $Q(p) = 3p^2 - 4\varepsilon p + \varepsilon^2$. The discriminant of Q is $\Delta = 16\varepsilon^2 - 12\varepsilon^2 = (2\varepsilon)^2$. Thus, the two roots of Q are $p_1 = \frac{2\varepsilon}{6} = \varepsilon/3$ and $p_2 = \frac{6\varepsilon}{6} = \varepsilon$. Hence, $Q(p) \geqslant 0$ when p is not in the interval $(\varepsilon/3, \varepsilon)$ and since $p \geqslant \varepsilon$ the claim follows. ■

Lemma 2. *Given a coin that succeeds with probability* $p \geqslant \varepsilon > 0$*, it holds that after* $m \geqslant 8/\varepsilon$ *trials the number of successes is not less than* $\varepsilon \cdot m/2$ *with probability at least* $1/2$*.*

Proof. We want to show that

$$\mathbf{Pr}\left(X < \frac{\varepsilon m}{2}\right) \leqslant \frac{1}{2}.$$

In order to do that we will use Proposition 6. For the given coin we have $\mu = \mathbf{E}\left[X\right] = pm \geqslant \varepsilon m$. We want $(1-\gamma)pm = \varepsilon m/2$. Thus, we set $\gamma = 1 - \frac{\varepsilon}{2p}$ and so by Proposition 6 we obtain

$$\begin{aligned}
\mathbf{Pr}\left(X < \tfrac{\varepsilon m}{2}\right) &\leqslant e^{-pm\left(1-\frac{\varepsilon}{2p}\right)^2/2} \\
&= e^{-\frac{1}{8}\cdot\left(4p-4\varepsilon+\varepsilon^2/p\right)\cdot m} \\
&\leqslant e^{-\frac{1}{8}\cdot p\cdot m} && \text{(Lemma 1)} \\
&\leqslant e^{-\varepsilon\cdot m/8} && \text{(since } \varepsilon \leqslant p\text{)} \\
&\leqslant e^{-\frac{\varepsilon}{8}\cdot\frac{8}{\varepsilon}} && \text{(since } 8/\varepsilon \leqslant m\text{)} \\
&= e^{-1} \\
&\leqslant 1/2
\end{aligned}$$

Recall that $\ln(2) \simeq 0.693 < 0.7$ and note that we could accomplish the same guarantee with *only* $m = \lceil 8\ln(2)/\varepsilon \rceil < \lceil 5.6/\varepsilon \rceil < \lceil 8/\varepsilon \rceil$ coin flips. For our purposes it will not matter, as m will be larger than $\lceil 8/\varepsilon \rceil$ anyway. ■

Exercise 1. *Prove Lemma 2 using Chebyshev's inequality (Proposition 3).*

References

[1] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[2] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

[3] Torben Hagerup and Christine Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33(6):305–308, February 1990.

[4] Wassily Hoeffding. Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[5] Michael Mitzenmacher and Eli Upfal. *Probability and Computing - Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, 2005.