Basic Tools and Techniques for Algorithms in Learning Theory

Dimitris Diochnos

Created: Spring 2016
Revised: Fall 2017, Fall 2020

Abstract

These notes were created for CS 4710 - Artificial Intelligence at the University of Virginia during the Spring of 2016. The primary scope of the notes is the exposition of Markov's inequality and Chebyshev's inequality, as these are common tools used in learning theory and more broadly in the analysis of randomized algorithms.

Revision (Fall 2017). The document was revised during the Fall of 2017 as I realized the existence of a particular line of work. See Section 4 for details; this section was added in the Fall of 2017.

Revision (Fall 2020). The document was revised during the Fall of 2020 when I was teaching CS 5970 – Computational Learning Theory at the University of Oklahoma. I am now simplifying the presentation in the historical remarks section as there is no reason to refer to connections to a homework assignment that was used in the Artificial Intelligence class that was taught at UVA.

1 Background from Probability Theory

Definition 1 (Sample Space)**.** The *sample space* of an experiment (or random trial) is the set of all possible outcomes (or results) of that experiment.

Definition 2 (Random Variable)**.** A random variable over a sample space Ω is a function that maps every sample point (that is, outcome) to a real number.

Definition 3 (Independent Random Variables)**.** Let Ω be a sample space. Two random variables X and Y are independent, if for all $x, y \in \Omega$,

$$\mathbf{Pr}\left(X = x \wedge Y = y\right) = \mathbf{Pr}\left(X = x\right) \cdot \mathbf{Pr}\left(Y = y\right) .$$

An alternate definition is the following.

Definition 4 (Independent Random Variables)**.** Let Ω be a sample space. Two random variables X and Y are independent, if for all $x, y \in \Omega$, such that $\mathbf{Pr}\left(Y = y\right) \neq 0$, we have

$$\mathbf{Pr}\left(X = x | Y = y\right) = \mathbf{Pr}\left(X = x\right) \cdot \mathbf{Pr}\left(Y = y\right) .$$

Proposition 1 (Expectation of Independent Random Variables)**.** *Let* X *and* Y *be two* independent *random variables. Then,*

$$\mathbf{E}\left[XY\right] = \mathbf{E}\left[X\right] \cdot \mathbf{E}\left[Y\right] .$$

Remark 1. *Note that Proposition 1 is really a property of the expectation operator for* independent *random variables. In general, Proposition 1 is* not true. *Actually, it is* covariance *(see Lemma 7) that measures by how much Proposition 1* fails.

Proposition 2 (Union Bound). *Let* $Y_1, Y_2, \ldots, Y_T$ *be* T *events in a probability space. Then,*

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

The inequality is equality for disjoint events Y_j.

1.1 Markov's Inequality

Theorem 3 (Markov's Inequality). *Any non-negative random variable* X *satisfies*

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

Proof. For any event E, let I_E be the indicator random variable of E; that is,

$$I_E = \begin{cases} 1 & , \text{ if } E \text{ occurs,} \\ 0 & , \text{ otherwise} \end{cases}$$

Let X be a non-negative random variable. We look at the event $E = (X \geqslant \alpha)$. In other words $I_{(X \geqslant \alpha)} = 1$, if $X \geqslant \alpha$ and $I_{(X \geqslant \alpha)} = 0$, if $X < \alpha$. But for $\alpha > 0$, it holds

$$\alpha I_{(X \geqslant \alpha)} \leqslant X .$$

We now take the expectation of the last inequality and we have

$$\mathbf{E}\left[\alpha I_{(X \geqslant \alpha)}\right] \leqslant \mathbf{E}\left[X\right] .$$

However, $\mathbf{E}\left[\alpha I_{(X \geqslant \alpha)}\right] = \alpha \mathbf{E}\left[I_{(X \geqslant \alpha)}\right] = \alpha \cdot (1 \cdot \mathbf{Pr}\left(X \geqslant \alpha\right) + 0 \cdot \mathbf{Pr}\left(X < \alpha\right)) = \alpha \cdot \mathbf{Pr}\left(X \geqslant \alpha\right)$. ■

1.2 Variance and Chebyshev's Inequality

Definition 5 (Variance). Let X be a random variable with expectation $\mu = \mathbf{E}\left[X\right]$. The *variance* $\sigma_X^2 = \mathbf{Var}\left[X\right]$ of X is defined to be

$$\sigma_X^2 = \mathbf{Var}\left[X\right] = \mathbf{E}\left[(X - \mu)^2\right] .$$

Remark 2 (Intuition on Variance). *Variance measures how far a set of observations are spread out.*

Proposition 4 (Variance). *For a random variable* X *with expectation* $\mu = \mathbf{E}\left[X\right]$, *we have*

$$\mathbf{Var}\left[X\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2 .$$

Proof. From Definition 5 we have

$$\begin{aligned}\mathbf{Var}[X] &= \mathbf{E}\left[(X-\mu)^2\right] \\ &= \mathbf{E}\left[X^2 - 2\mu X + \mu^2\right] \\ &= \mathbf{E}\left[X^2\right] - \mathbf{E}\left[2\mu X\right] + \mathbf{E}\left[\mu^2\right] \\ &= \mathbf{E}\left[X^2\right] - 2\mu\mathbf{E}[X] + \mu^2 \\ &= \mathbf{E}\left[X^2\right] - \mu^2 \quad \blacksquare\end{aligned}$$

Theorem 5 (Chebyshev's Inequality). *Let X be a random variable with expected value μ and variance σ^2. Then,*

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

Proof. Define the random variable $Y = (X-\mu)^2$.

Note that $\mathbf{E}[Y] = \mathbf{E}\left[(X-\mu)^2\right] = \mathbf{Var}[X]$.

Then, for $\alpha > 0$, the way we have defined Y, the event $(Y \geqslant \alpha^2)$ is the same as the event $(X-\mu)^2 \geqslant \alpha^2 \Leftrightarrow |X-\mu| \geqslant \alpha$. Thus, $\mathbf{Pr}\left(Y \geqslant \alpha^2\right) = \mathbf{Pr}\left(|X-\mu| \geqslant \alpha\right)$.

Clearly, again by the definition of Y, Y is non-negative. So, by Markov's inequality (Theorem 3),

$$\mathbf{Pr}\left(Y \geqslant \alpha^2\right) \leqslant \frac{\mathbf{E}[Y]}{\alpha^2} = \frac{\mathbf{Var}[X]}{\alpha^2}\,. \quad \blacksquare$$

Remark 3 (Chebyshev vs. Markov). *The Chebyshev inequality tends to give better bounds than the Markov inequality, because it also uses information about the variance of X.*

1.3 Covariance

Definition 6 (Covariance). Let X and Y be two jointly distributed random variables with finite variances. Then, the *covariance* of X and Y, $\mathbf{Cov}[X,Y]$, is defined to be

$$\mathbf{Cov}[X,Y] = \mathbf{E}\left[(X-\mathbf{E}[X])\cdot(Y-\mathbf{E}[Y])\right]\,.$$

Remark 4 (Intuition on Covariance). *Covariance is a measure of how much two random variables change together.*

Corollary 6 (Variance Seen as Covariance). *Let X be a random variable. Then,*

$$\mathbf{Var}[X] = \mathbf{Cov}[X,X]\,.$$

Proof. By Definition 6, $\mathbf{Cov}[X,X] = \mathbf{E}\left[(X-\mathbf{E}[X])\cdot(X-\mathbf{E}[X])\right] = \mathbf{E}\left[(X-\mathbf{E}[X])^2\right]$. However, by Definition 5, $\mathbf{Var}[X] = \mathbf{E}\left[(X-\mathbf{E}[X])^2\right]$. $\blacksquare$

Lemma 7 (Covariance). *Let X and Y be two jointly distributed random variables with finite variances. Then,*

$$\mathbf{Cov}[X,Y] = \mathbf{E}[X\cdot Y] - \mathbf{E}[X]\cdot\mathbf{E}[Y]\,.$$

Proof. By Definition 6,

$$\begin{aligned}
\mathbf{Cov}[X,Y] &= \mathbf{E}[(X-\mathbf{E}[X])\cdot(Y-\mathbf{E}[Y])] \\
&= \mathbf{E}[X\cdot Y - X\cdot\mathbf{E}[Y] - \mathbf{E}[X]\cdot Y + \mathbf{E}[X]\cdot\mathbf{E}[Y]] \\
&= \mathbf{E}[X\cdot Y] - \mathbf{E}[X\cdot\mathbf{E}[Y]] - \mathbf{E}[\mathbf{E}[X]\cdot Y] + \mathbf{E}[\mathbf{E}[X]\cdot\mathbf{E}[Y]] \\
&= \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y] - \mathbf{E}[X]\,\mathbf{E}[Y] + \mathbf{E}[X]\,\mathbf{E}[Y] \\
&= \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y] \quad \blacksquare
\end{aligned}$$

Corollary 8 (Covariance of Two Independent Random Variables)**.** *Let* X *and* Y *be two* independent *random variables. Then,*

$$\mathbf{Cov}[X,Y] = 0\,.$$

Proof. When X and Y are independent, by Proposition 1, $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$. But then, by Lemma 7,

$$\mathbf{Cov}[X,Y] = \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y] = \mathbf{E}[X]\,\mathbf{E}[Y] - \mathbf{E}[X]\,\mathbf{E}[Y] = 0\,. \quad \blacksquare$$

Remark 5 (Uncorrelated Random Variables)**.** *Note that when* $\mathbf{Cov}[X,Y] = 0$*, we say that* X *and* Y *are* uncorrelated*. So,* independent *variables are a special case of* uncorrelated *variables. We will not occupy ourselves further with uncorrelated variables though.*

1.4 Revisiting Variance

Lemma 9 (Variance of a Linear Combination of Two Random Variables)**.** *Let* $a, b \in \mathbb{R}$*. Let* X *and* Y *be two random variables. Then,*

$$\begin{cases} \mathbf{Var}[aX+bY] &= a^2\mathbf{Var}[X] + b^2\mathbf{Var}[Y] + 2ab\mathbf{Cov}[X,Y] \\ \mathbf{Var}[aX-bY] &= a^2\mathbf{Var}[X] + b^2\mathbf{Var}[Y] - 2ab\mathbf{Cov}[X,Y] \end{cases}$$

Proof. By Proposition 4,

$$\begin{aligned}
\mathbf{Var}[aX\pm bY] &= \mathbf{E}\left[(aX\pm bY)^2\right] - (\mathbf{E}[aX\pm bY])^2 \\
&= \mathbf{E}\left[a^2X^2 + b^2Y^2 \pm 2aXbY\right] - (\mathbf{E}[aX]\pm\mathbf{E}[bY])^2 \\
&= \mathbf{E}\left[a^2X^2\right] + \mathbf{E}\left[b^2Y^2\right] \pm \mathbf{E}[2abXY] - (a\mathbf{E}[X]\pm b\mathbf{E}[Y])^2 \\
&= a^2\mathbf{E}\left[X^2\right] + b^2\mathbf{E}\left[Y^2\right] \pm 2ab\mathbf{E}[XY] - \left(a^2\mathbf{E}[X]^2 + b^2\mathbf{E}[Y]^2 \pm 2ab\mathbf{E}[X]\,\mathbf{E}[Y]\right) \\
&= a^2\cdot\left(\mathbf{E}\left[X^2\right] - \mathbf{E}[X]^2\right) + b^2\cdot\left(\mathbf{E}\left[Y^2\right] - \mathbf{E}[Y]^2\right) \pm 2ab\cdot(\mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y]) \\
&= a^2\mathbf{Var}[X] + b^2\mathbf{Var}[Y] \pm 2ab\mathbf{Cov}[X,Y] \quad \blacksquare
\end{aligned}$$

Corollary 10 (Variance of Sum of Two Random Variables)**.** *Let* X *and* Y *be two random variables. Then,*

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}[X,Y]\,.$$

Proof. We use Lemma 9 with $a = b = 1$. $\blacksquare$

Corollary 11 (Variance of Sum of Two Independent Random Variables)**.** *Let* X *and* Y *be two indepedent random variables. Then,*

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]\,.$$

Proof. By Corollary 10, $\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y] + 2\mathbf{Cov}[X,Y]$. However, since X and Y are independent, by Corollary 8, $\mathbf{Cov}[X,Y] = 0$. ■

Theorem 12 (Variance of Sum of Scaled Independent Random Variables). *Let $Y_1, Y_2, \ldots, Y_T$ be T pairwise independent random variables and let $a_1, a_2, \ldots, a_T$ be T positive real constants. Further, let $X = \sum_{i=1}^{T}(a_i \cdot Y_i)$. Then,*

$$\mathbf{Var}[X] = \sum_{i=1}^{T}\left(a_i^2 \cdot \mathbf{Var}[Y_i]\right) .$$

Proof. First note that

$$\mu_X = \mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{T} a_i Y_i\right] = \sum_{i=1}^{T}\mathbf{E}[a_i Y_i] = \sum_{i=1}^{T} a_i \mathbf{E}[Y_i] = \sum_{i=1}^{T} a_i \mu_{Y_i} .$$

We have,

$$\begin{aligned}
\mathbf{Var}[X] &= \mathbf{E}\left[(X-\mu_X)^2\right] && \text{(Definition 5)}\\
&= \mathbf{E}\left[\left(\textstyle\sum_{i=1}^{T}(a_i Y_i) - \sum_{j=1}^{T}(a_i \mu_{Y_j})\right)^2\right] && \text{(given)}\\
&= \mathbf{E}\left[\left(\textstyle\sum_{i=1}^{T} a_i Y_i - \sum_{i=1}^{T} a_i \mu_{Y_i}\right)^2\right] && \text{(use same index)}\\
&= \mathbf{E}\left[\left(\textstyle\sum_{i=1}^{T} a_i \left(Y_i - \mu_{Y_i}\right)\right)^2\right] && \text{(merge the sums)}\\
&= \mathbf{E}\left[\left(\textstyle\sum_{i=1}^{T} a_i \left(Y_i - \mu_{Y_i}\right)\right) \cdot \left(\sum_{j=1}^{T} a_j \left(Y_j - \mu_{Y_j}\right)\right)\right] && \text{(rewrite the square)}\\
&= \mathbf{E}\left[\textstyle\sum_{i=1}^{T}\sum_{j=1}^{T} a_i a_j \left(Y_i - \mu_{Y_i}\right)\left(Y_j - \mu_{Y_j}\right)\right] && \text{(expand)}\\
&= \textstyle\sum_{i=1}^{T}\sum_{j=1}^{T} \mathbf{E}\left[a_i a_j \left(Y_i - \mu_{Y_i}\right) \cdot \left(Y_j - \mu_{Y_j}\right)\right] && \text{(linearity of expectation)}\\
&= \textstyle\sum_{i=1}^{T}\sum_{j=1}^{T} a_i a_j \mathbf{E}\left[\left(Y_i - \mu_{Y_i}\right) \cdot \left(Y_j - \mu_{Y_j}\right)\right] && \text{(property of expectation)}
\end{aligned}$$

We now look at the terms of the last quantity.

- When $i = j$, then
$$\mathbf{E}\left[\left(Y_i - \mu_{Y_i}\right) \cdot \left(Y_j - \mu_{Y_j}\right)\right] = \mathbf{E}\left[\left(Y_i - \mu_{Y_i}\right)^2\right] = \mathbf{Var}[Y_i] ,$$
where the last equality was obtained by the definition of variance (Definition 5).
- When $i \neq j$, then
$$\mathbf{E}\left[\left(Y_i - \mu_{Y_i}\right) \cdot \left(Y_j - \mu_{Y_j}\right)\right] = \mathbf{Cov}\left[Y_i, Y_j\right] ,$$
where the last equality was obtained by the definition of covariance (Definition 6). However, we also know that Y_i and Y_j are *independent*. Hence, by Corollary 8, $\mathbf{Cov}\left[Y_i, Y_j\right] = 0$.

Thus, with the above observations we obtain

$$\sum_{i=1}^{T}\sum_{j=1}^{T} a_i a_j \mathbf{E}\left[\left(Y_i - \mu_{Y_i}\right) \cdot \left(Y_j - \mu_{Y_j}\right)\right] = \sum_{i=1}^{T} a_i^2 \cdot \mathbf{Var}[Y_i]$$

The theorem follows. ■

1.5 Weak Law of Large Numbers

Theorem 13 (Weak Law of Large Numbers). *Let* $Y_1, \ldots, Y_T$ *be a sequence of* independent identically distributed *random variables, with expected value* μ *and bounded variance. For every* $\epsilon > 0$*,*

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

Proof. Since $Y_1, \ldots, Y_T$ are *independent identically distributed* random variables, let μ be the common expectation of them and σ their common variance. Now, define the random variable

$$X = \sum_{i=1}^{T} \frac{1}{T} \cdot Y_i = \frac{1}{T}\sum_{i=1}^{T} Y_i \,.$$

By linearity of expectation we get $\mathbf{E}[X] = \sum_{i=1}^{T} \frac{1}{T}\mathbf{E}[Y_i] = \mu$.

Since all the Y_i are independent, we use Theorem 12 with $a_1 = a_2 = \ldots = a_T = \frac{1}{T}$, and thus, $\mathbf{Var}[X] = \sum_{i=1}^{T}\left(\frac{1}{T^2}\mathbf{Var}[Y_i]\right) = \sum_{i=1}^{T}\frac{\sigma^2}{T^2} = \frac{\sigma^2}{T}$.

We now apply Chebyshev's inequality (Theorem 5) and obtain for any $\epsilon > 0$,

$$\mathbf{Pr}\left(|X - \mu| \geqslant \epsilon\right) \leqslant \frac{\sigma^2}{T\epsilon^2}\,. \quad \blacksquare$$

2 Applications on Identically Distributed Events

Examples with (biased) coins are our best friends on understanding the bounds mentioned earlier. For more examples and more advanced techniques see [12].

Lemma 14 (Variance of a Biased Coin). *We are given a coin that gives* HEADS *with probability* p*. Show that the variance of a single coin toss is* $p(1-p)$*.*

Proof. Let X be the (indicator) random variable that is 1 if we observe HEADS after a single coin toss and 0 otherwise. By Proposition 4, the variance of X is

$$\mathbf{Var}[X] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 \,.$$

However, $\mathbf{E}\left[X^2\right] = p \cdot 1^2 + (1-p)\cdot 0^2 = p$. Also, $\mathbf{E}[X] = p \cdot 1 + (1-p)\cdot 0 = p$. Hence, by substitution we have

$$\mathbf{Var}[X] = p - (p)^2 = p - p^2 \,.$$

In other words, $\mathbf{Var}[X] = p(1-p)$ as needed. $\blacksquare$

Corollary 15 (Upper Bound on Variance of a Single Coin Toss for *Any Biased* Coin). *We are given a biased coin that succeeds (gives* HEADS*) with probability* p*. Show that the variance of a single coin toss is at most* $\frac{1}{4}$*.*

Proof. Let X be the (indicator) random variable that is 1 if the outcome of a single coin toss is HEADS, and 0 otherwise. By Lemma 14 we know that $\mathbf{Var}[X] = p(1-p)$.

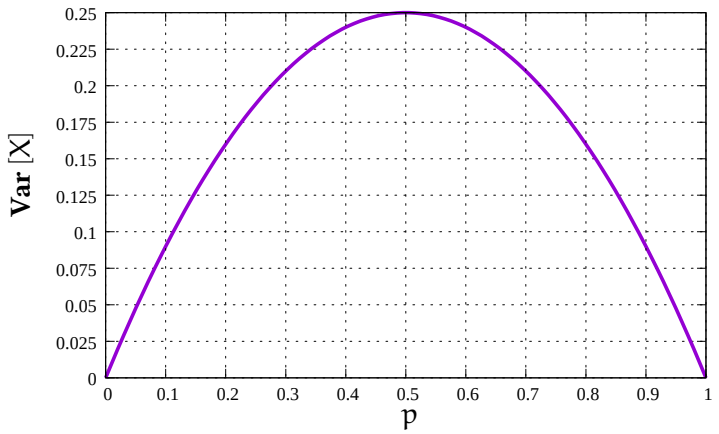

Figure 1: The function $f(p) = p(1-p)$, for $0 \leqslant p \leqslant 1$.

However, the function $f(p) = p(1-p)$ is a second degree polynomial for $0 \leqslant p \leqslant 1$. Moreover, $f(p)$ has roots $p_1 = 0$ and $p_2 = 1$ and obtains the maximum for $p = \frac{1}{2}$. Figure 1 shows the plot of $f(p)$ as p ranges in the $[0, 1]$ interval. Thus, for any $0 \leqslant p \leqslant 1$, it holds

$$f(p) \leqslant f(1/2) .$$

Hence, by substitution, $f(p) \leqslant f\left(\frac{1}{2}\right) = \frac{1}{2} \cdot \left(1 - \frac{1}{2}\right) = \frac{1}{4}$. In other words, $\mathbf{Var}[X] \leqslant \frac{1}{4}$. ■

Lemma 16 (Variance of Multiple Independent Identically Distributed Coin Tosses)**.** *We are given a coin that succeeds (gives* HEADS*) with probability* p*. We toss the coin* T *times. Show that the variance is* $T \cdot p \cdot (1-p)$*.*

Proof. Let Y_j be the indicator random variable for the j-th coin toss, where $j \in \{1, \ldots, T\}$. That is, $Y_j = 1$ if the outcome is HEADS for the j-th coin toss and $Y_j = 0$ otherwise (TAILS). By Lemma 14, $\mathbf{Var}\left[Y_j\right] = p(1-p)$. The random variable $X = \sum_{j=1}^{T} Y_j$ indicates how many successes we had after T coin tosses. Since each coin toss is *independent* of all the other ones, by Theorem 12 with $a_1 = a_2 = \ldots = a_T = 1$ we get

$$\mathbf{Var}[X] = \sum_{j=1}^{T} \mathbf{Var}\left[Y_j\right] . \tag{2}$$

Moreover, every time we toss *the same coin*, and thus the events are *identically distributed* with success probability p. Hence, by Lemma 14, the variance on *each coin toss* is $p(1-p)$. In other words, $\mathbf{Var}[Y_1] = \mathbf{Var}[Y_2] = \ldots = \mathbf{Var}[Y_T] = p(1-p)$. Thus, by (2), $\mathbf{Var}[X] = T \cdot p \cdot (1-p)$. ■

Corollary 17 (General Upper Bound on Variance of Multiple Independent Identically Distributed Coin Tosses)**.** *We are given a coin that succeeds (gives* HEADS*) with probability* p*. We toss the coin* T *times. Show that the variance is at most* $T/4$*.*

Proof. By Lemma 16, $\mathbf{Var}[X] = T \cdot p \cdot (1-p)$. Similar to Corollary 15, $f(p) = p(1-p) \leqslant f(1/2) = \frac{1}{4}$. Thus, regardless of what p is,

$$\mathbf{Var}[X] \leqslant \frac{T}{4} . \quad ■$$

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

Solution. Let $X = \sum_{i=1}^{100} X_i$ be the number of times that we observed HEADS, where the X_i's are indicator random variables indicating whether we obeserved HEADS or not in the i-th trial. Note that the expectation is $\mathbf{E}[X] = Tp = 100 \cdot (1/2) = 50$. Also note that by Lemma 16, $\mathbf{Var}[X] = Tp(1-p) = 100 \cdot (1/2) \cdot (1/2) = 25$. A direct computation for the probability $q = \mathbf{Pr}(X = 80)$ of such an event gives, $q = \binom{100}{80} \cdot 2^{-80} \cdot 2^{-20} = \binom{100}{80} \cdot 2^{-100} \approx 4.2 \cdot 10^{-10}$.

- Markov's inequality (Theorem 3) gives $\mathbf{Pr}(X \geqslant 80) \leqslant 50/80 = 0.625$.
- Chebyshev's inequality (Theorem 5) gives $\mathbf{Pr}(|X - 50| \geqslant 30) \leqslant \frac{25}{30^2} = 2.7 \cdot 10^{-2}$. ■

3 Sketch of Some Learning Theory Algorithms

Typically we derive an algorithm $\mathcal{A}$ to attack a particular problem and $\mathcal{A}$ works *with high probability*. Roughly, $\mathcal{A}$ computes a solution s such that

- either s achieves an *exact* optimum, or
- s achieves an *approximate (almost exact)* optimum.

Hence, depending on the guarantee that $\mathcal{A}$ has (that is, exactly correct or approximately correct), our big theorem is a statement of the form

$$\mathbf{Pr}(\mathcal{A} \text{ is } \textit{exactly} \text{ correct}) \geqslant 1 - \delta, \tag{3}$$

or

$$\mathbf{Pr}(\mathcal{A} \text{ is } \textit{approximately} \text{ correct}) \geqslant 1 - \delta. \tag{4}$$

Thus, δ is an upper bound for the probability that $\mathcal{A}$ will *fail* to deliver its guarantee at the end of the execution. We refer to δ as *confidence* even if in reality our true confidence is at least $1 - \delta$.

Remark 6 (δ is a parameter). *δ in (3) and (4) is a* parameter *for algorithm $\mathcal{A}$. This implies that one provides her favorite δ as part of the input for $\mathcal{A}$ and $\mathcal{A}$ will deliver a solution s that satisfies (3) or (4) depending on the case where $\mathcal{A}$ refers to.*

Remark 7 (PAC Learning). *We note that in order to capture the notion of* approximation *in (4), we introduce another variable ε. In this case we discuss about algorithms with probably approximately correct (PAC) guarantees.*

PAC learning, was introduced by Leslie Valiant[1] *in [16]. We will not occupy ourselves with PAC learning further here apart from the next remark. Our focus will be (3); that is* probably exactly correct *learning* [2]. *If you are interested in such kinds of questions around learning, sign up for the* CS 6501 - Learning Theory *course next semester.*

Remark 8 (On Running Time Efficiency). *An algorithm $\mathcal{A}$ that satisfies (3) or (4) is called* efficient *if its running time complexity has polynomial dependence on $\frac{1}{\delta}$, $\frac{1}{\varepsilon}$, and the rest of the input parameters (typically a notion of the dimension of the problem).*

[1] PAC learning was recognized as one of the major contributions that Leslie Valiant offered to theory of computation and Leslie Valiant received a Turing award in 2010. (The Turing award is the highest honor for anyone working in computer science, similar to the Nobel prize in other disciplines.)

[2] Please see Section 4 for some related discussion.

3.1 Typical Design of a PAC Learning Algorithm

The typical building blocks on the design of a learning algorithm are the following.

Step 1. Design $\mathcal{A}$ that works when all the (randomized) processes work in its favor.

Step 2. Identify which (randomized) processes may fail during the execution. We call such failures *bad events*.

Step 3. Count, or give an overestimate, of the processes for Step 2. Say that we have at most b such processes where things can go wrong.

Step 4. We now resort to the union bound (Proposition 2) and distribute the overall failure probability δ to all $\leqslant b$ processes where $\mathcal{A}$ may fail. The simplest way of doing that is by requiring each individual process to fail with probability at most $\frac{\delta}{b}$.

Hence, if we manage to bound the probability of *every bad event* by the quantity $\frac{\delta}{b}$, then we have an algorithm $\mathcal{A}$ that satisfies (3).

To see this, note that *some bad event* B_i will happen with probability

$$\mathbf{Pr}\left(B_1 \cup B_2 \cup \cdots \cup B_b\right) = \mathbf{Pr}\left(\bigcup_{i=1}^{b} B_i\right).$$

However, by the union bound,

$$\mathbf{Pr}\left(\bigcup_{i=1}^{b} B_i\right) \leqslant \sum_{i=1}^{b} \mathbf{Pr}\left(B_i\right).$$

Our hard work in Step 5 will guarantee $\mathbf{Pr}\left(B_i\right) \leqslant \frac{\delta}{b}$ for every $i \in \{1, \ldots, b\}$. Hence,

$$\mathbf{Pr}\left(\bigcup_{i=1}^{b} B_i\right) \leqslant \sum_{i=1}^{b} \mathbf{Pr}\left(B_i\right) \leqslant \sum_{i=1}^{b} \frac{\delta}{b} = \delta. \tag{5}$$

In other words, *some* bad event will happen with probability at most δ. Rephrasing, *none* bad event will happen with probability at least $1 - \delta$. This argument will give (3).

Step 5. Find a way to guarantee that for each bad event B_i, with $i \in \{1, \ldots, b\}$, it holds

$$\mathbf{Pr}\left(B_i\right) \leqslant \frac{\delta}{b}. \tag{6}$$

Thus, apart from counting in Step 3, our hard work is concentrated on trying to achieve (6) in Step 5. Arguing for Step 5 typically depends on the problem that we want to solve.

4 Historical Remarks and Further Reading

Section 3 presented a basic algorithmic idea that is behind several algorithms encountered in Valiant's PAC model of learning. However there are various models of learning within the broad spectrum of (computational) learning theory. One of them is the *membership query (MQ) model*[3] of Angluin [1].

[3] Section 1.2 in [1] contrasts briefly the membership query model and the PAC model of learning.

Roughly, in the MQ model the learner queries points from the domain and receives an answer as to whether or not these points belong to a particular concept or not (i.e., whether the points are *positive* or *negative*); the goal is to identify the target concept *precisely* (i.e., the function that determines the positive/negative label associated with each point of the domain). Note that there can be other models of learning that are somewhere between *exact identification* and *probably approximately correct* identification. Bshouty, Jackson and Tamon discuss such in-between models in [7]; they refer to our case of probably exactly correct learning as PExact learning.

A natural extension for various models of learning is the notion of *noise*. That is, we want to be able to discuss and argue about the guarantees for the algorithms that we design in the presence of uncertain information. One very natural variant of noise is along the lines of the homework; that is, the answer to *the label* of some queries can be wrong. A variant of such *misclassification noise* was considered by Angluin and Laird in [3], called *random misclassification noise*, in Valiant's PAC framework, where, according to the model, the label of each instance that is queried can have *incorrect label independently* with probability $p < 1/2$. It can be shown[4] that pretty much anything that is PAC learnable without noise is also PAC learnable with *random misclassification noise*; the notable exception is the *exclusive or (XOR)* function. There are other kinds of noise as well; e.g., the misclassification on the requested labels can be *malicious* and depend on the whole history of the labeled instances that have been previously returned to the learner. Sloan discusses this kind of noise in [14] and also describes other main variants of noise in the framework of PAC learning. In fact malicious noise can arise not only on the label of an instance but also on the instance[5] that is presented to the learner; this kind of malicious noise was considered by Valiant in [17] and was explored in detail by Kearns and Li in [11]. Beyond the models described by Sloan in [14], the *nasty noise* model of Bshouty, Eiron and Kushilevitz [6] should also be mentioned.

Similar to the random misclassification noise model of Angluin and Laird in the PAC learning framework, Sakakibara has considered this kind of noise in the MQ framework [13]. An interesting paper with some interesting problems along these lines is [8]. Note that even in the MQ model we can have various kinds of noise that one can study depending on the underlying model of a problem. For example, in [4] it is considered the case where the answer to the classification of an instance can be apart from positive and negative also *"I don't know"*; that is, *the teacher is not necessarily able to identify whether a particular instance belongs to a particular concept or not.* There are also other variants; for example a certain limited amount of *"I don't know"*'s are allowed in the responses of the teacher as in [15], or even some mistakes are malicious as in [2]. Another variant is that of the consistently ignorant teacher of [9], in the sense that the teacher can not answer a query with *"I don't know"* if the answer to the particular query can be inferred by answers to other queries that have been provided earlier in the learning process.

As a concluding remark we should perhaps mention that characterizing the sample size needed for PAC learnability is ultimately associated with the notion of the *Vapnik-Chervonenkis dimension (VC-dimension)* of the concept class being learnt [18]. In [5] a connection was established with PAC learning by using the techniques of the pioneering work of Vapnik and Chervonenkis on distribution-free convergence of empirical probability estimates.

[4] It is perhaps fair to say that we have understood better this kind of noise in Kearns' framework of *statistical queries* [10].

[5] In PAC learning the learner does not get to choose which instances are queried. Rather the instances are drawn following a distribution that is induced on the domain. It is in the MQ model that the learner has the power to select which instances to query for their labels.

References

[1] Dana Angluin. Queries and Concept Learning. *Machine Learning*, 2(4):319–342, 1987.

[2] Dana Angluin, Martins Krikis, Robert H. Sloan, and György Turán. Malicious Omissions and Errors in Answers to Membership Queries. *Machine Learning*, 28(2-3):211–255, 1997.

[3] Dana Angluin and Philip D. Laird. Learning From Noisy Examples. *Machine Learning*, 2(4):343–370, 1987.

[4] Dana Angluin and Donna K. Slonim. Randomly Fallible Teachers: Learning Monotone DNF with an Incomplete Membership Oracle. *Machine Learning*, 14(1):7–26, 1994.

[5] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, October 1989.

[6] Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

[7] Nader H. Bshouty, Jeffrey C. Jackson, and Christino Tamon. Exploring learnability between exact and PAC. *Journal of Computer and System Sciences*, 70(4):471–484, 2005.

[8] Thomas L. Dean, Dana Angluin, Kenneth Basye, Sean P. Engelson, Leslie Pack Kaelbling, Evangelos Kokkevis, and Oded Maron. Inferring Finite Automata with Stochastic Output Functions and an Application to Map Learning. *Machine Learning*, 18(1):81–108, 1995.

[9] Michael Frazier, Sally A. Goldman, Nina Mishra, and Leonard Pitt. Learning from a Consistently Ignorant Teacher. *Journal of Computer and System Sciences*, 52(3):471–492, 1996.

[10] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[11] Michael J. Kearns and Ming Li. Learning in the Presence of Malicious Errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[12] Michael Mitzenmacher and Eli Upfal. *Probability and computing - randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[13] Yasubumi Sakakibara. On Learning from Queries and Counterexamples in the Presence of Noise. *Information Processing Letters*, 37(5):279–284, 1991.

[14] Robert H. Sloan. Four Types of Noise in Data for PAC Learning. *Information Processing Letters*, 54(3):157–162, 1995.

[15] Robert H. Sloan and György Turán. Learning with Queries but Incomplete Information (Extended Abstract). In *COLT*, pages 237–245, 1994.

[16] Leslie G. Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[17] Leslie G. Valiant. Learning disjunctions of conjunctions. In *IJCAI*, pages 560–566, 1985.

[18] Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. Original publication appeared in 1968 in Russian in Dokl. Akad. Nauk SSSR, 181 (4): 781. 1968.