Physics of Learning Theory
Lecture 1: Probability

February $--$, 2025

Nilava Metya

1 Introduction

We will recall some probability theory and look at useful *deviation* or *concentration* bounds which are frequently used in analyzing algorithms (in learning theory). Recall that a (real-valued) random variable on a probability space $(\Omega, S, \mathbb{P})$ is nothing but a 'measurable function' $X : \Omega \to \mathbb{R}$. Here Ω is the universal or sample space where we think of events in, S is a collection of events in Ω and $\mathbb{P} : S \to [0, 1]$ assigns probability to each event in S. The space of events S is constrained to satisfy some obvious rules like Ω is an event, if A is an event then so is $\Omega \setminus A$ and that a countable union of events is an event which makes it sensible to work with the concept of assigning probabilities to each event. We will often say that $\mathbb{P}[A]$ is the probability that event A occurs. If $A = \{a\}$ is a singleton, we always write $\mathbb{P}[a]$ instead of $\mathbb{P}[\{a\}]$. The probability function $\mathbb{P}$ is also constrained to a couple of rules, namely, that the probability of the union of a mutually disjoint collection of events, which is an event, is the same as the sum of the probabilities of each of those events and that the probability that Ω occurs is 1. Roughly a random variable is to be thought of as a way of assigning points of the sample space to real numbers which are *really real* and are more tangible to work with, while respecting the rules of S. Such a random variable induces a map $X^{-1} : 2^{\mathbb{R}} \to S$ by $X^{-1}(A) \coloneqq \{x \in \Omega \mid X(x) \in S\}$ for any $A \subseteq \mathbb{R}$, and hence induces a probability on $\mathbb{R}$ given by $\mathbb{P}_{\mathbb{R}}[A] = \mathbb{P}[X^{-1}(A)]$ where A is any 'measurable' subset of $\mathbb{R}$. The random variable being a 'measurable function' precisely means that $X^{-1}(A)$ always lies in S.

$$
\begin{tikzcd}
S \arrow[d, "\mathbb{P}"'] & 2^{\mathbb{R}} \arrow[l, "X^{-1}"'] \arrow[dl, dashed, "\mathbb{P}_{\mathbb{R}}"] \\
{[0,1]} &
\end{tikzcd}
$$

1.1 Mean

The average or mean of a random variable X, often denoted as $\mathbb{E}[X]$, $\mu(X)$, or simply μ when the context is clear, is $\mathbb{E}[X] = \int_\Omega X \,\mathrm{d}\,\mathbb{P}$. For the discrete case, which we will mostly be interested in, this boils down to $\mathbb{E}[X] = \sum_{i\in\Omega} X(i)\mathbb{P}[i]$. Note that if X is an indicator random variable for event A, that is, $X = 1$ if A occurs and 0 otherwise, then $\mathbb{E}[X] = \mathbb{P}[A]$.

Example 1. Consider tossing a fair coin. Here $\Omega = \{\mathrm{H}, \mathrm{T}\}$. The probability function is $\mathbb{P}[\varnothing] = 0, \mathbb{P}[\mathrm{H}] = \mathbb{P}[\mathrm{T}] = 0.5, \mathbb{P}[\{\mathrm{H}, \mathrm{T}\}] = 1$. A natural random variable to consider is $X(i) = \mathbf{1}_{\mathrm{H}} \coloneqq \begin{cases} 1 & \text{if } i = \mathrm{H} \\ 0 & \text{if } i = \mathrm{T} \end{cases}$. The corresponding probability induced on $\mathbb{R}$ is given by $\mathbb{P}_{\mathbb{R}}[A] = \begin{cases} 0 & \text{if } 0 \notin A, 1 \notin A \\ 0.5 & \text{if } 0 \in A, 1 \notin A \\ 0.5 & \text{if } 0 \notin A, 1 \in A \\ 1 & \text{if } 0 \in A, 1 \in A \end{cases}$. In this case, $\mathbb{E}[X] = 1 \cdot \mathbb{P}[\mathrm{H}] + 0 \cdot \mathbb{P}[\mathrm{T}] = 0.5$

Example 2. Consider tossing n fair coins sequentially and independently. Here $\Omega = \{\mathrm{H}, \mathrm{T}\}^n$. So the singleton outcomes are tuples of H, T. The probability function is given by $\mathbb{P}[\boldsymbol{x}] = 2^{-n}$ for any element $x \in \Omega$ and then extending by countable additivity of $\mathbb{P}$. Consider n random variables $X_1, \cdots, X_n$ where $X_i(\boldsymbol{x}) \coloneqq \begin{cases} 1 & \text{if } x_i = \mathrm{H} \\ 0 & \text{if } x_i = \mathrm{T} \end{cases}$. Each X_i is the same random variable as the previous example after looking at the i^{th} coordinate. A natural variable to consider is the total number of heads obtained in one round of tossing, that is $X = X_1 + \cdots + X_n$. The corresponding probability induced on $\mathbb{R}$ is given by $\mathbb{P}_{\mathbb{R}}[k] = \begin{cases} \binom{n}{k} 2^{-n} & \text{if } k \in \{0, \cdots, n\} \\ 0 & \text{otherwise} \end{cases}$ and extend by countable additivity. Here $\mathbb{E}[X] = \frac{n}{2}$.

One useful result used for calculating expectations of sums of random variables is that if $a, b \in \mathbb{R}$ and X, Y are random variables then $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. It's worthy to note that sums and scalings of random variables are random variables. This result does **not** depend on 'independence' of X, Y. Independence plays an important role for the average of products of random variables (which is a random variable). We say random variables $X_1, \cdots, X_n$ are (*mutually*) *independent* if $\mathbb{P}\left[\bigcap_{i=1}^n \{X_i \le a_i\}\right] = \prod_{i=1}^n \mathbb{P}[X_i \le a_i] \ \forall\ a_i \in \mathbb{R}$. This is a stronger notion than *pairwise* independence where we demand that only every pair of them are independent. Note that mutual independence implies pairwise independence. If X, Y are independent then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

For a random variable $X \ge 0$ and $a \in \mathbb{R}$ let Y be the indicator random variable indicating whether $X \ge a$, that is, Y is 1 if $X \ge a$ and 0 otherwise. Then clearly $X \ge aY$. Indeed if $X \ge a$ then $Y = 1$ so $X \ge aY$ and if $X < a$ then $Y = 0$ so that $X \ge 0 = aY$. Expectation preserves inequalities, so $\mathbb{E}[X] \ge a\mathbb{E}[Y] = a\mathbb{P}[X \ge a]$. This establishes

Theorem 1 (Markov's inequality)
If X is a non-negative random variable and $a \in \mathbb{R}$ then $\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$.

1.2 Variance

Let's come to deviation now. One natural way to measure *deviation* is to look how on average much a random variable deviates either way from its mean (behavior). To look for deviation in either direction of $\mathbb{E}[X]$ we consider the random variable $(X - \mathbb{E}[X])^2$. Define the variance of a random variable X as $\text{Var}[X] := \mathbb{E}[(X - \mathbb{E}[X])^2]$. One useful result to compute variance is that if X, Y are independent then $\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y]$. This extends to n pairwise independent random variables. Another useful result is $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Applying Theorem 1 to $(X - \mathbb{E}[X])^2 \geq 0$ gives

Theorem 2 (Chebyshev's inequality)
If X is a random variable and $a \in \mathbb{R}_{\geq 0}$ then $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \frac{\text{Var}[X]}{a^2}$.

1.3 Higher moments

One might just ask why stop at the second power to measure deviation. What about the random variable $(X - \mathbb{E}[X])^k$ for $k \geq 2$? These are called higher centeral moments. Note that $\mathbb{E}\left[(X - \mathbb{E}[X])^k\right] = 0$ when k is odd and the distribution of X is symmetric about $\mathbb{E}[X]$. So it makes sense to consider the random variables $\mu_k := |X - \mathbb{E}[X]|^k$ instead. If we have access to such numbers, we can use the same trick as the proof of Chebyshev's inequality and get $\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \frac{\mu_k}{a^k}$. Knowing all higher moments means that we know something known as the 'characteristic function' (not yet defined) of X which uniquely determines X. But our aim was the study deviations using small information. Generally, higher moments are not known.