Probability - short review

1 The Basics

Definition 1.1. *A probability space has three components:*

1. *A set* Ω *of possible outcomes.*
2. *A collection of events* $\mathcal{F}$*, where each event* $E \in \mathcal{F}$ *is a subset of* Ω*. An event containing a single element of* Ω *is called* basic.
3. *A probability function* $\Pr : \mathcal{F} \to [0,1]$ *satisfying:*
   (a) $\Pr[\Omega] = 1$.
   (b) *For any countable sequence of pairwise mutually disjoint events* $\{E_i\}$*:*

$$\Pr[\cup_i E_i] = \sum_i \Pr[E_i].$$

Lemma 1.2. *Let* $E_1, E_2 \in \mathcal{F}$ *be events (not necessarily disjoint), then*

$$\Pr[E_1 \cup E_2] = \Pr[E_1] + \Pr[E_2] - \Pr[E_1 \cap E_2].$$

Proof:

$$\begin{aligned}
\Pr[E_1] &= \Pr[E_1 \setminus E_1 \cap E_2] + \Pr[E_1 \cap E_2], \text{ and} \\
\Pr[E_2] &= \Pr[E_2 \setminus E_1 \cap E_2] + \Pr[E_1 \cap E_2], \text{ and} \\
\Pr[E_1 \cup E_2] &= \Pr[E_1 \setminus E_1 \cap E_2] + \Pr[E_1 \cap E_2] + \Pr[E_2 \setminus E_1 \cap E_2]. \\
\text{Hence, } \Pr[E_1 \cup E_2] &= \Pr[E_1] + \Pr[E_2] - \Pr[E_1 \cap E_2].
\end{aligned}$$

□

Corollary 1.3. *For any countable sequence of events* $\{E_i\}$*:*

$$\Pr[\cup_i E_i] \le \sum_i \Pr[E_i].$$

Definition 1.4. *Events* E, F *are called* independent *if* $\Pr[E \cap F] = \Pr[E] \cdot \Pr[F]$ *holds.*

Definition 1.5. *The* conditional probability *of* E *given* F *is defined as*

$$\Pr[E|F] = \frac{\Pr[E \cap F]}{\Pr[F]}.$$

Corollary 1.6. *For independent events E, F:*

$$\Pr[E|F] = \Pr[E].$$

Corollary 1.7 (Bayes' Rule). *For any two events E, F:*

$$\Pr[E|F] = \Pr[F|E] \cdot \frac{\Pr[E]}{\Pr[F]}$$

Definition 1.8 (Random Variables)**.** *A random variable is a function $X : \Omega \to \mathbb{R}$. For a (discrete) random variable X and a real number a, the event $X = a$ corresponds to the set of basic events on which the variable X is assigned the value a:*

$$\Pr[X = a] = \sum_{\omega \in \Omega : X(\omega) = a} \Pr[\omega].$$

2 Expectancy, Variance, and higher moments

Definition 2.1. *The* expectancy *(or* expectation *or* mean*) of a (discrete) random variable X is*

$$\mathbf{E}[X] = \sum_a a \cdot \Pr[X = a].$$

Theorem 2.2 (Linearity of Expectation)**.** *Let X, Y be random variables, then*

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

Proof:

$$\begin{aligned}
\mathbf{E}[X+Y] &= \sum_a \sum_b (a+b) \Pr[X = a \cap Y = b] \\
&= \sum_a a \sum_b \Pr[X = a \cap Y = b] + \sum_b b \sum_a \Pr[X = a \cap Y = b] \\
&= \sum_a a \Pr[X = a] + \sum_b b \Pr[Y = b] \\
&= \mathbf{E}[X] + \mathbf{E}[Y].
\end{aligned}$$

□

Definition 2.3. *The* conditional expectancy *of X given event E is*

$$\mathbf{E}[X|E] = \sum_a a \Pr[X = a|E].$$

Definition 2.4. *The t-th* moment *of a random variable X is $\mathbf{E}[X^t]$.*

Definition 2.5. *The* variance *of a random variable X is*

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

The standard deviation is

$$\sigma_X = \sqrt{\mathbf{Var}[X]}.$$

Corollary 2.6. *For independent random variables X, Y :*

$$\mathbf{E}[X \cdot Y] = \mathbf{E}[X] \cdot \mathbf{E}[Y] \quad \textit{and} \quad \mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y].$$

Definition 2.7 (Moment Generating Function)**.** *The* moment generating function *of a random variable X is*

$$m_X(t) = \mathbf{E}[e^{tX}],$$

for $t \in R$ for which the expectation exists.

Noting that $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots = \sum_{j=0}^{\infty} \frac{x^j}{j!}$, we get that

$$m_X(t) = \sum_i \Pr[X=i] \cdot \sum_{j=0}^{\infty} \frac{(xt)^j}{j!} = \sum_{j=0}^{\infty} \sum_i \Pr[X=i] \cdot \frac{(xt)^j}{j!} = \sum_{j=0}^{\infty} \mathbf{E}[x^j] \cdot \frac{t^j}{j!}.$$

Derivating at $t = 0$, we get that

$$\left.\frac{d^j m_X(t)}{dt^j}\right|_{t=0} = \mathbf{E}[x^j],$$

i.e., the jth moment of X.

We will use the moment generating function in deriving bounds on the sum of independent random variables. Let $S = \sum_i X_i$ where the random variables X_i are independent. We get that

$$m_S(t) = \mathbf{E}[e^{tS}] = \mathbf{E}[e^{t \cdot \sum_i X_i}] = \mathbf{E}[\prod_i e^{tX_i}] = \prod_i \mathbf{E}[e^{tX_i}] = \prod_i m_{x_i}(t).$$

3 Continuous Variables

For a continuous random variable X define the probability density function (PDF) $\mathsf{PDF}_X(x) : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that

$$\int_{-\infty}^{\infty} \mathsf{PDF}_X(x) dx = 1.$$

Informally, we will write

$$\Pr[X \in S] = \int_S \mathsf{PDF}_X(x) dx.$$

The cumulative distribution function (CDF) of X is the function defined as

$$\mathsf{CDF}(x) = \Pr[X < x] = \int_{-\infty}^{x} \mathsf{PDF}(x) dx$$

4 Some Random Variables

Below are a list of random variables that we will repeatedly use in this course.

Bernoulli Random Variable. A discrete random variable. X is called a *Bernoulli random variable*, denoted $X \sim Ber(p)$ if X takes only two values $0, 1$ with $p = \Pr[X = 1]$. The expectation of a Bernoulli random variable is $\mathbf{E}[X] = p$ and its variance is $p - p^2$.
Bernoulli random variables are often called indicators. For any event E we can associate a corresponding Bernoulli r.v. where $X = 1$ if E holds, and $X = 0$ otherwise.

Uniform $[0, 1]$**.** When X is a r.v. chosen uniformly at random (u.a.r) from the interval $[0, 1]$, denoted $X \sim U_{[0,1]}$ then for any $x \in [0, 1]$ we have that $\mathsf{PDF}(x) = 1$ and 0 everywhere else (so the PDF indeed integrates to 1) and $\mathsf{CDF}(x) = x$ on the $[0, 1]$ interval. The expected value of X is $\mathbf{E}[X] = 1/2$ and $\mathbf{Var}[X] = 1/3 - (1/2)^2 = 1/12$.

Exponential Random Variable. A continuous random variable X is called *exponential*, denoted $X \sim Exp(\lambda)$ if its PDF is defined as: $\mathsf{PDF}(x) = \lambda e^{-\lambda x}$ for any $x \geq 0$. In this case $\mathsf{CDF}_X(x) = 1 - e^{-\lambda x}$ for any $x \geq 0$, the expectation $\mathbf{E}[X] = 1/\lambda$ and the variance $\mathbf{Var}(X) = 1/\lambda^2$.

Laplace Random Variable. A continuous random variable X is called *Laplace*, denoted $X \sim Lap(\lambda)$ if its PDF is defined as: $\mathsf{PDF}(x) = \frac{1}{2\lambda} e^{-|x|/\lambda}$ for any $x \in \mathbb{R}$. Observe that one way to sample a r.v. $X \sim Lap(\lambda)$ is to pick $Y \sim Exp(1/\lambda)$ and then set $X = Y$ w.p. $1/2$ and $X = -Y$ w.p. $1/2$. The mean of a Laplace random variable is $\mathbf{E}[X] = 0$ and its variance is $\mathbf{Var}(X) = 2\lambda^2$.

Gaussian Random Variable. A continuous random variable X is called *Gaussian*, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$ if its PDF is defined as: $\mathsf{PDF}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu)^2}$ for any $x \in \mathbb{R}$. The mean of a Gaussian random variable is $\mathbf{E}[X] = \mu$ and its variance is $\mathbf{Var}(X) = \sigma^2$. A r.v. $X \sim \mathcal{N}(0, 1)$ is called a *normal* random variable.

5 Tail inequalities

Theorem 5.1 (Markov's Inequality)**.** *For a non-negative random variable X,*

$$\Pr[X \geq a] \leq \frac{\mathbf{E}[X]}{a}.$$

Proof: Define a new random variable to be the indicator function:

$$I = \begin{cases} 1 & X \geq a \\ 0 & \text{otherwise} \end{cases}$$

Note that $I \leq X/a$, hence $\mathbf{E}[I] \leq \mathbf{E}[X]/a$. We get:

$$\Pr[X \geq a] = \Pr[I = 1] = \mathbf{E}[I] \leq \mathbf{E}[X]/a.$$

□

Theorem 5.2 (Chebyshev Inequaltiy)**.** *For any random variable X,*

$$\Pr[|X - \mathbf{E}[X]| > a] \leq \frac{\mathbf{Var}[X]}{a^2}$$

Proof: Let Y denote the (non-negative) r.v. $Y = (X - \mathbf{E}[X])^2$. Then

$$\Pr[|X - \mathbf{E}[X]| > a] = \Pr[(X - \mathbf{E}[X])^2 > a^2] = \Pr[Y > a^2] \leq \frac{\mathbf{E}[Y]}{a^2} = \frac{\mathbf{Var}[X]}{a^2}$$

□

Theorem 5.3 (Chernoff-Hoeffding Inequalities)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $X_i \in [0,1]$ and $\mathbf{E}[X_i] = \mu$ and denote $S = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for all $\epsilon > 0$:*

$$\text{(Hoeffding:)}\ \Pr[S > \mu + \epsilon] \le e^{-2n\epsilon^2} \quad \text{and} \quad \Pr[S < \mu - \epsilon] \le e^{-2n\epsilon^2}$$
$$\text{(Chernoff:)}\ \Pr[S > (1+\epsilon)\mu] \le e^{-n\mu\epsilon^2/3} \quad \text{and} \quad \Pr[S < (1-\epsilon)\mu] \le e^{-n\mu\epsilon^2/2}$$

5.1 Applying Tail Inequalities

Imagine we toss a fair coin n (many) times. We know that w.p. $1/2$ we see Heads, which means that roughly $1/2$ of our tosses are likely to come out Heads and half should come out as Tails. So, what is the probability that we see a lot of heads? Say, more than $\frac{1+\epsilon}{2}$ fraction of the tosses are Heads?

Let X_i be the Bernoulli random variable which is 1 if the i-th coin toss comes out Heads. Let $X = \sum_{i=1}^n X_i$. Then for every i we have that $\mathbf{E}[X_i] = \frac{1}{2}$ and because of Linearity of Expectation we have that $\mathbf{E}[X] = \frac{n}{2}$. So now we can use Markov's Inequality and deduce that

$$\Pr[X > \tfrac{1+\epsilon}{2}n] \le \frac{n/2}{n(1+\epsilon)/2} = \frac{1}{1+\epsilon}$$

This bound is really not tight. First of all, it is pretty close to 1. More importantly, it's not improving with n.

Let us now try to bound this event using Chebyshev. Well, $\mathbf{Var}[X_i] = 1/4$ for any i and since all coin tosses are independent then $\mathbf{Var}[X] = n/4$. So we now how that

$$\Pr[X > \tfrac{1+\epsilon}{2}n] \le \Pr[|X - \tfrac{n}{2}| > \epsilon n/2] \le \frac{n/4}{\epsilon^2 n^2/4} = \frac{1}{\epsilon^2 n}$$

This is already much better. It means that when $n = 2/\epsilon^2$ then this event happens w.p. $< 1/2$. Yet, what if we want this probability to be really small? Not just $1/2$ but rather $1/20,000$? This means we have to set $n = 20,000/\epsilon^2$. In general, if we want this probability to be at most δ, then we need to toss the coin $1/\delta\epsilon^2$ times.

To improve on this, we use the Chernoff-Hoeffding bounds. We can use the Chernoff bound and deduce this probability is at most $e^{-n\epsilon^2/6}$. So, if we want this probability to be at most δ then we need to set $n = 6\ln(1/\delta)/\epsilon^2$. Observe that n now depends on $\log(1/\delta)$ rather than $1/\delta$. The Hoeffding bound gives a similar result $n = O(\log(1/\delta)/\epsilon^2$.[1]

Why is this logarithmic dependence in δ important? Imagine the following scenario. Preparing to the upcoming elections we are conducting a phone survey, and we ask randomly and independently chosen people whether they are pro or con k different current issues. How many people do we need to survey to know the true answer for *all* queries up to, say, 5%-error?

We formalize this problem as follows. Let n denote the size of our survey and for any $j \in \{1, 2, \ldots, k\}$ we define X_i^j, which is a Bernoulli r.v. indicating whether the i-th person is supporting the j-th issue. Let $X^j = \frac{1}{n}\sum_i X_i^j$. What is $E[X^j] = E[X_i^j]$? That is the fraction of the people in the population that are favor of the j-th issue. Observe that since we pick the survey participants randomly, then for any j it holds that $\{X_1^j, X_2^j, \ldots, X_n^j\}$ are all mutually independent. (But do note that X_i^1 and X_i^2 are not independent since it is the same person answering both questions.)

[1] In general, this quadratic dependence on $1/\epsilon$ is unavoidable. However, if we know that $p = O(\epsilon)$ then the Chernoff bound outperforms the Heoffding bound: whereas the Hoeffding bound has dependence of ϵ^{-2}, the Chernoff have n depending only on $1/\epsilon$.

Our goal is to lower-bound the probability $\Pr[\forall j,\ |X^j - E[X^j]| \leq \epsilon]$, which is equivalent to upper bounding the probability $\Pr[\exists j,\ |X^j - \mathbf{E}[X^j]| > \epsilon]$. That is, we want to have it so that *no* question has a bad estimation.

Note that we can't directly use Chernoff-Hoeffding, because not all events are independent. Instead, we can use the following argument. Fix j. Now the events $X_1^j, \ldots, X_n^j$ are independent and we can use Hoeffding's inequality to deduce that for one issue, $\Pr[|X^j - \mathbf{E}[X^j]| > \epsilon] < 2e^{-2n\epsilon^2}$. The next step is to use the Union Bound — since if there exists a j s.t. $|X^j - \mathbf{E}[X^j]| > \epsilon$ then this j is either 1, or 2, or 3,..., or k. So

$$\Pr[\exists j,\ |X^j - \mathbf{E}[X^j]| > \epsilon] \leq \sum_{j=1}^{k} \Pr[|X^j - \mathbf{E}[X^j]| > \epsilon] \leq 2ke^{-2n\epsilon^2}$$

Therefore, if we want that w.p. $1 - \delta$ all estimations to all k queries are within an error of ϵ is suffices to set $n = \ln(2k/\delta)/(2\epsilon^2)$. In other words, if we want to be 99% confident we know the answer to all k questions up to ϵ accuracy, then it suffices to have a sample of size $n = O(\ln(k)/\epsilon^2)$.

This argument will recur quite frequently throughout the semester. We will often abbreviate it by saying "using Chernoff and union we get..."