E0 206: Theorist's Toolkit October 1, 2020

Lecture 1: Probability Refresher

Instructor: Arindam Khan *Scribe: Deepesh Hada*

This lecture will serve as a refresher on basic probability needed for the rest of the course lectures.

1 Fundamentals

Definition 1 (Probability Space)**.** A probability space consists of three components:

1. a sample space Ω, which is the set of all possible outcomes of the random process modelled by the probability space;
2. a family of sets $\mathcal{F}$ representing the allowable events, where each set in $\mathcal{F}$ is a subset of the sample space Ω; and
3. a probability function $\Pr : \mathcal{F} \to \mathbf{R}$, satisfying *Definition 3*.

An element of Ω is called a *simple* or *elementary* event.

Definition 2 (σ-field)**.** A collection of subsets, $\mathcal{F}$ is referred to as a σ-field (or sigma-field) if:

1. $\Omega \in \mathcal{F}$;
2. for an event $A \in \mathcal{F} \implies A^c \in \mathcal{F}$; and
3. for a sequence of events $A_1, A_2, ..., A_n \in \mathcal{F} \implies \bigcup_{i=1}^{n} \in \mathcal{F}$.

σ-field is a collection of subsets of the sample space, Ω, that is used to formally measure probability. The sets in the σ-field constitute the events from Ω. A usual choice of $\mathcal{F}$ is 2^{Ω}, i.e., all the possible events in Ω.

Definition 3 (Probability Function)**.** A probability function is any function $\Pr : \mathcal{F} \to \mathbf{R}$ that satisfies the following conditions:

1. for any event $E, 0 \leq Pr(E) \leq 1$;
2. $\Pr(\Omega) = 1$; and
3. for any finite or countably infinite sequence of pairwise mutually disjoint events $E_1, E_2, E_3, ...,$

$$\Pr\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} \Pr(E_i).$$

Lemma 4 (Inclusion-Exclusion Principle)**.** *Let $E_1, E_2, ..., E_n$ be any n events. Then,*

$$\Pr\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \Pr(E_i) - \sum_{i<j} \Pr(E_i \cap E_j) + \sum_{i<j<k} \Pr(E_i \cap E_j \cap E_k) - ... + (-1)^{l+1} \sum_{i_1<i_2<...<i_l} \Pr\left(\bigcap_{r=1}^{l} E_{i_r}\right).$$

Corollary 5 (Union Bound)**.** *For any finite or countably infinite sequence of events $E_1, E_2, ...,$ which may not necessarily be pairwise disjoint,*

$$\Pr\left(\bigcup_{i\geq 1} E_i\right) \leq \sum_{i\geq 1} \Pr(E_i).$$

Definition 6 (Independence)**.** Two events E and F are independent if and only if,,

$$\Pr(E \cap F) = \Pr(E).\Pr(F).$$

More generally, events $E_1, E_2, ..., E_k$ are mutually independent if and only if, for any subset $I \subseteq [1, k]$,

$$\Pr(\bigcap_{i \in I} E_i) = \prod_{i \in I} \Pr(E_i).$$

Mutual independence is a stronger definition than **pairwise independence** since it *implies* pairwise independence. It is possible and easy to construct an example that shows that a collection of events can be pairwise independent but not mutually independent.

Definition 7 (Conditional Probability)**.** The conditional probability that event E occurs given that event F occurs is,

$$Pr(E|F) = \frac{\Pr(E \cap F)}{\Pr(F)}.$$

$\Pr(E|F)$ is well-defined if and only if $\Pr(F) > 0$. In the conditional universe where the event F has already occurred, $\Pr(E|F)$ gives us the updated probability of occurrence of event E.

By the **Chain Rule**, the joint occurrence of a sequence of k events, $E_1, E_2, ..., E_k$ is given as,

$$\Pr\left(\bigcap_{i=1}^{k} E_i\right) = \prod_{i=1}^{k} \Pr\left(E_i \,\middle|\, \bigcap_{j=1}^{i-1} E_j\right).$$

Theorem 8 (Law of Total Probability)**.** *Let $E_1, E_2, ..., E_n$ be mutually disjoint events in the sample space Ω, and let $\bigcup_{i=1}^{n} E_i = \Omega$. Then*

$$\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap E_i) = \sum_{i=1}^{n} \Pr(B|E_i)\Pr(E_i).$$

Theorem 9 (Bayes' Theorem)**.** *Let $E_1, E_2, ..., E_n$ be mutually disjoint events in the sample space Ω such that $\bigcup_{i=1}^{n} E_i = \Omega$. Then*

$$\Pr(E_j|B) = \frac{\Pr(E_j \cap B)}{\Pr(B)} = \frac{\Pr(B|E_j)\Pr(E_j)}{\sum_{i=1}^{n} \Pr(B|E_i)\Pr(E_i)},$$

where the last equality comes from the Chain Rule and the Law of Total Probability.

2 Discrete Random Variables

Definition 10 (Random Variable)**.** A random variable (R.V.) X is a function $X : \Omega \to \mathbf{R}$. A discrete R.V. takes finite or countably infinite number of values.

A random variable is neither random nor a variable, hence, a misnomer. It is a deterministic function that maps every element in the sample space to a real line.

Definition 11 (Expectation)**.** The Expectation of a random variable X is given as

$$\mathbb{E}[X] = \sum_{i} i\Pr[X = i],$$

where the summation is over all values in the range of $\mathbf{X}$.

Another handy rule comes in the form of the **Law of the Unconscious Statistician** (LOTUS). LOTUS is used to calculate the expected value of a function $g(X)$ of a random variable X when one knows the probability distribution of X, but not the distribution of $g(X)$. The expected value of $g(X)$ is

$$\mathbb{E}[g(X)] = \sum_{i} g(i)\Pr[X = i].$$

2.1 Properties of Expectation

Theorem 12 (Linearity of Expectation)**.** *For any finite collection of discrete random variables $X_1, X_2, ..., X_n$ with finite expectations,*

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

These random variables need not be independent.

Proof. We prove the statement for two random variables X and Y. Note that $X + Y$ is also a random variable with the domain being Ω. So,

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{\omega\in\Omega} (X+Y)(\omega)\Pr(\omega) \\
&= \sum_{\omega\in\Omega} (X(\omega)+Y(\omega))\Pr(\omega) \\
&= \sum_{\omega\in\Omega} (X(\omega)\Pr(\omega)+Y(\omega)\Pr(\omega)) \\
&= \mathbb{E}[X]+\mathbb{E}[Y].
\end{aligned}$$

The general case with more than 2 random variables follows by induction. □

Lemma 13. *For any constant c and a discrete random variable X,*

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

2.2 Bernoulli Distribution

Suppose that we run an experiment that succeeds with probability p and fails with probability $1-p$. Let X be a random variable such that

$$X = \begin{cases} 1 & \text{if the experiment succeeds,} \\ 0 & \text{otherwise.} \end{cases}$$

A Bernoulli random variable has a parameter p, which is the probability of *success* (defining *success* is subjective and depends on the experiment). Since it is an indicator of success, the Bernoulli random variable is also known as an *Indicator random variable.*

2.2.1 Expectation of a Bernoulli R.V.

For a Bernoulli random variable,

$$\mathbb{E}[X] = p.1 + (1-p).0 = p = \Pr(X = 1).$$

2.3 Binomial Distribution

Definition 14. A binomial random variable X with parameters n and p is defined by the following probability distribution on $k = 0, 1, 2, ..., n$:

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

. Here, X counts the number of successes. $\Pr(X = k)$ gives the probability of exactly k successes and $n-k$ failures in n independent trials, each of which is successful with probability p.

2.3.1 Expectation of a Binomial R.V.

The expectation of a Binomial R.V. can be found in multiple ways. The most elegant way it to use independent identically distributed (iid) indicator random variables. If X is a binomial random variable with parameters n and p, then X is the number of successes in n trials, where each trial is successful with probability p. We define a set of n indicator random variables $X_1, X_2, ..., X_n$, where $X_i = 1$ if the i^{th} trial is successful and 0 otherwise.

Clearly, $\mathbb{E}[X_i] = p$ and $X = \sum_{i=1}^{n} X_i$ and so, by the linearity of expectations,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = np.$$

2.4 Geometric Distribution

Definition 15. A geometric random variable X with parameter p is given by the following probability distribution on $n = 1, 2, ...,$:

$$\Pr(X = n) = (1-p)^{n-1}p.$$

The geometric random variable X counts the number of repeated independent trials until the first success (inclusive). Hence, for X to equal n, there must be $n-1$ failures, followed by a success.

Geometric random variables are said to be memoryless because the probability that the first success will be reached n trials from now is independent of the number of failures experienced till now. Informally, one can ignore past failures because they do not change the distribution of the number of future trials until first success. Formally, we have the following statement.

Lemma 16 (Memoryless property)**.** *For a geometric random variable X with parameter p and for $n > 0$,*

$$\Pr(X = n + k | X > k) = \Pr(X = n).$$

2.4.1 Expectation of Geometric R.V.

Consider a geometric random variable X. If we get success in the very first trial, the number of failures is 0, and X will take a value of 1. However, if this trial results in a failure, then the number of trials will increase by 1, and we'll start all over again because of the memoryless property of X. Based on this, we get the recurrence relation:

$$\mathbb{E}[X] = 1.p + (1 + \mathbb{E}[X])(1-p),$$

solving which we get

$$\mathbb{E}[X] = \frac{1}{p}.$$

2.5 Poisson Distribution

The Poisson distribution is a widely used probability distribution, typically used when counting the occurrences of certain events in an interval of time or space. The average number of events is given by the rate parameter, *lambda*.

Definition 17. A random variable X is said to be a Poisson random variable with parameter λ if its range is the set of non-negative integers, and its PMF is given by

$$\Pr(X = k) = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{for } k \in \{0, 1, 2, ...\}, \\ 0 & \text{otherwise.} \end{cases}$$

2.5.1 Expectation of Poisson R.V.

Since a Poisson random variable is non-negative,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\
&= e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^{(j+1)}}{j!} && \text{(by letting } j = k-1\text{)} \\
&= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \\
&= \lambda e^{-\lambda} e^{\lambda} && \text{(Taylor series for } e^{\lambda}\text{)} \\
&= \lambda.
\end{aligned}$$

It is not surprising that the expected value of the Poisson random variable is $\mathbb{E}[X] = \lambda$, since we introduced its parameter λ as the average number of events.

2.6 Conditional Expectation

Definition 18. Consider two random variables Y and Z. The conditional expectation is given as

$$\mathbb{E}[Y|Z = z] = \sum_{y} y \Pr(Y = y|Z = z),$$

where the summation is over all y in the range of Y.

The difference between unconditional and conditional expectation is that, now each value is weighted by the *conditional probability* that the variable assumes that value.

Also, $Z = z$ is an event. We can similarly define the conditional expectation of a random variable Y conditioned on an event E too.

Lemma 19. *For any random variables X and Y, the unconditional expectation is given as*

$$\mathbb{E}[X] = \sum_{y} \Pr(Y = y)\mathbb{E}[X|Y = y],$$

where the sum is over all values in the range of Y and all of the expectations exist.

Proof.

$$\begin{aligned}
\sum_{y} \Pr(Y = y)\mathbb{E}[X|Y = y] &= \sum_{y} \Pr(Y = y) \sum_{x} \Pr(X = x|Y = y) \\
&= \sum_{x} \sum_{y} x \Pr(X = x|Y = y)\Pr(Y = y) \\
&= \sum_{x} \sum_{y} x \Pr(X = x \cap Y = y) \\
&= \sum_{x} \Pr(X = x) = \mathbb{E}[X].
\end{aligned}$$

The linearity of expectations and multiplication of a constant to a random variable also extends to conditional expectations. □

Definition 20. The expression $\mathbb{E}[Y|Z]$ is a random variable $f(Z)$ that takes on the value $\mathbb{E}[Y|Z=z]$ when $Z=z$.

It is important to note that $\mathbb{E}[Y|Z]$ is not a real value, but is rather a function of the random variable Z. Hence, $\mathbb{E}[Y|Z]$ is itself a function from the sample space to the real numbers and can therefore be thought of as a random variable. On the other hand, $\mathbb{E}[Y|Z=z]$ is not a random variable, as the event $Z=z$ has already happened.

Theorem 21 (Law of Iterated (Total) Expectation)**.** *The Law of Iterated Expectation or the Adam's Law states that*

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|Z]].$$

Proof. We know that $\mathbb{E}[Y|Z]$ is a function of Z that takes on the value of $\mathbb{E}[Y|Z=z]$ when $Z=z$, and is hence, a random variable. Applying *LOTUS* on this random variable,

$$\mathbb{E}[\mathbb{E}[Y|Z]] = \sum_z \mathbb{E}[Y|Z=z]\Pr(Z=z).$$

By Lemma 19, the right-hand side equals $\mathbb{E}[Y]$. □

2.6.1 Properties of Conditional Expectations

Let X, Y and Z be random variables, $a, b \in \mathbf{R}, g : \mathbf{R} \to \mathbf{R}$, then assuming all the expectations exist,

1. (Linearity) $\mathbb{E}[\alpha X + \beta Y|Z] = \alpha\mathbb{E}[X|Z] + \beta\mathbb{E}[Y|Z]$.
2. (Monotonicity) $X \leq Y \implies E[X|Z] \leq E[Y|Z]$. In particular, if $X \geq 0 \implies E[X] \geq 0$.
3. (Independence) If X and Z are independent, then $E[X|Z] = E[X]$.
4. (Law of Total Expectation/Adam's Law) $\mathbb{E}_Y[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.
5. (Law of Total Variance/Eve's Law) $Var(X) = \mathbb{E}_Y[Var(X|Y)] + Var_Y(\mathbb{E}[X|Y])$.
6. $\mathbb{E}[Xg(Y)|Y] = g(Y)\mathbb{E}[X|Y]$. In particular, $\mathbb{E}[g(Y)|Y] = g(Y)$.
7. $\mathbb{E}[X|Y, g(Y)] = \mathbb{E}[X|Y]$.
8. $\mathbb{E}[\mathbb{E}[X|Y,Z]|Y] = \mathbb{E}[X|Y]$.

3 Moments and Deviation

Definition 22. The k^{th} moment a random variable X is $\mathbb{E}[X^k]$.

The **moments** of a random variable (or of its distribution) are the expected values of powers or related functions of the random variable. In particular, the first moment is the mean, $\mu_X = E[X]$.

Definition 23. The **variance** of a random variable X is defined as:

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

The standard deviation of a random variable X is $\sigma(X) = \sqrt{Var(X)}$.

If a random variable X is constant (so that it always assumes the same value), then its variance and standard deviation are both zero. Also, the variance of a random variable is always a positive value.

Definition 24. The covariance of two random variables X and Y is

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Theorem 25. *For any two random variables X and Y,*

$$Var[X+Y] = Var[X] + Var[Y] + 2Cov(X,Y)$$

Proof.

$$\begin{aligned}
Var(X+Y) &= \mathbb{E}[(X+Y-\mathbb{E}[X+Y])^2] \\
&= \mathbb{E}[(X+Y-\mathbb{E}[X]-\mathbb{E}[Y])^2] \\
&= \mathbb{E}[(X-\mathbb{E}[X])2 + (Y-\mathbb{E}[Y])2 + 2(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \mathbb{E}[(X-\mathbb{E}[X])2] + \mathbb{E}[(Y-\mathbb{E}[Y])2] + 2\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= Var[X] + Var[Y] + 2Cov(X,Y).
\end{aligned}$$

The extension of this theorem to a sum of any finite number of random variables can also be proven. □

Theorem 26. *If X and Y are two independent random variables, then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Proof. In the summations that follow, let i take on all values in the range of X, and let j take on all values in the range of Y:

$$\begin{aligned}
\mathbb{E}[X.Y] &= \sum_i \sum_j (i.j)\Pr((X=i)\cap(Y=j)) \\
&= \sum_i \sum_j (i.j)\Pr(X=i)\Pr(Y=j) \\
&= \left(\sum_i i\Pr(X=i)\right)\left(\sum_j j\Pr(Y=j)\right) \\
&= \mathbb{E}[X]\mathbb{E}[Y],
\end{aligned}$$

where the independence of X and Y is used in the second line. □

Corollary 27. *If X and Y are independent random variables, then*

$$Cov(X,Y) = 0$$

and

$$Var[X+Y] = Var[X] + Var[Y].$$

Proof.

$$\begin{aligned}
Cov(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \mathbb{E}[X-\mathbb{E}[X]].\mathbb{E}[Y-\mathbb{E}[Y]] \\
&= \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]].\mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y]] \\
&= (\mathbb{E}[X]-\mathbb{E}[X])(\mathbb{E}[Y]-\mathbb{E}[Y]) \\
&= 0.
\end{aligned}$$

For the last equation, we use the fact that for any random variable Z, $\mathbb{E}[\mathbb{E}[Z]] = \mathbb{E}[Z]$. Also, since $Cov(X,Y) = 0$, we have $Var(X+Y) = Var(X) + Var(Y)$. □

3.1 Cauchy-Schwarz Inequality

The same Cauchy-Schwarz inequality in Linear Algebra is valid for random variables.

Definition 28. For any two random variables X and Y, we have

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]},$$

where the equality holds if and only if $X = \alpha Y$, for some constant $\alpha \in \mathbf{R}$.

3.2 Jensen's Inequality

Recall that variance of every random variable X is a positive value, *i.e.*,

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \geq 0.$$

Thus,

$$\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2.$$

Definition 29 (Convex Function)**.** Consider a function $g : I \to R$, where I is an interval in R. We say that g is a *convex* function if, for any two points x and y in I and any $\alpha \in [0, 1]$, we have

$$g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y).$$

We say that g is *concave* if

$$g(\alpha x + (1 - \alpha)y) \geq \alpha g(x) + (1 - \alpha)g(y).$$

Note that in the above definition the term $\alpha x + (1 - \alpha)y$ is the weighted average of x and y. Also, $\alpha g(x) + (1 - \alpha)g(y)$ is the weighted average of $g(x)$ and $g(y)$.

Theorem 30. *If $g(x)$ is a convex function, and $\mathbb{E}[g(X)]$ and $g(\mathbb{E}[X])$ are finite, then*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]).$$

Jensen's inequality can be proven using second order Taylor's approximation around the mean of the random variable X, $\mathbb{E}[X]$.

To use Jensen's inequality, we need to determine if the function g is convex. A useful method is to use the second derivative (Hessian) test.

3.3 Markov's Inequality

We now examine techniques for bounding the *tail distribution*, *i.e.*, the probability that a random variable assumes values that are far from its expectation.

Theorem 31 (Markov's Inequality)**.** *Let X be a random variable that assumes only non-negative values. Then, for any $a > 0$,*

$$\Pr(X > a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. For some $a > 0$, consider an indicator random variable I,

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

Since I is an indicator random variable, $\mathbb{E}[I] = \Pr(I = 1) = \Pr(X \geq a)$ Also, since X is a non-negative random variable, $X \geq 0$. So,

$$I \leq \frac{X}{a}$$

Taking expectations on both sides,

$$\mathbb{E}[I] = \Pr(X \geq a) \leq \mathbb{E}[\frac{X}{a}] = \frac{\mathbb{E}[X]}{a}.$$ □

Markov's inequality is often too weak to yield useful results, but it is a fundamental tool in developing more sophisticated bounds.

3.4 Chebyshev's Inequality

Using the expectation and the variance of the random variable, a significantly stronger tail bound known as Chebyshev's inequality can be derived.

Theorem 32 (Chebyshev's Inequality)**.** *For any $a > 0$,*

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{Var(X)}{a^2}.$$

Proof. Observe that

$$\Pr(|X - \mathbb{E}[X]| \geq a) = \Pr((X - \mathbb{E}[X])^2 \geq a^2).$$

Notice that $(X - \mathbb{E}[X])^2$ is a non-negative random variable. We can now apply Markov's inequality to get:

$$\Pr((X - \mathbb{E}[X])^2 \geq a^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} = \frac{Var(X)}{a^2}.$$ □

Example 33. *Consider n independent unbiased coin tosses and a random variable X that counts the number of heads. We are interested in obtaining the probability bound of getting more than $3n/4$ heads in the coin flip sequence.*

Here, X is binomial random variable with parameters n and $p = 1/2$. Also, $\mathbb{E}[X] = \frac{n}{2}$ and $Var(X) = \frac{n}{4}$.

1. **Markov's Inequality:** Applying Markov's inequality, we obtain

   $$\Pr(X \geq 3n/4) \leq \frac{\mathbb{E}[X]}{3n/4} = \frac{n/2}{3n/4} = \frac{2}{3}.$$

   Markov's inequality gives us a very crude bound which is not dependent on the value of n. Naturally, as n keeps on growing, the probability of X assuming a value far from its expected value will keep on decreasing (Law of Large Numbers). We would like our bounds to reflect this property.

2. **Chebyshev's Inequality:** Recall that we can model the binomial random variable X as sum of n iid indicator random variables, such that, $X_i = 1$ if the i^{th} coin flip is heads and 0 otherwise. To use Chebyshev's inequality we need to compute the variance of X. Also, observe that since X_i is an indicator random variable, X_i only assumes the values 0 and 1, and so, $X_i^2 = X_i$. Thus,

   $$\mathbb{E}[X_i^2] = \mathbb{E}[X_i] = \frac{1}{2}.$$

   $$Var(X_i) = \mathbb{E}[X_i^2] - (\mathbb{E}[X_i])^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

   Applying Chebyshev's Inequality yields,

   $$\begin{aligned}
   \Pr(X \geq 3n/4) &\leq \Pr(|X - n/4| \geq 3n/4 - n/4) \\
   &= \Pr(|X - \mathbb{E}[X]| \geq n/4) \\
   &\leq \frac{Var(X)}{(n/4)^2} \\
   &= \frac{n/4}{(n/4)^2} \\
   &= \frac{4}{n}.
   \end{aligned}$$

Chebyshev's inequality gives a significantly better bound than Markov's inequality for large n.

4 Chernoff and Hoeffding Bounds

This section introduces large deviation bounds, commonly called *Chernoff* and *Hoeffding* bounds. These bounds are extremely powerful, giving exponentially decreasing bounds on the tail distribution. These bounds are derived by applying Markov's inequality to the Moment Generating Function (MGF) of a random variable, so we look at MGFs first.

4.1 Moment Generating Functions

Definition 34. The moment generating function of a random variable X is

$$M_X(t) = \mathbb{E}[e^{tX}]$$

From the Taylor Series of e^{tX}, it can be shown that the k^{th} moment of X is the coefficient of $\frac{t^k}{k!}$ in the Taylor series of $M_X(t)$. Also, from calculus, we know that the coefficient of $\frac{t^k}{k!}$ in the Taylor series of $M_X(t)$ is obtained by taking the k^{th} derivative of $M_X(t)$ and evaluating it at $t = 0$. Thus,

$$E[X^k] = \frac{d^k}{dt^k} M_X(t) \Big|_{t=0}.$$

Theorem 35. *If X and Y are independent random variables, then*

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

Proof.

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX} e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = M_X(t) M_Y(t).$$

Here we have used that X and Y are independent, and hence, e^{tX} and e^{tY} are independent, to conclude that

$$\mathbb{E}[e^{tX} e^{tY}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}].$$

□

4.2 Chernoff Bounds for the Sum of Poisson Trials

The most commonly used version of the Chernoff bound is for the tail distribution of a sum of independent *(not necessarily identical)* indicator random variables, also known as *Poisson trials*. Note that Poisson trials are different from Poisson random variables. *Bernoulli trials* (as in the Binomial distribution) are a special case of Poisson trials where the independent indicator random variables have the same distribution (hence, iid).

We now derive Let $X_1, X_2, ..., X_n$ be a sequence of independent Poisson trials with $\Pr(X_i = 1) = p_i$. Also, let $X = \sum_{i=1}^{n} X_i$. Then,

$$\mu = \mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p_i.$$

For some $\delta > 0$, we're interested in finding the probability that X deviates from its expectation μ by $\delta\mu$ or more, i.e., the bounds on $\Pr(X \geq (1+\delta)\mu)$ and $\Pr(X \leq (1-\delta)\mu)$.

To do so, we first determine a bound on the MGF of X. For each X_i,

$$M_{X_i}(t) = \mathbb{E}[e^{tX_i}] = p_i e^t + (1-p_i)e^0 = 1 + p_i(e^t - 1) \leq e^{p_i(e^t - 1)},$$

where the last inequality results from the fact that, for any y, $1+y \leq e^y$. Since these n trials are independent, the MGF of X is the product of the n MGFs of the Poisson trials:

$$\begin{aligned} M_X(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &\leq \prod_{i=1}^{n} e^{p_i(e^t-1)} \\ &= \exp\left\{\sum_{i=1}^{n} p_i(e^t-1)\right\} \\ &= e^{(e^t-1)\mu}. \end{aligned}$$

Using the bound on the MGF, we now state the Chernoff bound for a sum of Poisson trials.

Theorem 36 (Chernoff bounds on deviation above the mean)**.** *Let $X_1, X_2, ..., X_n$ be independent Poisson trials such that $\Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. Then,*

1. *for any $\delta > 0$,*
$$\Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^\mu;$$
2. *for $0 < \delta \leq 1$,*
$$\Pr(X \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/3};$$
3. *for $R \geq 6\mu$,*
$$\Pr(X \geq (1+\delta)\mu) \leq 2^{-R}.$$

Theorem 37 (Chernoff bounds on deviation below the mean)**.** *Let $X_1, X_2, ..., X_n$ be independent Poisson trials such that $\Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. Then, for $0 < \delta < 1$:*

1. $\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^\mu$;
2. $\Pr(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}$.

Corollary 38. *Let $X_1, X_2, ..., X_n$ be independent Poisson trials such that $\Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. Then, for $0 < \delta < 1$:*

$$\Pr(|X - \mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}.$$

Theorem 39 (Additive bounds when indicators are iid)**.** *Let $X_1, X_2, ..., X_n$ be iid trials such that for all X_i, $\Pr(X_i = 1) = p$. Then, for any $\varepsilon > 0$,*

1. $\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + \varepsilon\right) \leq \left(\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right)^n$;
2. $\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \leq p - \varepsilon\right) \leq \left(\left(\frac{p}{p-\varepsilon}\right)^{p-\varepsilon}\left(\frac{1-p}{1-p+\varepsilon}\right)^{1-p+\varepsilon}\right)^n$.

4.3 Special cases of Chernoff Bounds

Theorem 40. *Let $X_1, X_2, ..., X_n$ be independent random variables (not 0-1) with $X = \sum_{i=1}^{n} X_i$ and $\Pr(X_i = 1) = \Pr(X_i = -1) = \frac{1}{2}$, then for all $a > 0$,*

1. $\Pr(X \geq a) \leq e^{-a^2/2n}$,
2. $\Pr(X \leq -a) \leq e^{-a^2/2n}$,
3. $\Pr(|X| \geq a) \leq 2e^{a^2/2n}$.

Applying the transformation $Y_i = (X_i + 1)/2$ allows us to prove the following.

Corollary 41. *Let $Y_1, Y_2, ..., Y_n$ be independent random variables with*

$$\Pr(Y_i = 1) = \Pr(Y_i = 0) = \frac{1}{2}.$$

Let $Y = \sum_{i=1}^{n} Y_i$ and $\mu = \mathbb{E}[Y] = n/2$.

1. *For any $a > 0$,*
$$\Pr(Y \geq \mu + a) \leq e^{-2a^2/n}.$$
2. *For any $\delta > 0$,*
$$\Pr(Y \geq (1+\delta)\mu) \leq e^{-\delta^2\mu}.$$
3. *For any $0 < a < \mu$,*
$$\Pr(Y \leq \mu - a) \leq e^{-2a^2/n}.$$
4. *For any $0 < \delta < 1$,*
$$\Pr(Y \leq (1-\delta)\mu) \leq e^{-\delta^2\mu}.$$

These bounds hold for the more general setting of the sum of Poisson trials and hence, will also hold for the Binomial distribution.

4.4 Example: Coin Flips

Let X be the number of heads in a sequence of n independent fair coin flips. Applying the Chernoff bound $\Pr(|X - \mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3}$, we have

$$\begin{aligned}\Pr\left(\left| X - \frac{n}{2} \right| \geq \frac{1}{2}\sqrt{6n \ln n}\right) &\leq 2\exp\left\{-\frac{1}{3}\frac{n}{2}\frac{6\ln n}{n}\right\} \\ &= \frac{2}{n}.\end{aligned}$$

4.5 The Hoeffding Bound

Hoeffding's bound extends the Chernoff bound technique to general random variables with a bounded range.

Theorem 42. *Let $X_1, X_2, ..., X_n$ be independent random variables such that for all $1 \leq i \leq n$, $\mathbb{E}[X_i] = \mu$ and $\Pr(a \leq X_i \leq b) = 1$. Then,*

$$\Pr\left(\left| \frac{1}{n}\sum_{i=1}^{n} X_i - \mu \right| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2/(b-a)^2}.$$

The above theorem bounds the deviation of the *average* of the n random variables. The next theorem bounds the deviation of the *sum* of n random variables.

Theorem 43. *Let $X_1, X_2, ..., X_n$ be independent random variables with $\mathbb{E}[X_i] = \mu_i$ and $\Pr(a_i \leq X_i \leq b_i) = 1$ for constants a_i and b_i. Then,*

$$\Pr\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right| \geq \varepsilon\right) \leq 2e^{-2\varepsilon^2 / \sum_{i=1}^{n} (b_i - a_i)^2}.$$

4.6 Independent Bounded Difference Inequality

Let $X = (X_1, X_2, ..., X_n)$ be a family of independent random variables with $X_j \in A_j$ for $j = 1, 2, ..., n$, and $\phi : \prod_{j=1}^{n} A_j \to \mathbf{R}$ be a function such that $\phi(x) - \phi(x') \leq c_j$ whenever the vectors x and x' differ only in the j^{th} coordinate. Then $\forall\ t \geq 0$,

$$\Pr(\phi(x) - \mathbb{E}[\phi(x)] \geq t) \leq \frac{e^{-2t^2}}{\sum_{i=1}^{n} c_j^2}.$$