1 Probability basics

2 Basics of Convexity

3 Basic Probabilistic Inequalities

4 Concentration Inequalities

5 Relative Entropy Inequalities
Sample Space and Events

- **Random Experiment**: An experiment whose outcome is unknown.
- **Sample Space** \( \equiv \Omega \): Set of all possible outcomes of a random experiment.
- **Events**: Any subset of the Sample Space.
  - E.g. Roll a die. \( E_1 = \{ \text{Outcome} \in (1, 2) \} \), \( E_2 = \{ \text{Outcome} \leq 4 \} \).
Frequentist viewpoint. To compute the probability of an event $A \subset \Omega$, count the number of occurrences of $A$ in $N$ random experiments. Then $P(A) = \lim_{n \to \infty} \frac{N(A)}{N}$.

E.g. A coin was tossed 100 times and 51 heads were counted. $P(A) \sim \frac{51}{100}$. By W.L.L.N (see later slides) this estimate converges to the actual probability of heads. If its a fair coin, this should be 0.5.

If $A$ and $B$ are two disjoint events, $N(A \cup B) = N(A) + N(B)$. Thus, the frequentist viewpoint seems to imply $P(A \cup B) = P(A) + P(B)$ if $A, B$ are disjoint events.
Axioms of Probability

1. \( P(\emptyset) = 0 \)
2. \( P(\Omega) = 1 \)
3. If \( A_1, A_2, \ldots, A_k \) are a collection of mutually disjoint events (i.e. \( A_i \subset \Omega, A_i \cap A_j = \emptyset \)). Then, \( P(\bigcup A_i) = \sum_{i=1}^{k} P(A_i) \).

E.g.: Two simultaneous ‘fair’ coin tosses. Sample Space: \( \{(H, H), (H, T), (T, H), (T, T)\} \).
\[ P(\text{Atleast one head}) = P((H, H) \cup (H, T) \cup (T, H)) = 0.75. \]

E.g. Roll of a die. \( P(\text{Outcome} \leq 3) = P(\text{Outcome} = 1) + P(\text{Outcome} = 2) + P(\text{Outcome} = 3) = 0.5. \)
Conditional Probability, Independence, R.V.

- **Conditional Probability**: Let $\Omega$ be a sample space associated with a random experiment. Let $A, B$ be two events. Then define the *Conditional Probability of $A$ given $B$* as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- **Independence**: Two events $A, B$ are independent if

$$P(A \cap B) = P(A)P(B).$$

Stated another way, $A, B$ are independent if

$$P(A|B) = P(A), P(B|A) = P(B).$$

Note that $A, B$ are *independent* necessarily means that $A, B$ are not mutually exclusive i.e. $A \cap B \neq \emptyset$. 
**Example**: Consider simultaneous flips of two ‘fair’ coins. Event A: First Coin was a head. Event B: Second coin was a tail. Note that $A$ and $B$ are indeed *independent* events but are not mutually exclusive. On the other hand, let Event C: First coin was a tail. Then $A$ and $C$ are not independent. Infact they are mutually exclusive - i.e. knowing that one has occurred (e.g. heads) implies the other has not (tails).
Bayes’ rule

A simple outcome of the definition of Conditional Probability is the very useful **Bayes’ rule**:

\[
P(A \cap B) = P(A|B)P(B) = P(B|A)Prob(A)
\]

Let’s say we are given \(P(A|B_1)\) and we need to compute \(P(B_1|A)\). Let \(B_1, \ldots B_k\) be disjoint events the probability of whose union is 1. Then,

\[
P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A)} = \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{k} P(A \cap B_i)} = \frac{P(A|B_1)P(B_1)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}
\]

With additional information on \(P(B_i)\) and \(P(A|B_i)\), the required conditional probability is easily computed.
Random Variables

- **Random Variable**: A random variable is a mapping from the sample space to the real line, i.e. \( X : \Omega \rightarrow \mathbb{R} \). Let \( x \in \mathbb{R} \), then \( P(X(w) = x) = P\{w \in \Omega : X(w) = x\} \). Although random variable is a function, in practice it is simply represented as \( X \). It is very common to directly define probabilities over the range of the random variable.

- **Cumulative Distribution Function (CDF)**: Every random variable \( X \) has associated with it a CDF: \( F : \mathbb{R} \rightarrow [0, 1] \) such that \( P(X \leq x) = F(x) \). If \( X \) is a continuous random variable, then its Probability density function is given by \( f(x) = \frac{d}{dx} F(x) \).
**Probability Mass Function (p.m.f):** Every discrete random variable has associated with it, a probability mass function which outputs the probability of the random variable taking a particular value. E.g. if $X$ denotes the R.V. corresponding to the role of a die, $P(X = i) = \frac{1}{6}$ is the p.m.f associated with $X$.

**Expectation:** The Expectation of a discrete random variable $X$ with *probability mass function* $p(X)$ is given by $E[X] = \sum_x p(x)x$. Thus expectation is the weighted average of the values that a random variable can take (where the weights are given by the probabilities). E.g. Let $X$ be a bernoulli random variable with $P(X = 0) = p$. Then $E[X] = p \times 0 + (1 - p) \times 1 = 1 - p$. 
More on Expectation

- **Conditional Expectation**: Let $X, Y$ be two random variables. Given that $Y = y$, we can talk of the expectation of $X$ *conditioned* on the information that $Y = y$:

$$E[X|Y = y] = \sum_{x} P(X = x|Y = y)x$$

- **Law of Total Expectation**: Let $X, Y$ be two random-variables. The Law of Total Expectation then states that:

$$E[X] = E[E[X|Y]]$$

The above equation can be interpreted as: The Expectation of $X$ can be computed by *first* computing the Expectation of the *randomness* present only in $X$ (but not in $Y$) and *next* computing the Expectation with respect to the randomness in $Y$. Thus, the ‘Total Expectation’ w.r.t $X$ can be computed by evaluating a ‘sequence of partial Expectations’.
Stochastic Process

- **Stochastic Process**: Is defined as an indexed collection of random variables. A stochastic process is denoted by $\{X_i\}_{i=1}^N$.

- E.g. In $N$ flips of a fair coin, each flip is a random variable. Thus the collection of these $N$ random variables forms a stochastic process.

- **Independent and Identically distributed (i.i.d)**. A stochastic process is said to be i.i.d if each of the random variables, $X_i$ have the same CDF and in addition the random variables $(X_1, X_2, \ldots, X_N)$ are independent. Independence implies that $\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_N) = \Pr(X_1 = x_1)\Pr(X_2 = x_2)\ldots\Pr(X_N = x_N)$. 
Binomial Distribution. Consider the i.i.d stochastic process \( \{X_i\}_{i=1}^{N} \), where \( X_i \) is a bernoulli random variable. I.e. \( X_i \in \{0, 1\} \) and \( P(X_i = 1) = p \). Let \( X = \sum_{i=1}^{n} X_i \). Consider the probability, \( P(X = k) \) for some \( 0 \leq k \leq n \). Then,

\[
P(X = k) = P(\cup_{1 \leq i_1 < i_2 < \ldots < i_k \leq n}(X_1 = 0, X_2 = 0, \ldots, X_{i_1} = 1, X_{i_1+1} = 0, \ldots, X_{i_2} = 1, \ldots, X_{i_k} = 1, \ldots, X_n = 0)).
\]

Thus the probability to be computed is a union of \( \binom{n}{k} \) disjoint events and hence can be expressed as sum of probabilities of \( \binom{n}{k} \) events. Also, since the stochastic process is i.i.d, each event in the sum has a probability of \( p^k(1 - p)^{n-k} \). Combining the two, we have that \( P(X = k) = \binom{n}{k} p^k(1 - p)^{n-k} \), which is exactly the probability distribution associated with a Binomial Random Variable.
Convex Set

A set $C$ is a convex set if for any two elements $x, y \in C$ and any $0 \leq \theta \leq 1$, $\theta x + (1 - \theta)y \in C$. 

![Diagram of convex sets](image-url)
Convex Function

A function $f : X \rightarrow Y$ is convex if $X$ is convex and for any two $x, y \in X$ and any $0 \leq \theta \leq 1$ it holds that, $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$.

Examples: $e^x, -\log(x), x\log x$.

Second order condition

A function $f$ is convex if and only if its hessian, $\nabla^2 f \succeq 0$.

For functions in one variable, this reduces to checking if second derivative is non-negative.
First order condition

\( f : X \rightarrow Y \) is convex if and only if it is bound below by its first order approximation at any point:

\[
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle \quad \forall x, y \in X
\]
Operations preserving convexity

- Intersection of Convex sets is convex. E.g.: Every polyhedron is an intersection of finite number of half-spaces and is therefore convex.

- Sum of Convex functions is convex. E.g. The squared 2-norm of a vector, \( \|x\|_2^2 \) is the sum of quadratic functions of the form \( y^2 \) and hence is convex.

- If \( h, g \) are convex functions and \( h \) is increasing, then the composition \( f = h \circ g \) is convex. E.g. \( f(x) = \exp(Ax + b) \) is convex.

- Max of convex functions is convex. E.g.

\[
\begin{align*}
  f(y) &= \max_x x^T y \\
  &\text{s.t. } x \in C.
\end{align*}
\]

- Negative Entropy: \( f(x) = \sum_i x_i \log(x_i) \) is convex.
**Cauchy-Schwartz Inequality**

**Theorem 3.1**

Let $X$ and $Y$ be any two random variables. Then it holds that
\[ |E[XY]| \leq \sqrt{E[X^2]} \sqrt{E[Y^2]} . \]

**Proof.**

Let $U = X / \sqrt{E[X^2]}$, $V = Y / \sqrt{E[Y^2]}$. Note that $E[U^2] = E[V^2] = 1$. We need to show that $|E[UV]| \leq 1$. For any two $U$, $V$ it holds that
\[ |UV| \leq \frac{|U|^2 + |V|^2}{2} . \]
Now, $|E[UV]| \leq E[|UV|] \leq \frac{1}{2} E[|U|^2 + |V|^2] = 1. \]
Theorem 3.2 (Jensen’s inequality)

Let $X$ be any random variable and $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Then, it holds that $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.

Proof.

The proof uses the convexity of $f$. First note from the first order conditions for convexity, for any two $x, y \in \mathbb{R}$, $f(x) \geq f(y) + f'(y)(x - y)$.

Now, let $x = X$, $y = \mathbb{E}[X]$. Then we have that $f(X) \geq f(\mathbb{E}[X]) + f'(\mathbb{E}[X])(X - \mathbb{E}[X])$. Taking expectations on both sides of the previous inequality, it follows that, $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$. \qed
Applications of Jensen’s

- General Mathematical Inequalities: Cauchy Schwartz, Holder’s Inequality, Minkowski’s Inequality, AM-GM-HM inequality (HW 0).
- Economics - Exchange rate (HW 0).
- Statistics - Inequalities related to moments: Lyapunov’s inequality and sample-population standard deviations (HW 0).
- Information Theory - Relative Entropy, lower bounds, etc.
Application of Jensen’s Inequality

Note that Jensen’s inequality even holds for compositions. Let $g$ be any function, $f$ be a convex function. Then, $f(\mathbb{E}[g(X)]) \leq \mathbb{E}[f(g(X))]$.

Lemma 3.3 (Lyapunov’s inequality)

Let $X$ be any random variable. Let $0 < s < t$. Then

$$(\mathbb{E}[|X|^s])^{1/s} \leq (\mathbb{E}[|X|^t])^{1/t}.$$

Proof.

Since $s < t$, the function $f(x) = x^{t/s}$ is convex on $\mathbb{R}_+$. Thus,

$$(\mathbb{E}[|X|^s])^{t/s} = f(\mathbb{E}[|X|^s]) \leq \mathbb{E}[f(|X|^s)] = \mathbb{E}[|X|^t].$$

The result now follows by taking the $t^{th}$ root on both sides of the inequality.

Application of Jensen’s: AM-GM inequality

As a consequence of Lyapunov’s inequality, the following often used bound follows: $E[|X|] \leq \sqrt{E[|X|^2]}$. The AM-GM inequality also follows from Jensen’s. Let $f(x) = \log x$. Let $a_1, a_2, \ldots, a_n \geq 0$ and $p_i \geq 0$, $1^T p = 1$. Let $X$ be a random variable with distribution $p$ over $a_1, a_2, \ldots, a_n$. Then, $\sum_i p_i \log a_i = E[f(X)] \leq f(E[X]) = \log(\sum_i p_i a_i)$.

Or in other words,

$$\prod_i a_i^{p_i} \leq \sum_i p_i a_i$$

Setting $p_i = 1/n \ \forall i = 1, 2, \ldots, n$ the AM-GM inequality follows:

$$(\prod_{i=1}^n a_i)^{\frac{1}{n}} \leq \frac{\sum_i a_i}{n}.$$
The GM-HM inequality also follows from Jensen’s. I.e., for any \( a_1, a_2, \ldots, a_n > 0 \) it holds that:

\[
\left( \prod_{i=1}^{n} a_i \right)^\frac{1}{n} \geq \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}
\]

Exercise: Show this.
Markov’s Inequality

Theorem 4.1 (Markov’s inequality)

Let $X$ be any non-negative random variable. Then for any $t > 0$ it holds that,

$$P(X > t) \leq \frac{E[X]}{t}$$

Proof.

$$E[X] = \int_{0}^{\infty} x f(x) \, dx$$
$$= \int_{0}^{t} x f(x) \, dx + \int_{t}^{\infty} x f(x) \, dx$$
$$\geq t P(X > t)$$
**Theorem 4.2 (Chebyshev’s inequality)**

Let $X$ be any random variable. Then for any $s > 0$ it holds that,

$$P(|X - E[X]| > s) \leq \frac{\text{Var}[X]}{s^2}$$

**Proof.**

This is an application of Markov’s inequality. Set $Z = |X - E(X)|^2$ and apply Markov’s inequality to $Z$ with $t = s^2$. Note that $E[Z] = \text{Var}[X]$. 

□
Application to Weak Law of Large Numbers

As a consequence of the Chebyshev inequality we have the *Weak Law of Large numbers*.

**WLLN**

Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables. Then,

$$\lim \frac{1}{n} \sum_{i} X_i \rightarrow^p E[X_i]$$

**Proof.**

W.log assume $X_i$ has mean 0 and variance 1. Then variance of $X = \frac{1}{n} \sum_{i=1}^{n} X_i$ equals $\frac{1}{n}$. By Chebyshev’s inequality as $n \rightarrow \infty$, $X$ concentrates around its mean w.h.p.
Relative entropy or KL-divergence

Let $X$ and $Y$ be two discrete random variables with probability distributions $p(X), q(Y)$. Then the relative entropy between $X$ and $Y$ is given by

$$D(X||Y) = E_{p(X)} \log \frac{p(X)}{q(X)}$$
Non-negativity of Relative Entropy

Lemma 5.1

For any two discrete random variables, $X, Y$ with distributions $p(X), q(Y)$ respectively it holds that $D(X||Y) \geq 0$.

Proof.

The proof follows from a simple application of Jensen’s inequality (Theorem 3.2). Note that $f(x) = \log x$ is concave. Let $g(X) = \frac{q(X)}{p(X)}$. Then,

$$-D(X||Y) = \mathbb{E}_{p(X)} \log \frac{q(X)}{p(X)}$$

$$= \mathbb{E}_{p(X)} \log g(X)$$

$$\leq \log \mathbb{E}_{p(X)} g(X)$$

$$= 0$$
A tighter bound

Lemma 5.2

Let $X, Y$ - Binary R.V.s with probability of success $p, q$ respectively. Then

$$D(p||q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q} \geq 2(p - q)^2.$$ 

Proof.

Note that $x(1 - x) \leq \frac{1}{4}$ for all $x$. Hence,

$$D(p||q) = \int_q^p \left( \frac{p}{x} - \frac{1-p}{1-x} \right) dx$$

$$= \int_q^p \frac{p(1-x)-(1-p)(x)}{x(1-x)} dx$$

$$\geq 4 \int_q^p (p - x) dx$$

$$= 2(p - q)^2$$
Summary

- We reviewed the basics of probability: Sample Space, Random variables, independence, conditional probability, expectation, stochastic process, etc.

- Convexity is fundamental to some important probabilistic inequalities: E.g. Jensen’s inequality and Relative Entropy Inequality.

- Jensen’s inequality has many applications. A simple one being the AM-GM-HM inequality.

- Markov’s inequality is fundamental to concentration inequalities. The Weak Law of Large Numbers essentially follows from a variant of Markov’s inequality (Chebyshev’s Inequality).