Information Theory (5XSE0)
Ch.0: Mathematical Preliminaries

Hamdi Joudeh

TU/e (Q3 2020-2021)

1 Probability

Sets: Sets are denoted using calligraphic font, e.g. $\mathcal{A} = \{1, 2, 3, 4, 5\}$. If a is an element of $\mathcal{A}$, we write $a \in \mathcal{A}$. The number of elements in $\mathcal{A}$ (i.e. cardinality) is denoted by $|\mathcal{A}|$. If $\mathcal{S}$ is contained in $\mathcal{A}$, we write $\mathcal{S} \subseteq \mathcal{A}$, i.e. $\mathcal{S}$ is a subset of $\mathcal{A}$. Here $\mathcal{S}$ can be equal to $\mathcal{A}$. If $\mathcal{S}$ is *strictly* contained in $\mathcal{A}$, we write $\mathcal{S} \subset \mathcal{A}$. For a pair of sets $\mathcal{A}$ and $\mathcal{B}$, the *set difference* is defined as

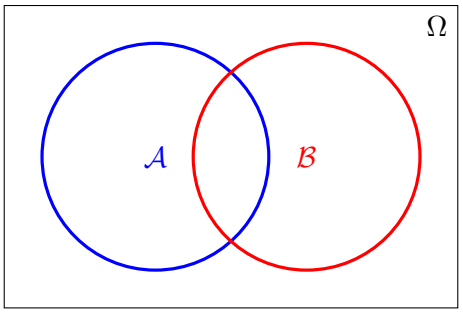
$$\mathcal{A} \setminus \mathcal{B} \equiv \{\text{elements in } \mathcal{A} \text{ and not in } \mathcal{B}\}.$$

The union of two sets is denoted by $\mathcal{A} \cup \mathcal{B}$, while their intersection is denoted by $\mathcal{A} \cap \mathcal{B}$. The empty set, which contains no elements, is denoted by $\emptyset$. If two sets do not overlap, then their intersection is the empty set, e.g. $\{1, 2, 3\} \cap \{4, 5, 6\} = \emptyset$. For any set $\mathcal{A}$, we have $\emptyset \cup \mathcal{A} = \mathcal{A}$, $\emptyset \cap \mathcal{A} = \emptyset$, and $\emptyset \subseteq \mathcal{A}$.

Probability Space: Consider a random experiment, e.g. tossing a coin, or rolling a dice. The *sample space* Ω is the set of all possible outcomes of the random experiment in question. In this course, we mostly work with discrete sample spaces, where Ω is finite or countably infinite, i.e. $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ for some $n \leq \infty$. An *event* is a subset of Ω. Events include the certain event Ω, and the empty event $\emptyset$ (or impossible event). An event $\mathcal{A}$ occurs when the outcome of the random experiment is an element in $\mathcal{A}$. A probability measure $\mathbb{P}$ assigns a real number between 0 and 1 to each event, such that

$$\mathbb{P}[\Omega] = 1, \quad \text{and} \quad \mathbb{P}[\mathcal{A} \cup \mathcal{B}] = \mathbb{P}[\mathcal{A}] + \mathbb{P}[\mathcal{B}] \text{ if } \mathcal{A} \cap \mathcal{B} = \emptyset.$$

The above implies that $\mathbb{P}[\emptyset] = 0$. Note that we write $\mathbb{P}[\mathcal{A}]$ or $\mathbb{P}\{\mathcal{A}\}$ to denote the probability of event $\mathcal{A}$. For an event $\mathcal{A} \subseteq \Omega$, the complement $\mathcal{A}^{\text{c}}$ is defined as $\mathcal{A}^{\text{c}} \equiv \Omega \setminus \mathcal{A}$.



Example 1. Consider a random experiment involving tossing a coin and observing the outcome. Here we have $\Omega = \{\text{H}, \text{T}\}$, where H denotes heads and T denotes tails. The set of all events (i.e. the *event space*) is given by $\{\emptyset, \{\text{H}\}, \{\text{T}\}, \{\text{H}, \text{T}\}\}$. If the coin is *fair*, then we will have $\mathbb{P}\big[\{\text{H}\}\big] = \mathbb{P}\big[\{\text{T}\}\big] = 0.5$.

Joint Probability: For two events $\mathcal{A}$ and $\mathcal{B}$, their joint probability is given by $\mathbb{P}[\mathcal{A} \cap \mathcal{B}]$, i.e. the probability that both $\mathcal{A}$ and $\mathcal{B}$ occur. In the above example, $\mathbb{P}[\{\text{H}\} \cap \{\text{T}\}] = 0$, as a coin cannot be simultaneously heads and tails. These are called *mutually exclusive* or *disjoint* events.

Independence: Two events $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if their joint probability is equal to the product of their probabilities, i.e.

$$\mathcal{A} \text{ and } \mathcal{B} \text{ are independent } \iff \mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{A}]\mathbb{P}[\mathcal{B}].$$

Note that mutual exclusiveness and independence are not the same thing.

Example 2. Consider another experiment of tossing a pair of fair coins and observing the outcomes. Here each outcome is an ordered pair and the sample space is given by $\Omega = \big\{(\mathrm{H},\mathrm{H}), (\mathrm{H},\mathrm{T}), (\mathrm{T},\mathrm{H}), (\mathrm{T},\mathrm{T})\big\}$. The event space consists of all $2^4 = 16$ subsets of Ω (including $\emptyset$ and Ω itself). Assuming coins are independent, each of the 4 outcomes will have a probability of 0.25. The event of observing at least one head is given by $\big\{(\mathrm{H},\mathrm{H}), (\mathrm{H},\mathrm{T}), (\mathrm{T},\mathrm{H})\big\}$. This has a probability of 0.75.

Conditional Probability: For events $\mathcal{A}$ and $\mathcal{B}$, the conditional probability of $\mathcal{A}$ given $\mathcal{B}$ is defined as

$$\mathbb{P}[\mathcal{A}|\mathcal{B}] \equiv \frac{\mathbb{P}[\mathcal{A} \cap \mathcal{B}]}{\mathbb{P}[\mathcal{B}]}.$$

Note that $\mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{A}|\mathcal{B}]\mathbb{P}[\mathcal{B}]$. Since $\mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{B} \cap \mathcal{A}]$, we can exchange the order and write $\mathbb{P}[\mathcal{A} \cap \mathcal{B}] = \mathbb{P}[\mathcal{B} \cap \mathcal{A}] = \mathbb{P}[\mathcal{B}|\mathcal{A}]\mathbb{P}[\mathcal{A}]$. This leads to Bayes' rule

$$\mathbb{P}[\mathcal{B}|\mathcal{A}] = \frac{\mathbb{P}[\mathcal{A}|\mathcal{B}]\mathbb{P}[\mathcal{B}]}{\mathbb{P}[\mathcal{A}]}.$$

Law of Total Probability: Let $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_n$ be a partition of the sample space, i.e. all n events are disjoint and their union is the sample space. Then

$$\mathbb{P}[\mathcal{A}] = \sum_{i=1}^{n} \mathbb{P}[\mathcal{A} \cap \mathcal{B}_i] = \sum_{i=1}^{n} \mathbb{P}[\mathcal{A} \mid \mathcal{B}_i]\mathbb{P}[\mathcal{B}_i].$$

As a special case, we have $\mathbb{P}[\mathcal{A}] = \mathbb{P}[\mathcal{A} \mid \mathcal{B}]\mathbb{P}[\mathcal{B}] + \mathbb{P}[\mathcal{A} \mid \mathcal{B}^{\mathrm{c}}]\mathbb{P}[\mathcal{B}^{\mathrm{c}}]$.

Example 3. Assume that there is a 1% chance of being infected with a particular virus and you take a test which is 95 % accurate, i.e. $\mathbb{P}[+\text{ve test} \mid \text{infected}] = \mathbb{P}[-\text{ve test} \mid \text{not infected}] = 0.95$. Given that you test positive, what is the probability that you have the virus?

Let $\mathcal{I}$ denote the event of being infected, and $\mathcal{P}$ be the event of testing positive. Note that $\mathcal{I}^{\mathrm{c}}$ is the event of not being infected, while $\mathcal{P}^{\mathrm{c}}$ is the event of testing negative. We have $\mathbb{P}[\mathcal{I}] = 0.01$ and $\mathbb{P}[\mathcal{P} \mid \mathcal{I}] = \mathbb{P}[\mathcal{P}^{\mathrm{c}} \mid \mathcal{I}^{\mathrm{c}}] = 0.95$. We want to find $\mathbb{P}[\mathcal{I} \mid \mathcal{P}]$. This is calculated as follows

$$\begin{aligned}
\mathbb{P}[\mathcal{I} \mid \mathcal{P}] &= \frac{\mathbb{P}[\mathcal{P} \mid \mathcal{I}]\mathbb{P}[\mathcal{I}]}{\mathbb{P}[\mathcal{P}]} \\
&= \frac{\mathbb{P}[\mathcal{P} \mid \mathcal{I}]\mathbb{P}[\mathcal{I}]}{\mathbb{P}[\mathcal{P} \mid \mathcal{I}]\mathbb{P}[\mathcal{I}] + \mathbb{P}[\mathcal{P} \mid \mathcal{I}^{\mathrm{c}}]\mathbb{P}[\mathcal{I}^{\mathrm{c}}]} \\
&= \frac{\mathbb{P}[\mathcal{P} \mid \mathcal{I}]\mathbb{P}[\mathcal{I}]}{\mathbb{P}[\mathcal{P} \mid \mathcal{I}]\mathbb{P}[\mathcal{I}] + \big(1 - \mathbb{P}[\mathcal{P}^{\mathrm{c}} \mid \mathcal{I}^{\mathrm{c}}]\big)\big(1 - \mathbb{P}[\mathcal{I}]\big)} \\
&= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.1 \times 0.99} = 0.16.
\end{aligned}$$

Despite testing positive, the probability that you are infected is only 16%.

Chain Rule: From the definition of the conditional probability, we get

$$\mathbb{P}[\mathcal{A}_1 \cap \mathcal{A}_2 \cap \cdots \cap \mathcal{A}_n] = \mathbb{P}[\mathcal{A}_1]\mathbb{P}[\mathcal{A}_2|\mathcal{A}_1]\mathbb{P}[\mathcal{A}_3|\mathcal{A}_1 \cap \mathcal{A}_2] \cdots \mathbb{P}[\mathcal{A}_n|\mathcal{A}_1 \cap \mathcal{A}_2 \cap \cdots \cap \mathcal{A}_{n-1}].$$

Note that this expansion can also be written in reverse order (or any other order of events).

Union Bound: The following inequality holds (verify it!)

$$\mathbb{P}[\mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_n] \leq \sum_{i=1}^{n} \mathbb{P}[\mathcal{A}_i].$$

Random Variables: A random variable is a function that assigns numerical values to outcomes of a random experiment, e.g. $X : \Omega \to \mathcal{X}$ maps each outcome ω in Ω to a corresponding value $X(\omega)$ from the set $\mathcal{X}$. For instance, we can define a binary random variable for the coin toss experiment from earlier, such that $\mathcal{X} = \{0, 1\}$, $X(\mathrm{H}) = 1$ and $X(\mathrm{T}) = 0$. From now on, the outcome argument ω in $X(\omega)$ will be dropped, and a random variable $X(\omega)$ will be simply denoted by X.

2 Random Variables

Random variables are denoted by uppercase letters, e.g. X, Y, Z. A random variable X takes values on a set $\mathcal{X}$, which we often refer to as the *alphabet* of X. In this course, we mainly focus on real-valued discrete random variables, i.e. $\mathcal{X}$ is a countable set and its elements are real numbers.

Probability Mass Function: The probability mass function (pmf) of X is given by

$$p_X(x) \equiv \mathbb{P}[X = x], \text{ for all } x \in \mathcal{X}$$

which assigns a probability between 0 and 1 to each value in $\mathcal{X}$. A pmf satisfies

$$\sum_{x \in \mathcal{X}} p_X(x) = 1 \quad \text{and} \quad p_X(x) \geq 0, \text{ for all } x \in \mathcal{X}.$$

A pmf $p_X(x)$ is also referred to as the *distribution* of X. For brevity, we will often drop the subscript in $p_X(x)$ and write it as $p(x)$, where the argument x identifies $p(x)$ as a pmf of X.

Expectation: The expected value of a random variable X is defined as

$$\mathbb{E}X \equiv \sum_{x \in \mathcal{X}} x p_X(x).$$

We sometimes use brackets as $\mathbb{E}[X]$. If $f(X)$ is some function of X, then $\mathbb{E}\big[f(X)\big] = \sum_{x \in \mathcal{X}} p_X(x) f(x)$.

Variance: The variance of a random variable X is defined as

$$\mathrm{var}(X) \equiv \mathbb{E}\left[(X - \mathbb{E}X)^2\right]$$

The variance is also equivalently given by $\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Verify this!

Example 4. (Bernoulli) A Bernoulli random variable is a binary random variable with an alphabet $\mathcal{X} = \{0, 1\}$ and a pmf of

$$p_X(x) = \begin{cases} 1 - p, & x = 0 \\ p, & x = 1 \end{cases}$$

where $0 \leq p \leq 1$ is a parameter. We use $\mathrm{Bern}(p)$ to denote a Bernoulli distribution with parameter p, and $X \sim \mathrm{Bern}(p)$ means that X has a distribution $\mathrm{Bern}(p)$. A Bernoulli random variable represents a coin toss, which is *fair* when $p = 0.5$ and *biased* otherwise. The expected value and variance are given by $\mathbb{E}X = p$ and $\mathrm{var}(X) = p(1 - p)$, respectively (verify this!).

Joint pmf: Consider a pair of random variables X and Y taking values on $\mathcal{X}$ and $\mathcal{Y}$, respectively. The joint pmf is given by

$$p_{XY}(x,y) \equiv \mathbb{P}[X=x, Y=y], \text{ for all } (x,y) \in \mathcal{X} \times \mathcal{Y}.$$

Note that $\mathcal{X} \times \mathcal{Y} \equiv \{(x,y) : x \in \mathcal{X}, y \in \mathcal{Y}\}$ is the set of all pairs (x,y), known as the *Cartesian product* of $\mathcal{X}$ and $\mathcal{Y}$. The pair (X,Y) can be seen as a vector-valued random variable with a pmf $p_{XY}(x,y)$, which satisfies $\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{XY}(x,y) = 1$ and $p_{XY}(x,y) \geq 0$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$.

Marginal pmfs: From a joint pmf $p_{XY}(x,y)$, we obtain the *marginal* pmfs of X and Y as follows

$$p_X(x) = \sum_{y\in\mathcal{Y}} p_{XY}(x,y) \quad \text{and} \quad p_Y(y) = \sum_{x\in\mathcal{X}} p_{XY}(x,y).$$

Independence: The two random variables X and Y are independent if and only if their joint pmf is equal to the product of their marginal pmfs, i.e.

$$p_{XY}(x,y) = p_X(x)p_Y(y).$$

Conditional pmfs: The conditional pmf of X given $Y=y$ is defined as

$$p_{X|Y}(x|y) \equiv \mathbb{P}[X=x|Y=y] = \frac{p_{XY}(x,y)}{p_Y(y)}.$$

Similarly, the conditional pmf of Y given $X=x$ is given by

$$p_{Y|X}(y|x) \equiv \mathbb{P}[Y=y|X=x] = \frac{p_{XY}(x,y)}{p_X(x)}.$$

Using brief notation, $p_{X|Y}(x|y)$ and $p_{Y|X}(y|x)$ are denoted by $p(x|y)$ and $p(y|x)$, respectively. We use this brief notation when there is no ambiguity. Note that each conditional pmf is a distribution, e.g. each value $y \in \mathcal{Y}$ defines a new (conditional) distribution for X given by $p(x|y)$.

Note that if X and Y are independent, then we have

$$p(x|y) = p(x).$$

In this case, regardless of the value Y takes, the distribution of X remains unchanged. Similarly, we will also have $p(y|x) = p(y)$.

Example 5. Consider a pair of Bernoulli (or binary) random variables X and Y with a joint distribution given in the following table

	$Y=0$	$Y=1$
$X=0$	0.5	0.1
$X=1$	0.2	0.2

Here the marginal distributions are given by

$$p_X(x) = \begin{cases} 0.6, & x=0 \\ 0.4, & x=1 \end{cases} \quad \text{and} \quad p_Y(y) = \begin{cases} 0.7, & y=0 \\ 0.3, & y=1. \end{cases}$$

The conditional distributions of X given $Y=0$ and X given $Y=1$ are given by

$$p_{X|Y}(x|0) = \begin{cases} 5/7, & x=0 \\ 2/7, & x=1 \end{cases} \quad \text{and} \quad p_{X|Y}(x|1) = \begin{cases} 1/3, & x=0 \\ 2/3, & x=1. \end{cases}$$

Work out the conditional distributions of Y given X.

Chain Rule: Let $X_1, X_2, \ldots, X_n$ have a joint pmf of $p(x_1, x_2, \ldots, x_n)$. We have

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots p(x_n|x_1, x_2, \ldots, x_{n-1})$$

which follows from the definition of the conditional pmf. This expansion can be written more briefly as $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^n p(x_i|x_1, \ldots, x_{i-1})$. When $i = 1$ here, we have $p(x_i|x_1, \ldots, x_{i-1}) = p(x_i) = p(x_1)$, i.e. the sequence $x_1, \ldots, x_{i-1}$ is empty in this case.

Independent and Identically Distributed (i.i.d.): A sequence of random variables $X_1, X_2, \ldots, X_n$ is i.i.d. if all random variables are independent and have the same marginal distribution $p_X(x)$, i.e.

$$p(x_1, x_2, \ldots, x_n) = p_X(x_1)p_X(x_2)\cdots p_X(x_n) = \prod_{i=1}^n p_X(x_i).$$

Example 6. (Binomial) Consider an i.i.d. sequence $X_1, X_2, \ldots, X_n$ in which each entry is a Bernoulli random variable with parameter p (i.e. has a Bern(p) distribution). The number of ones in this random sequence is another random variable given by $K = \sum_{i=1}^n X_i$, which takes values on the set $\{0, 1, 2, \ldots, n\}$. K has a binomial distribution with parameters (n, p) and a pmf given by

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

where $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ is the binomial coefficient (i.e. n choose k). The expected value and variance are given by $\mathbb{E}K = np$ and $\text{var}(K) = np(1-p)$, respectively (verify this!).

Weak Law of Large Numbers Consider a random variable X with distribution $p(x)$ and a finite expected value, i.e. $-\infty < \mathbb{E}X < \infty$. Now let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. copies of X, i.e. each has the same distribution of X. The empirical mean of this sequence is defined as

$$\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^n X_i.$$

In case we do not know the true expected value $\mathbb{E}X$, we can estimate it as $\overline{X}_n$, given that we have access to an i.i.d. sample. How good is this estimate? The weak law of large numbers says that the probability that the empirical mean estimate is bad goes to zero as n grows large. This is formally stated as follows.

Theorem 1. (Weak Law of Large Numbers). For every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\Big\{ \big|\overline{X}_n - \mathbb{E}X\big| > \epsilon \Big\} = 0.$$

Proof. This proof is part of the first assignment. In particular, you will be guided through the steps of showing that the following inequality holds:

$$\mathbb{P}\Big\{ \big|\overline{X}_n - \mathbb{E}X\big| > \epsilon \Big\} \leq \frac{\text{var}(X)}{n\epsilon^2}.$$

Theorem 1 follows directly as the right-hand-side of the above inequality converges to 0 for any $\epsilon > 0$. This simple proof of the weak law of large number also requires that $\text{var}(X) < \infty$. □

3 Convexity

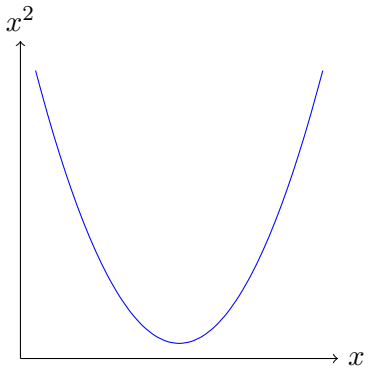
Let's take a small break from probability, and look at convexity. We will combine the two through Jensen's inequality further on. An open interval on the real line is denote by (a, b), for some $a < b$. Similarly, a closed interval is denoted by $[a, b]$. We use $f(x)$ to denote a real-valued function of a real variable x.

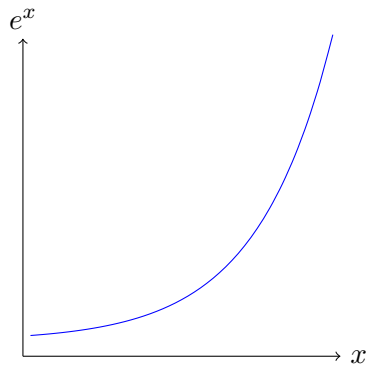
Convex Function: A function $f(x)$ is said to be *convex* over an interval (a, b) if we have

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$. The function $f(x)$ is *strictly convex* if the inequality is strict for all $\lambda \in (0, 1)$, and equality holds only if $\lambda = 0$ or $\lambda = 1$.

Roughly speaking, convex functions curve upwards and often look like a cup (i.e. $\cup$). This is seen from the definition by noting that a convex function always lies below any chord. Examples of convex functions are x^2, e^x and $|x|$ (see below). The opposite of convex functions are concave functions.



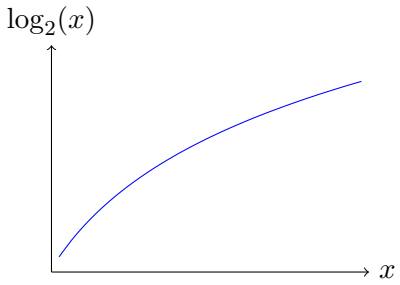


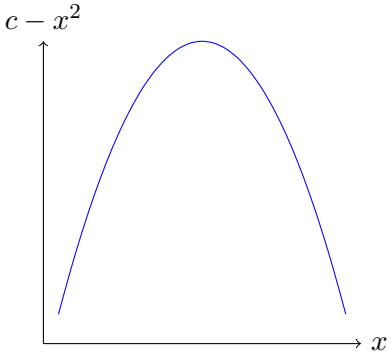
Concave Function: A function $f(x)$ is said to be *concave* over an interval (a, b) if we have

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \geq \lambda f(x_1) + (1-\lambda) f(x_2)$$

for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$. The function $f(x)$ is *strictly concave* if the inequality is strict for all $\lambda \in (0, 1)$, and equality holds only if $\lambda = 0$ or $\lambda = 1$.

From the above definition, it follows that a function $f(x)$ is concave if $-f(x)$ is convex. Hence roughly speaking, concave functions curve downwards and often look like a cap (i.e. $\cap$). A concave function always lies above any chord. Examples of concave functions are $\log(x)$, $\sqrt{x}$ and $c - x^2$ (see below). Note that an affine (or linear) function given by $cx + d$ is both convex and concave (verify this!).





Second Derivative: Now assume that the a function $f(x)$ is twice differentiable, i.e. its second derivative $f''(x)$ exists. If $f''(x) \geq 0$ on an interval (a, b), then the function is convex on that interval. Strictness in the inequality implies strict convexity. That is

$$f''(x) \geq 0 \implies f(x) \text{ is convex.} \quad f''(x) > 0 \implies f(x) \text{ is strictly convex.} \tag{1}$$

The second derivative captures the change in the slope of the function. The slope of a convex function is non-decreasing, or increasing when the function is strictly convex (check this for x^2 and e^x). Note that (1) immediately implies the following statement for concave functions

$$f''(x) \leq 0 \implies f(x) \text{ is concave.} \quad f''(x) < 0 \implies f(x) \text{ is strictly concave.} \tag{2}$$

We now give a proof* (no need to learn this proof) for the statement in (1).

Proof. Suppose that we have a function $f(x)$ with $f''(x) \geq 0$ for all $x \in (a, b)$. From the Taylor theorem, for any $x_0 \in (a, b)$, we can express $f(x)$ as follows:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^\star)}{2}(x - x_0)^2$$

where $x^\star$ is between x_0 and x. Since $x^\star$ is also in (a, b), we have $f''(x^\star) \geq 0$. Therefore, by removing the non-negative term $\frac{f''(x^\star)}{2}(x - x_0)^2$, we may only reduce $f(x)$, that is

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0). \tag{3}$$

Now let's set $x_0 = \lambda x_1 + (1 - \lambda)x_2$ for some $x_1, x_2 \in (a, b)$ and $\lambda \in [0, 1]$. On the other hand, we take x to be either x_1 or x_2. Combing with (3), we obtain the two following inequalities

$$x = x_1 \implies f(x_1) \geq f(x_0) + f'(x_0)(1 - \lambda)(x_1 - x_2) \tag{4}$$
$$x = x_2 \implies f(x_2) \geq f(x_0) + f'(x_0)\lambda(x_2 - x_1). \tag{5}$$

Multiplying (4) by λ and (5) by $(1 - \lambda)$ and adding the resulting inequalities, we obtain

$$\begin{aligned}
\lambda f(x_1) + (1 - \lambda)f(x_2) &\geq \lambda\big[f(x_0) + f'(x_0)(1 - \lambda)(x_1 - x_2)\big] + (1 - \lambda)\big[f(x_0) + f'(x_0)\lambda(x_2 - x_1)\big] \\
&= f(x_0) + f'(x_0)\lambda(1 - \lambda)(x_1 - x_2) + f'(x_0)(1 - \lambda)\lambda(x_2 - x_1) \\
&= f(x_0) \\
&= f(\lambda x_1 + (1 - \lambda)x_2).
\end{aligned}$$

This holds for any $x_1, x_2 \in (a, b)$ and $\lambda \in [0, 1]$, as we can always choose x_0 from (a, b) for the Taylor expansion such that $x_0 = \lambda x_1 + (1 - \lambda)x_2$. Therefore, $f(x)$ is convex, which completes the proof. □

Jensen's Inequality: We now put together probability and convexity.

Theorem 2. (Jensen's Inequality). Let X be a random variable and let f be a convex function. We have

$$\mathbb{E}f(X) \geq f(\mathbb{E}X).$$

If f is strictly convex, then the inequality is strict. Moreover, if f is strictly convex and $\mathbb{E}f(X) = f(\mathbb{E}X)$, then we must have $X = \mathbb{E}X$ (i.e. X is a constant).

Proof. Let's assume that X takes values in $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$ with probabilities $p_1, p_2, \ldots, p_k$. If $k = 2$, the inequality follows directly from the definition of convexity. In particular, we have $\lambda = p_1$ and $1 - \lambda = 1 - p_1 = p_2$, for which we obtain

$$\mathbb{E}f(X) = p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) = f(\mathbb{E}X). \tag{6}$$

Now we wish to show that the statement holds for any k. We do this by induction.

In particular, let's suppose that the statement holds for any $k - 1$ mass points. This means that for a pmf with $k - 1$ mass points given by $p'_1, p'_2, \ldots, p'_{k-1}$, we have

$$\sum_{i=1}^{k-1} p'_i f(x_i) \geq f\left(\sum_{i=1}^{k-1} p'_i x_i\right). \tag{7}$$

Now let's go back to our pmf with k mass points $p_1, p_2, \ldots, p_k$. We split this pmf into $p_1, p_2, \ldots, p_{k-1}$ and p_k. By normalizing $p_1, p_2, \ldots, p_{k-1}$, we obtain a pmf with $k - 1$ mass points as follows:

$$p'_i = \frac{p_i}{\sum_{j=1}^{k-1} p_j} = \frac{p_i}{1 - p_k}, \quad \text{for all } i \in \{1, 2, \ldots, k - 1\}.$$

We now proceed as follows

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + \sum_{i=1}^{k-1} p_i f(x_i) \\
&= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} \frac{p_i}{1-p_k} f(x_i) \\
&= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) \\
&\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i \right) \qquad (8) \\
&\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i \right) \qquad (9) \\
&= f\left(\sum_{i=1}^{k} p_i x_i \right).
\end{aligned}$$

Note that the inequality in (8) is due to our hypothesis that the statement holds for $k-1$ (see (7)). On the other hand, (9) follows from the definition of convexity (see (6)).

Based on the hypothesis (or assumption) that Jensen's inequality holds for $k-1$, we have shown that it must also hold for k. Since we know that it holds for the case with 2 mass points, this means that it must hold for the case with 3 mass points, and so on until we reach any k. This completes the proof. □