Lecture 2 - Random Variables

Jan Bouda

FI MU

March 27, 2012

Part I

Motivation and definition

Random variable - motivation

- In many situation outcomes of a random experiment are numbers.
- In other situations we want to assign to each outcome a number (in addition to probability).
- It may e.g. quantify financial or energetic cost of a particular outcome.
- We will define the random variable to to develop methods for studying random experiments with outcomes that can be described numerically.
- E.g. in case of Bernoulli trials we may be interested only in number of 'successes' and not in actual sequence of 'successes' and 'failures'.
- Almost all real probabilistic computation is done using random variables.

Random variable - definition

A random variable is a rule that assigns a numerical value to each outcome of an experiment.

Definition

A **random variable** $\mathbf{X}$ on a sample space $\mathcal{S}$ is a function $\mathbf{X} : \mathcal{S} \rightarrow \mathbb{R}$ that assigns a real number $\mathbf{X}(s)$ to each sample point $s \in \mathcal{S}$.

We define the **image** of a random variable $\mathbf{X}$ as the set $\mathbf{Im}(\mathbf{X}) = \{\mathbf{X}(s) | s \in \mathcal{S}\}$. This definition is similar to image of any other function.

Random variable - definition

A random variable partitions the sample space into a set of mutually exclusive and collectively exhaustive events. For a random variable $\mathbf{X}$ and a real number x we define the event $A_x =$ "$\mathbf{X} = x$" (sometimes called the **inverse image** of the set $\{x\}$) to be the set of all events from $\mathcal{S}$ to which $\mathbf{X}$ assigns the value x

$$A_x = \{s \in \mathcal{S} | \mathbf{X}(s) = x\}.$$

Whenever you are not sure what some operation with random variable means, always recall the basic definition of the random variable. In example, the statement $\mathbf{X} \leq \mathbf{Y}$ means that $\mathbf{X}$ and $\mathbf{Y}$ are defined on the same sample space $\mathcal{S}$ and for every sample point $s \in \mathcal{S}$ it holds that $\mathbf{X}(s) \leq \mathbf{Y}(s)$.

Random variable - definition

Obviously $A_x \cap A_y = \emptyset$ iff $x \neq y$ and

$$\bigcup_{x \in \mathbb{R}} A_x = S.$$

Therefore the set of events $\{A_x\}_{x \in \mathbb{R}}$ defines an event space and we will often prefer to work in this event space rather than in the original sample space. We usually abbreviate A_x as $[X = x]$.

The image of a discrete random variable (this is the case in this course) is at most countable.

Following the definition of $A_x = [X = x]$ we calculate its probability as

$$P([X = x]) = P(\{s | X(s) = x\}) = \sum_{X(s)=x} P(s).$$

Random variable - probability distribution

Definition

Probability distribution of a random variable X is a function $p_X : \mathbb{R} \to [0,1]$ satisfying the properties:

(p1) $0 \leq p_X(x) \leq 1$ for all $x \in \mathbb{R}$

(p2) For a discrete random variable X, the set $\{x | p_X(x) > 0\}$ is a finite or countable infinite subset of real numbers. Let us denote it by $\{x_1, x_2, \dots\}$. We require that

$$\sum_i p_X(x_i) = 1.$$

A real valued function $p_X(x)$ defined on $\mathbb{R}$ is a probability distribution of some random variable if it satisfies properties (p1) and (p2).

Random variable - probability distribution

When the random variable is clear from the context, we denote the probability distribution as $p(x)$.
Do not mistake the probability distribution with the **distribution function**, which is a non-decreasing function which tends to 0 as $x \to -\infty$ and to 1 as $x \to \infty$.

Distribution functions

We often are interested in computing the probability of the set $\{s|X(s) \in A\}$ for some subset $A \subseteq \mathbb{R}$. We know that

$$\{s|X(s) \in A\} = \bigcup_{x_i \in A} \{s|X(s) = x_i\} \overset{def}{=} [X \in A].$$

If $-\infty < a < b < \infty$ and A is an interval $A = (a, b)$, we usually write $P(a < X < b)$ instead of $P(X \in (a, b))$. If $A = (a, b]$, then $P(X \in A)$ will be written as $P(a < X \leq b)$. Of special interest is the infinite interval $A = (-\infty, x]$ and we denote it by $[X \leq x]$. We calculate the probability of A as

$$P(X \in A) = \sum_{x_i \in A} p_X(x_i).$$

Probability distribution function

Definition

The **probability distribution function** (or simply **distribution function**) of a random variable X is

$$F_X(t) = P(-\infty < X \leq t) = P(X \leq t) = \sum_{x \leq t} p_X(x), \ -\infty < t < \infty.$$

It follows that

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

If X is an integer-valued random variable, then

$$F(t) = \sum_{-\infty < x \leq \lfloor t \rfloor} p_X(x).$$

Probability distribution function - properties

(F1) $0 \leq F(x) \leq 1$ for $-\infty < x < \infty$

(F2) $F(x)$ is a monotone increasing function of x, that is if $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$. It is easy to see that $(-\infty, x_1] \subseteq (-\infty, x_2]$ if $x_1 \leq x_2$ and we have

$$P(-\infty < X \leq x_1) \leq P(-\infty < X \leq x_2)$$

giving $F(x_1) \leq F(x_2)$.

(F3) $\lim_{x\to-\infty} F(x) = 0$, and $\lim_{x\to\infty} F(x) = 1$. If the random variable X has a finite image, then there exist $u, v \in \mathbb{R}$ such that $F(x) = 0$ for all $x < u$ and $F(x) = 1$ for all $x \geq v$.

(F4) $F(x)$ has a positive increase equal to $p_X(x_i)$ at $i = 1, 2, \ldots$ and in the interval $[x_i, x_{i+1})$ it has a constant value. Thus

$$F(x) = F(x_i) \text{ for } x_i \leq x < x_{i+1}$$

and

$$F(x_{i+1}) = F(x_i) + p_X(x_{i+1}).$$

Probability distribution function

- Any function satisfying properties (F1)-(F4) is the distribution function of some discrete random variable.
- In most cases we simply forget the theoretical background (random experiment, sample space, events,...) and examine random variables, probability distributions and probability distribution functions.
- Often the initial information is *we have a random variable X with the probability distribution $p_X(x)$*. We can construct probability space consistent with the random variable as follows. Let $S = \mathbb{R}$, $X(s) = s$ for $s \in S$, $\mathcal{F}$ is a union of inverse images A_x of all subsets $\{x\}$ and

$$P(A) = \sum_{x \in A} p_X(x).$$

Part II

Examples of probability distributions

Examples of probability distributions

In this part of the lecture we introduce the most common probability distributions occurring in practical situations. In fact, we can always derive the distributions and all related results ourselves, however, it is anyway useful to remember these distributions and situations they describe both as examples and to speed up our calculations. These probability distributions are so important that they have specific names and sometimes also notation.

Constant random variable

- For $c \in \mathbb{R}$ the function defined for all $s \in S$ by $X(s) = c$ is a discrete random variable with $P(X = c) = 1$.
- The probability distribution of this variable is

$$p_X(x) = \begin{cases} 1 & \text{if } x = c \\ 0 & \text{otherwise.} \end{cases}$$

- Such a random variable is called the **constant random variable**.
- The corresponding distribution function is

$$F_X(x) = \begin{cases} 0 & \text{for } x < c \\ 1 & \text{for } x \geq c. \end{cases}$$

Discrete uniform probability distribution

- Let X be a discrete random variable with a finite image $\{x_1, x_2 \ldots x_n\}$ and let us assign to all elements of the image the same probability $p_X(x_i) = p$.
- From the requirement that the probabilities must sum to 1 we have

$$1 = \sum_{i=1}^{n} p_X(x_i) = \sum_{i=1}^{n} p = np$$

  and the probability is

$$p_X(x_i) = \begin{cases} 1/n & x_i \in \mathbf{Im}(X) \\ 0 & \text{otherwise.} \end{cases}$$

- Such a random variable is said to have the **uniform probability distribution**.
- This concept cannot be extended to random variable with countably infinite image.

Discrete uniform probability distribution

- If $\mathbf{Im}(X) = \{1, 2, \dots n\}$ with $p_X(i) = 1/n$, $1 \le i \le n$, the probability distribution function is

$$F_X(x) = \sum_{i=1}^{\lfloor x \rfloor} p_X(i) = \frac{\lfloor x \rfloor}{n}, \quad 1 \le x \le n.$$

Bernoulli probability distribution

- The Bernoulli probability distribution of a random variable X origins from the random experiment consisting of a single bernoulli trial (e.g. a coin toss).
- The only possible values of the random variable X are 0 and 1 (often denoted as *failure* and *success*, respectively).
- The distribution is given by

$$\begin{aligned} p_X(0) =& p_0 = P(X=0) = q \\ p_X(1) =& p_1 = P(X=1) = p = 1-q \end{aligned}$$

- The corresponding probability distribution function is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ q & \text{for } 0 \leq x < 1 \\ 1 & \text{for } x \geq 1. \end{cases}$$

Bernoulli probability distribution

Example

Let X be a Bernoulli random variable with parameter p and image $\{0,1\}$. X is the indicator of the event

$$A = \{s | X(s) = 1\}$$

and its probability distribution is $p_X(0) = 1 - p$ and $P_X(1) = p$.

Binomial probability distribution

- The Binomial probability distribution of a random variable Y_n is the number of successes (outcomes 1) in n consecutive Bernoulli trial with the same fixed probability p of success in each trial.
- The domain of the random variable Y_n are all n–tuples of 0s and 1s. The image is $\{0, 1, 2, \dots n\}$.
- As already demonstrated in the previous lecture, the probability distribution of Y_n is

$$\begin{aligned} p_k =& P(Y_n = k) = p_{Y_n}(k) \\ =& \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{for } 0 \leq k \leq n \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

- The binomial distribution is often denoted as $b(k; n, p) = p_k$ and represents the probability that there are k successes in a sequence of n bernoulli trials with probability of success p.
- In example, $b(3; 5, 0.5) = \binom{5}{3}(1/2)^3(1/2)^2 = 0.3125$

Binomial probability distribution

After specifying the distribution of a random variable we should verify that this function is a valid probability distribution, i.e. to verify properties (p1) and (p2). While (p1) is usually clear (it is easy to see that the function is nonnegative), the property (p2) may be not so straightforward and should be verified explicitly.

We can apply the binomial model when the following conditions hold:

- Each trial has exactly two mutually exclusive outcomes.
- The probability of 'success' is constant on each trial.
- The outcomes of successive trials are mutually independent.

Binomial probability distribution

- The name 'binomial' comes from the equation verifying that the probabilities sum to 1

$$\sum_{i=0}^{n} p_i = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$
$$= [p + (1-p)]^n = 1.$$

- The corresponding distribution function, denoted by $B(t; n, p)$ is given by

$$B(t; n, p) = F_{Y_n}(t) = \sum_{i=0}^{\lfloor t \rfloor} \binom{n}{i} p^i (1-p)^{n-i}.$$

Part III

Discrete random vectors

Discrete random vectors

- Suppose we want to study relationship between two or more random variables defined on a given sample space.
- Let $X_1, X_2, \ldots X_r$ be r discrete random variables defined on a sample space S.
- For each sample point $s \in S$, each of the random variables $X_1, X_2, \ldots X_r$ takes on one of its possible values

$$X_1(s) = x_1, X_2(s) = x_2, \ldots X_r(s) = x_r.$$

- The random vector $\mathbf{X} = (X_1, X_2, \ldots X_r)$ is an r-dimensional vector-valued function $\mathbf{X} : S \to \mathbb{R}^r$ with $\mathbf{X}(s) = \mathbf{x} = (x_1, x_2, \ldots x_r)$.

Discrete random vectors

Definition

The **joint** (or **compound**) **probability distribution** of a random vector $\mathbf{X}$ is defined to be

$$p_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \dots X_r = x_r).$$

The properties of random vectors are

(j1) $p_{\mathbf{X}}(\mathbf{x}) \geq 0,\ \mathbf{x} \in \mathbb{R}^r$.

(j2) $\{\mathbf{x} | p_{\mathbf{x}}(\mathbf{x}) \neq 0\}$ is a finite or countably infinite subset of $\mathbb{R}^r$, which will be denoted as $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$.

(j3) $P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} p_{\mathbf{X}}(\mathbf{x})$.

(j4) $\sum_i p_{\mathbf{X}}(\mathbf{x}_i) = 1$.

Marginal probability distributions

- In situation when we are examining more that one random variable, the probability distribution of a single variable, e.g. $p_X(x)$, is referred to as **marginal probability distribution** (in contrast to joint probability distribution).
- Considering joint probability distribution $p_{X,Y}(x,y)$ of random variables X and Y we can calculate the marginal probability distribution of X as

$$\begin{aligned} p_X(x) =& P(X=x) = P\left(\bigcup_j \{X=x, Y=y_j\}\right) \\ &= \sum_j P(X=x, Y=y_j) = \sum_j p_{X,Y}(x,y_j). \end{aligned}$$

Similarly we obtain the marginal probability distribution of Y

$$p_Y(y) = \sum_i p_{X,Y}(x_i,y).$$

Marginal probability distributions

- While it is relatively easy to calculate the marginal probability distributions from the joint distribution, in general there is no way how to determine the joint distribution from corresponding marginal distributions.
- The only exception are independent random variables (see below), when the joint probability distribution is the product of marginal distributions.

Multinomial probability distribution

- Interesting example of joint probability distribution is the multinomial distribution.
- Consider a sequence of n generalized bernoulli trials, where each of them has a finite number r of outcomes having probabilities $p_1, p_2, \ldots, p_r$.
- Let us define the random vector $\mathbf{X} = (X_1, X_2, \ldots X_r)$ such that X_i is the number of trials that resulted in ith outcome.
- Then the compound probability distribution of $\mathbf{X}$ is

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{n}) =& P(X_1 = n_1, X_2 = n_2, \ldots X_r = n_r) \\ =& \binom{n}{n_1, n_2, \ldots n_r} p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r}, \end{aligned}$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_r)$ and $\sum_{i=1}^{r} n_i = n$.

Multinomial probability distribution

The marginal probability distribution of X_i may be computed by

$$
\begin{aligned}
p_{X_i}(n_i) = & \sum_{\mathbf{n}:\left[\left(\sum_{j\neq i} n_j\right)=n-n_i\right]} \binom{n}{n_1, n_2 \dots n_r} p_1^{n_1} p_2^{n_2} \dots p_r^{n_r} \\
= & \frac{n! p_i^{n_i}}{(n-n_i)! n_i!} \sum_{\mathbf{n}:\left[\left(\sum_{j\neq i} n_j\right)=n-n_i\right]} \frac{(n-n_i)! p_1^{n_1} \dots p_{i-1}^{n_{i-1}} p_{i+1}^{n_{i+1}} \dots p_r^{n_r}}{n_1! n_2! \dots n_{i-1}! n_{i+1}! \dots n_r!} \\
= & \binom{n}{n_i} p_i^{n_i} (p_1 + \cdots + p_{i-1} + p_{i+1} + \cdots + p_r)^{n-n_i} \\
= & \binom{n}{n_i} p_i^{n_i} (1-p_i)^{n-n_i}.
\end{aligned}
$$

Part IV

Independent random variables

Independent random variables

Definition

Two discrete random variables are **independent** provided their joint probability distribution is a product of the marginal probability distributions, i.e.

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \text{ for all } x \text{ and } y.$$

- If X and Y are two independent random variables, then for any two subsets $A, B \subseteq \mathbb{R}$ the events $X \in A$ and $Y \in B$ are independent:

$$P(X \in A \cap Y \in B) = P(X \in A)P(Y \in B)$$

To see this

$$\begin{aligned} P(X \in A \cap Y \in B) &= \sum_{x\in A}\sum_{y \in B} p_{X,Y}(x,y) \\ &= \sum_{x\in A}\sum_{y \in B} p_X(x)p_Y(y) \end{aligned}$$

Independent random variables

Definition

Let $X_1, X_2, \dots X_r$ be discrete random variables with probability distributions $p_{X_1}, p_{X_2}, \dots p_{X_r}$. These random variables are **pairwise independent** if

$$\forall 1 \leq i < j \leq r,\ \forall x_i \in \mathbf{Im}(X_i), x_j \in \mathbf{Im}(X_j),\ p_{X_i,X_j}(x_i, x_j) = p_{X_i}(x_i)p_{X_j}(x_j).$$

Definition

Let $X_1, X_2, \dots X_r$ be discrete random variables with probability distributions $p_{X_1}, p_{X_2}, \dots p_{X_r}$. These random variables are **mutually independent** if for all $x_1 \in \mathbf{Im}(X_1), x_2 \in \mathbf{Im}(X_2), \dots, x_r \in \mathbf{Im}(X_r)$

$$p_{X_1,X_2,\dots X_r}(x_1, x_2, \dots x_r) = p_{X_1}(x_1)p_{X_2}(x_2)\dots p_{X_r}(x_r).$$

Note that pairwise independence of a set of random variables does not imply their mutual independence.

Independent random variables

- Let X and Y be non-negative independent random variables. Then the probability distribution of the random variable $Z = X + Y$ is

$$p_Z(t) = p_{X+Y}(t) = \sum_{x=0}^{t} p_X(x) p_Y(t-x).$$

In case X and Y can also take negative values, the sum should go from $-\infty$ instead of 0.

Independent random variables

Theorem

Let $X_1, X_2, \ldots X_r$ be mutually independent. If X_i has the binomial distribution with parameters n_i and p, then $\sum_{i=1}^{r} X_i$ has the binomial distribution with parameters $n_1 + n_2 + \cdots + n_r$ and p.

Part V

Functions of a random variable

Functions of a random variable

Given a random variable $\mathbf{X}$ and a function $\Phi : \mathbb{R} \to \mathbb{R}$ we define the transformed random variable $\mathbf{Y} = \Phi(\mathbf{X})$ as

- Random variables $\mathbf{X}$ and $\mathbf{Y}$ are defined on the same sample space, moreover, $\mathbf{Dom}(\mathbf{X}) = \mathbf{Dom}(\mathbf{Y})$.
- $\mathbf{Im}(\mathbf{Y}) = \{\Phi(x) | x \in \mathbf{Im}(\mathbf{X})\}$.
- The probability distribution of $\mathbf{Y}$ is given by

$$p_{\mathbf{Y}}(y) = \sum_{x \in \mathbf{Im}(\mathbf{X}); \Phi(x) = y} p_{\mathbf{X}}(x).$$

In fact, we may define it by $\Phi(\mathbf{X}) = \Phi \circ \mathbf{X}$, where $\circ$ is the usual function composition.

Part VI

Expectation

Expectation

- The probability distribution or probability distribution function completely characterize properties of a random variable.
- Often we need description that is less accurate, but much shorter - single number, or a few numbers.
- First such characteristic describing a random variable is the **expectation**, also known as the **mean value**.

Definition

Expectation of a random variable X is defined as

$$E(X) = \sum_i x_i p(x_i)$$

provided the sum is absolutely (!) convergent. In case the sum is convergent, but not absolutely convergent, we say that no finite expectation exists. In case the sum is not convergent the expectation has no meaning.

Median; Mode

- The **median** of a random variable X is any number x such that $P(X < x) \leq 1/2$ and $P(X > x) \geq 1/2$.
- The **mode** of a random variable X is the number x such that

$$p(x) = \max_{x' \in \mathbf{Im}(X)} p(x').$$

Expectation of a functional transformation

Theorem

Let $X_1, X_2, \ldots X_n$ be random variables defined on the same probability space and let $Y = \Phi(X_1, X_2, \ldots X_n)$. Then

$$E(Y) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} \Phi(x_1, x_2, \ldots, x_n) p(x_1, x_2, \ldots, x_n).$$

Theorem (Linearity of expectation)

Let X and Y be random variables. Then

$$E(X + Y) = E(X) + E(Y).$$

Linearity of expectation (proof)

Linearity of expectation.

$$
\begin{aligned}
E(X+Y) &= \sum_i \sum_j (x_i + y_j) p(x_i, y_j) = \\
&= \sum_i x_i \sum_j p(x_i, y_j) + \sum_j y_j \sum_i p(x_i, y_j) = \\
&= \sum_i x_i p_X(x_i) + \sum_j y_j p_Y(y_j) = \\
&= E(X) + E(Y).
\end{aligned}
$$

□

Linearity of expectation

The linearity of expectation can be easily generalized for any linear combination of n random variables, i.e.

Theorem (Linearity of expectation)

Let $X_1, X_2, \ldots X_n$ be random variables and $a_1, a_2, \ldots a_n \in \mathbb{R}$ constants. Then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i).$$

Proof is left as a home exercise :-).

Expectation of independent random variables

Theorem

If X and Y are independent random variables, then

$$E(XY) = E(X)E(Y).$$

Proof.

$$\begin{aligned}
E(XY) =& \sum_i \sum_j x_i y_j p(x_i, y_j) = \\
=& \sum_i \sum_j x_i y_j p_X(x_i) p_Y(y_j) = \\
=& \sum_i x_i p_X(x_i) \sum_j y_j p_Y(y_j) = \\
=& E(X)E(Y).
\end{aligned}$$

□

Expectation of independent random variables

The expectation of independent random variables can be easily generalized for any n–tuple $X_1, X_2, \ldots X_n$ of mutually independent random variables:

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i).$$

If $\Phi_1, \Phi_2, \ldots \Phi_n$ are functions, then

$$E\left[\prod_{i=1}^{n} \Phi_i(X_i)\right] = \prod_{i=1}^{n} E[\Phi_i(X_i)].$$

Part VII

Jensen's inequality

Convex and concave functions

Before introducing Jensen's inequality, let us briefly refresh definitions of convex and concave function, which are crucial in this part.

Definition

A function $f(x)$ is said to be **convex** on a set $\mathbf{S}$ if for every $x_1, x_2 \in \mathbf{S}$ and $0 \leq \lambda \leq 1$

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

A function is **strictly convex** if the equality holds only if $\lambda = 0$ or $\lambda = 1$. A function f is **concave** if $-f$ is convex. A function f is **strictly concave** if $-f$ is strictly convex.

Theorem

If the function has a second derivative which is nonnegative (positive) everywhere, then the function is convex (strictly convex).

Convex and Concave Functions

Proof.

We use the Taylor series expansion of the function around x_0

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

where x^* lies between x_0 and x. By our initial assumption the term $f''(x^*)$ is always nonnegative and the same holds for the last addend. Let $x_0 = \lambda x_1 + (1 - \lambda)x_2$, $\lambda \in [0, 1]$ and $x = x_1$ and we have

$$f(x_1) \geq f(x_0) + f'(x_0)[(1 - \lambda)(x_1 - x_2)]. \tag{1}$$

Similarly, taking $x = x_2$ we obtain

$$f(x_2) \geq f(x_0) + f'(x_0)[\lambda(x_2 - x_1)]. \tag{2}$$

Multiplying (1) by λ and (2) by $(1 - \lambda)$ and adding we obtain the convexity. The proof for the strict convexity is analogous.

Convex and Concave Functions

Proof.

Multiplying (1) by λ and (2) by $(1-\lambda)$ and adding we obtain the convexity

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

The proof for the strict convexity is analogous. □

Jensen's Inequality

Last theorem shows immediately the strict convexity for x^2, e^x and $x \log x$ for $x \geq 0$, and the strict concavity of $\log x$ and $\sqrt{x}$ for $x \geq 0$.
The following inequality is behind most of the fundamental theorems in information theory and in mathematics in general.

Theorem (Jensen's inequality)

If f is a convex function and X is a random variable, then

$$E[f(X)] \geq f(E(X)). \tag{3}$$

Moreover, if f is strictly convex, the equality in (3) implies that $X = E(X)$ occurs with probability 1, i.e. X is a constant.

Jensen's Inequality

Proof.

We prove this inequality by induction on the number of elements in $\mathbf{Im}(X)$. For probability distribution on two points we have

$$E(f(X)) = p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) = f(E(X)) \tag{4}$$

what follows directly from convexity. Suppose the theorem holds for $k-1$ points. Then we put $p_i' = p_i/(1-p_k)$ for $i = 1, 2 \ldots, k-1$ and we have

$$\begin{aligned} E(f(X)) = \sum_{i=1}^{k} p_i f(x_i) =& p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) \geq \\ \geq& p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) \geq \end{aligned} \tag{5}$$

Jensen's Inequality

Proof.

$$
\begin{aligned}
&\geq f\left(p_k x_k + (1-p_k)\sum_{i=1}^{k-1} p'_i x_i\right) = \\
&= f\left(\sum_{i=1}^{k} p_i x_i\right) = f(E(X)),
\end{aligned}
$$

where the first inequality follows from the induction hypothesis and the second one from convexity of f. □