Information-Theoretic Modeling

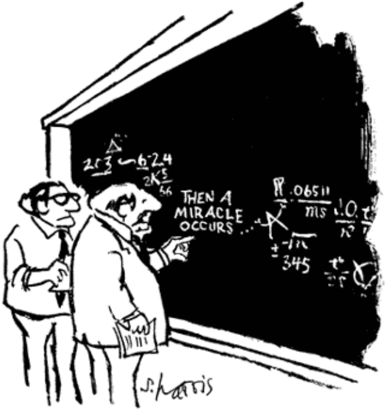
Lecture 2: Mathematical Preliminaries

Teemu Roos

Department of Computer Science, University of Helsinki

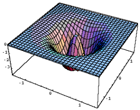
Fall 2014

Lecture 2: Mathematical Preliminaries

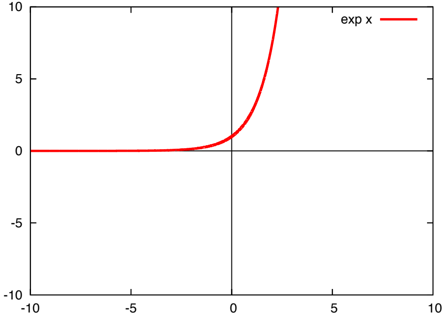


"I think you should be more explicit here in step two."

1. Calculus
   - Limits and Convergence
   - Convexity
2. Probability
   - Probability Space and Random Variables
   - Joint and Conditional Distributions
   - Expectation
   - Law of Large Numbers
3. Inequalities
   - Jensen's Inequality
   - Gibbs's Inequality



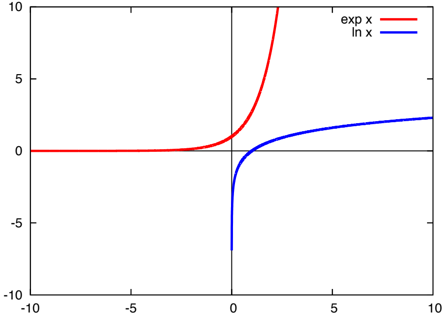
Exponent Function



Exponent function $\exp : \mathbb{R} \to \mathbb{R}^+$, $\exp k = e^k = \overbrace{e \times e \times \ldots \times e}^{k}$: multiplicative growth (nuclear reaction, "interest on interest", ...)

$\exp x \cdot \exp y = \exp(x + y)$ Derivative $\frac{d \exp x}{dx} = \exp x$.

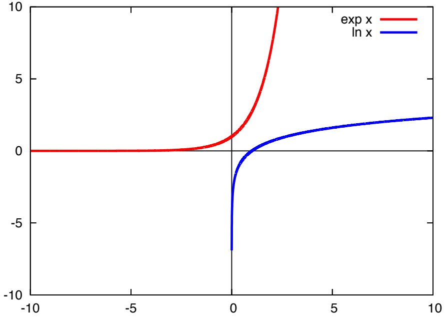
Examples: Logarithm



Natural logarithm $\ln\ :\ \mathbb{R}^+ \to \mathbb{R}$, $\ln \exp x = x$:
time to grow to x, number of digits ($\log_{10}$).

General (base a) logarithm, $\log_a a^x = x$: $\qquad \log_a x = \dfrac{1}{\ln a} \ln x$

Logarithm Function



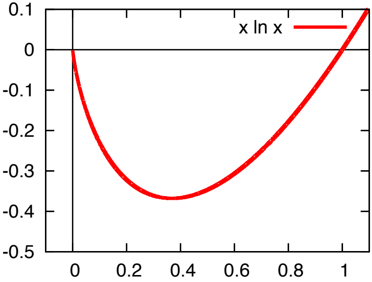
$$\ln xy = \ln x + \ln y \quad \ln x^r = r \ln x \quad \ln \frac{1}{x} = -\ln x \quad \ln \frac{x}{y} = \ln x - \ln y$$

$$\frac{d \ln x}{dx} = \frac{1}{x}$$

Limits and Convergence

- A sequence of values $(x_i \ : \ i \in \mathbb{N})$ *converges* to *limit* L, $\lim_{i\to\infty} x_i = L$, iff for any $\epsilon > 0$ there exists a number $N \in \mathbb{N}$ such that
$$|x_i - L| < \epsilon \quad \text{for all } i \geq N \ .$$
- $f(x)$ has a *limit* L as x approaches c, $\lim_{x\to c} f(x) = L$, (from above c^+/below c^-) iff for any $\epsilon > 0$ there exists a number $\delta > 0$ such that
$$|f(x) - L| < \epsilon \quad \text{for all} \begin{cases} c < x < c + \delta & \text{‘above’} \\ c - \delta < x < c & \text{’below’} \\ 0 < |x - c| < \delta & \text{—} \end{cases}$$

Example: Logarithm Again



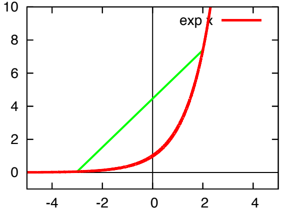
Even though $x \ln x$ is undefined at $x = 0$, we have (by l'Hôpital's rule):

$$\lim_{x \to 0^+} x \ln x = 0 \quad .$$

Convexity

Function $f : \mathcal{X} \to \mathbb{R}$ is said to be **convex** iff for any $x, y \in \mathcal{X}$ and any $0 \leq t \leq 1$ we have

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \ .$$



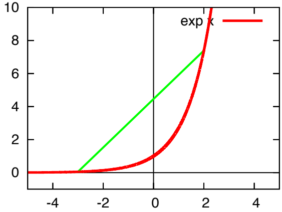
Function f is **strictly convex** iff the above inequality holds strictly ('$<$' instead of '$\leq$') when $0 < t < 1$.

Function f is (strictly) **concave** iff the above holds for $-f$.

Convexity and Derivatives

Theorem

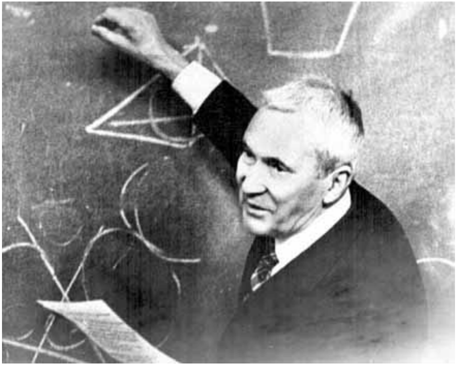
If function f has a second derivative f'', and f'' is non-negative (≥ 0) for all x, then f is convex. If f'' is positive (> 0) for all x, then f is *strictly* convex.



e^x is conve^x!

Example: $f'(x) = \dfrac{d \exp x}{dx} = \exp x \;\Rightarrow f''(x) = \exp x > 0$. Hence exp is strictly convex.

Probability



A.N. Kolmogorov, 1903–1987

1. Calculus
   - Limits and Convergence
   - Convexity
2. Probability
   - Probability Space and Random Variables
   - Joint and Conditional Distributions
   - Expectation
   - Law of Large Numbers
3. Inequalities
   - Jensen's Inequality
   - Gibbs's Inequality

Probability Space

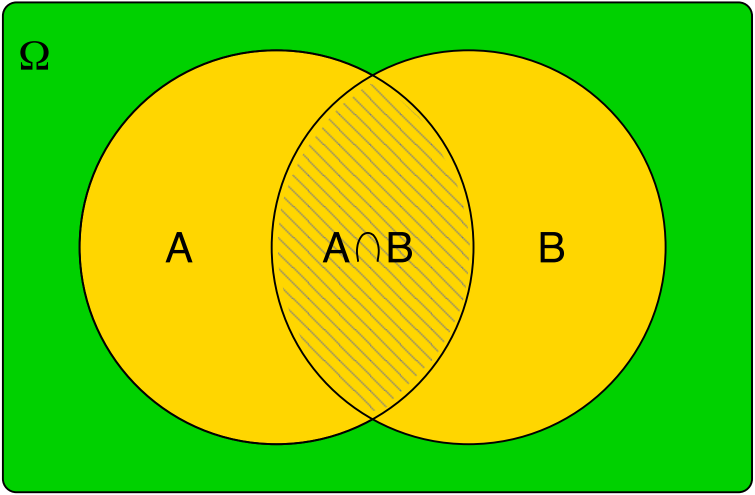
A probability space $(\Omega, \mathcal{F}, P)$ is defined by

- the **sample space** Ω whose elements are called outcomes ω,
- a sigma algebra $\mathcal{F}$ of subsets of Ω, whose elements are called **events** E, and
- a measure P which determines the **probabilities of events**, $P : \mathcal{F} \to [0, 1]$.

Measure P has to satisfy the **probability axioms**: $P(E) \geq 0$ for all $E \in \mathcal{F}$, $P(\Omega) = 1$, and $P(E_1 \cup E_2 \cup \ldots) = \sum_i P(E_i)$ if (E_i) is a countable sequence of *disjoint* events.

These axioms imply the usual rules of **probability calculus**, e.g., $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, $P(\Omega \backslash E) = 1 - P(E)$, etc.

Venn Diagrams



Probability Calculus

1. The **conditional probability** of event B given that event A occurs is defined as
$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad \text{for } A \text{ such that } P(A) > 0.$$
2. $P(A \cap B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$.
3. Bayes' rule: $P(B \mid A) = \dfrac{P(A \mid B) \cdot P(B)}{P(A)}$.
4. Chain rule:
$$\begin{aligned} P(\cap_{i=1}^{N} E_i) &= \prod_{i=1}^{N} P(E_i \mid \cap_{j=1}^{i-1} E_j) \\ &= P(E_1) \cdot P(E_2 \mid E_1) \cdot P(E_3 \mid E_1 \cap E_2) \cdot \ldots \\ &\qquad \cdot P(E_N \mid E_1 \cap \ldots \cap E_{N-1}) \ . \end{aligned}$$

Random Variables

Technically, a random variable is a (measurable) function $X : \Omega \to \mathbb{R}$ from the sample space to the reals.

The probability measure P on Ω determines the distribution of X:

$$P_X(A) = \Pr[X \in A] = P(\{\omega : X(\omega) \in A\}) \ ,$$

where $A \subseteq \mathbb{R}$.

It is often more natural to relabel the outcomes and denote them, for instance, by letters, A, B, C, ..., or words `red`, `black`, ...

In practice, we often forget about the underlying probability space Ω, and just speak of random variable X and its distribution P_X.

Random Variables

The distribution of a random variable can *always* be represented as a *cumulative distribution function* (cdf) $F_X(x) = \Pr[X \le x]$.

In addition:

- A **discrete** random variable X with countable alphabet $\mathcal{X}$ has a *probability mass function* (pmf) p_X such that $\Pr[X = x] = p_X(x)$.
- A **continuous** random variable Y has a *probability density function* (pdf) f_Y such that $\Pr[Y \in A] = \int_A f_Y(x)\, dy$.

There are also *mixed* random variables that are neither discrete nor continuous. They don't have a pmf or pdf, but they do have a cdf.

We often omit the subscripts $X, Y, \ldots$ and write $p(x), f(y)$, etc.

Random Variables

Since random variables are functions, we can define more random variables as functions of random variables: if f is a function, and X and Y are r.v.'s, then $f(X) : \Omega \to \mathbb{R}$ is a r.v., $X + Y$ is a r.v., etc.

Example: Let r.v. X be the outcome of a die.

- The pmf of X is given by $p_X(x) = 1/6$ for all $x \in \{1, 2, 3, 4, 5, 6\}$.
- The pmf of r.v. X^2 is given by $p_{X^2}(x) = 1/6$ for all $x \in \{1, 4, 9, 16, 25, 36\}$.

! In particular, a pmf p_X is a function, and hence, $p_X(X)$ is also a random variable. Further, $p_X^2(X), \ln p_X(X)$, etc. are random variables.

Multivariate Distributions

The probabilistic behavior of two or more random variables is described by multivariate distributions.

The **joint distribution** of r.v.'s X and Y is

$$\begin{aligned} P_{X,Y}(A,B) &= \Pr[X \in A \ \wedge \ Y \in B] \\ &= P(\{\omega \ : \ X(\omega) \in A, Y(\omega) \in B\}) \ . \end{aligned}$$

For each multivariate distribution $P_{X,Y}$, there are unique **marginal distributions** P_X and P_Y such that

$$P_X(A) = P_{X,Y}(A,\mathbb{R}), \qquad P_Y(B) = P_{X,Y}(\mathbb{R},B) \ ,$$

$$\text{pmf: } p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x,y) \quad \text{pdf: } f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,dx \ .$$

Multivariate Distributions

The **conditional distribution** is defined similar to *conditional probability*:

$$P_{Y|X}(B \mid A) = \frac{P_{X,Y}(A,B)}{P_X(A)} \quad \text{for } A \text{ such that } P_X(A) > 0.$$

For discrete/continuous variables we have:

- *discrete* r.v.'s:

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \ , \quad p_X(x) > 0 \ ,$$

- *continuous* r.v.'s:

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \ , \quad f_X(x) > 0 \ .$$

Independence

Variable X is said to be **independent** of variable Y ($X \perp\!\!\!\perp Y$) iff

$$P_{X,Y}(A, B) = P_X(A) \cdot P_Y(B) \quad \text{for all } A, B \subseteq \mathbb{R}.$$

This is equivalent to

$$P_{X|Y}(A \mid B) = P_X(A) \quad \text{for all } B \text{ such that } P(B) > 0,$$

and

$$P_{Y|X}(B \mid A) = P_Y(B) \quad \text{for all } A \text{ such that } P(A) > 0.$$

In words, knowledge about one variable tells nothing about the other. Note that independence is symmetric, $X \perp\!\!\!\perp Y \Leftrightarrow Y \perp\!\!\!\perp X$.

Expectation

The **expectation** (or expected value, or mean) of a discrete random variable is given by

$$E[X] = \sum_{x \in \mathcal{X}} p(x)\, x \ .$$

The expectation of a continuous random variable is given by

$$E[X] = \int_{\mathcal{X}} f(x)\, x \, dx \ .$$

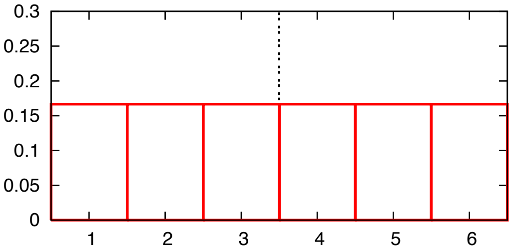
In both cases, it is possible that $E[X] = \pm\infty$.

$$E[kX] = kE[X] \qquad E[X+Y] = E[X] + E[Y]$$

$$E[XY] = E[X]E[Y] \quad \text{if } X \perp\!\!\!\perp Y$$

Law of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of independent outcomes of a die, so that $p_{X_i}(x) = 1/6$ for all $i \in \mathbb{N}, x \in \{1, 2, 3, 4, 5, 6\}$.



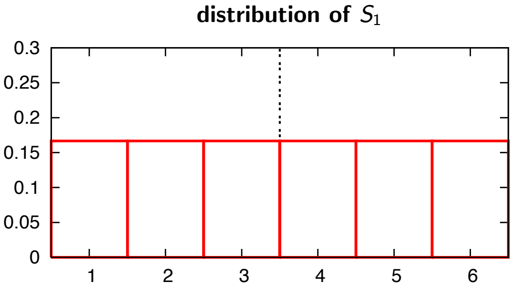
$$E[X_i] = \sum_{x=1}^{6} \frac{1}{6} x = \frac{21}{6} = 3.5 \quad \text{for all } i \in \mathbb{N}.$$

Law of Large Numbers

Let $S_n = \sum_{i=1}^{n} X_n$ be the sum of the first n outcomes.

The distribution of S_n is given by

$$P_{S_n}(x) = \frac{\#\text{ of ways to get sum } x \text{ with } n \text{ dice}}{6^n}$$

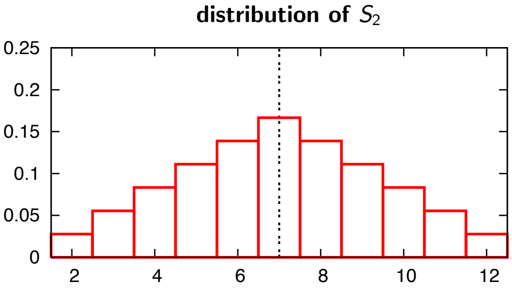


Law of Large Numbers

Let $S_n = \sum_{i=1}^{n} X_n$ be the sum of the first n outcomes.

The distribution of S_n is given by

$$P_{S_n}(x) = \frac{\#\text{ of ways to get sum } x \text{ with } n \text{ dice}}{6^n}$$

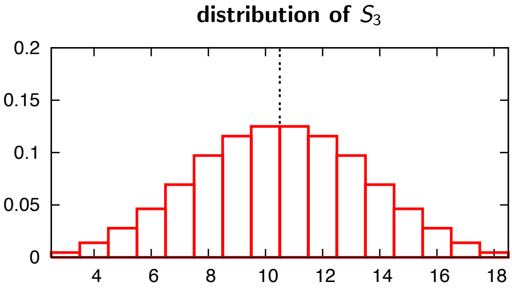


Law of Large Numbers

Let $S_n = \sum_{i=1}^{n} X_n$ be the sum of the first n outcomes.

The distribution of S_n is given by

$$P_{S_n}(x) = \frac{\#\text{ of ways to get sum } x \text{ with } n \text{ dice}}{6^n}$$

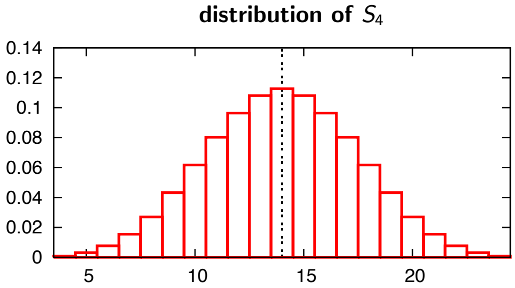


Law of Large Numbers

Let $S_n = \sum_{i=1}^{n} X_n$ be the sum of the first n outcomes.

The distribution of S_n is given by

$$P_{S_n}(x) = \frac{\#\text{ of ways to get sum } x \text{ with } n \text{ dice}}{6^n}$$

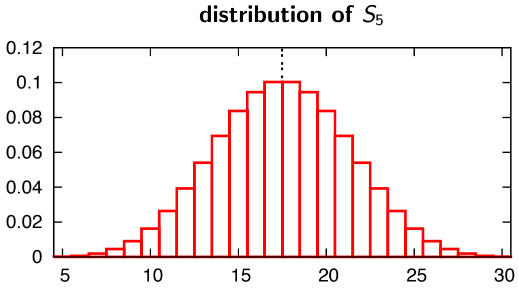


Law of Large Numbers

Let $S_n = \sum_{i=1}^{n} X_n$ be the sum of the first n outcomes.

The distribution of S_n is given by

$$P_{S_n}(x) = \frac{\#\text{ of ways to get sum } x \text{ with } n \text{ dice}}{6^n}$$

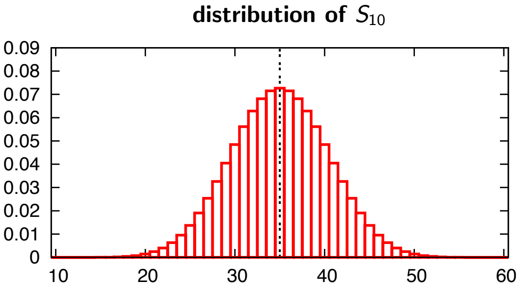


Law of Large Numbers

Let $S_n = \sum_{i=1}^{n} X_n$ be the sum of the first n outcomes.

The distribution of S_n is given by

$$P_{S_n}(x) = \frac{\#\text{ of ways to get sum } x \text{ with } n \text{ dice}}{6^n}$$

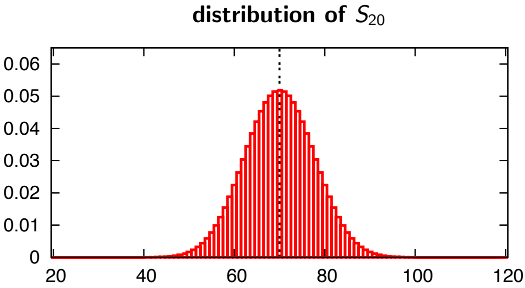


Law of Large Numbers

Let $S_n = \sum_{i=1}^{n} X_n$ be the sum of the first n outcomes.

The distribution of S_n is given by

$$P_{S_n}(x) = \frac{\#\text{ of ways to get sum } x \text{ with } n \text{ dice}}{6^n}$$



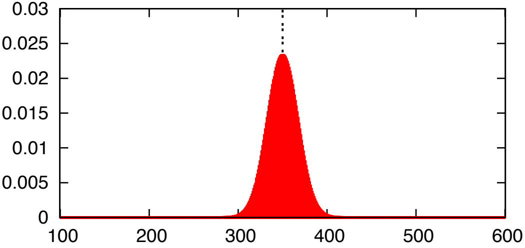
Law of Large Numbers

Let $S_n = \sum_{i=1}^{n} X_n$ be the sum of the first n outcomes.

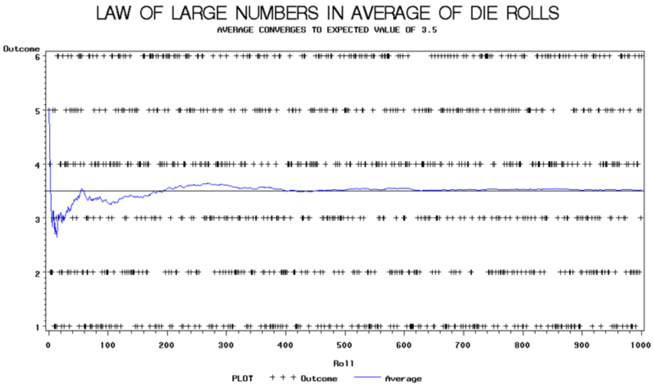
The distribution of S_n is given by

$$P_{S_n}(x) = \frac{\#\text{ of ways to get sum } x \text{ with } n \text{ dice}}{6^n}$$

distribution of S_{100}



Law of Large Numbers



Source: Wikipedia

Law of Large Numbers

Weak Law of Large Numbers

For a sequence of independent and identically distributed (i.i.d.) random variables with finite mean μ, the average $\frac{1}{n}S_n$ converges in probability to μ:

$$\lim_{n\to\infty} \Pr\left[\left|\frac{S_n}{n} - \mu\right| < \epsilon\right] = 1 \quad \text{for all } \epsilon > 0.$$

We will use the LLN to prove a result known as the Asymptotic Equipartition Property (AEP), which is a central result in information theory (we'll return to it soon enough).

Jensen's inequality

J.L.W.V. Jensen, 1859–1925

Inequalities: Jensen

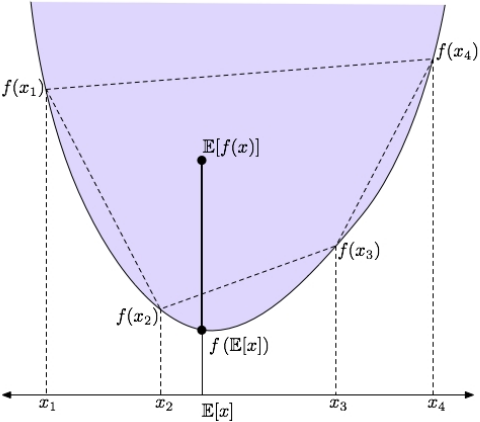
Jensen's inequality

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

$$E[f(X)] \geq f(E[X]) \ .$$

Moreover, if f is strictly convex, the inequality holds as an equality if and only if $X = E[X]$ with probability 1.

Inequalities: Jensen



Source: *Inductio Ex Machina*, `mark.reid.name/iem/`

Inequalities: Jensen

Jensen's inequality

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

$$E[f(X)] \geq f(E[X]) \ .$$

Moreover, if f is strictly convex, the inequality holds as an equality if and only if $X = E[X]$ with probability 1.

We give a proof for the first part of the theorem in the special case where X has a finite domain.

For two mass points, we have $p(x_2) = 1 - p(x_1)$, and the claim holds by definition of convexity:

$$p(x_1)\, f(x_1) + p(x_2)\, f(x_2) \geq f(p(x_1)\, x_1 + p(x_2)\, x_2) \ .$$

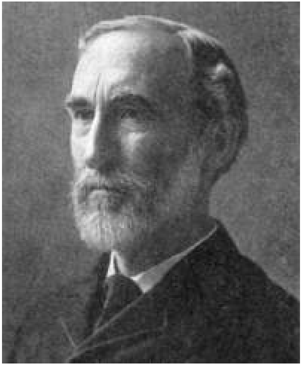
Inequalities: Jensen

Induction: Assume that $(*)$ the theorem holds for $N-1$ mass points.

$$\begin{aligned}
\sum_{i=1}^{N} p(x_i)\, f(x_i) &= p(x_N)\, f(x_N) + (1 - p(x_N)) \sum_{i=1}^{N-1} p'(x_i)\, f(x_i) \\
&\geq p(x_N)\, f(x_N) + (1 - p(x_N))\, f\left(\sum_{i=1}^{N-1} p'(x_i)\, x_i\right) \quad (*) \\
&\geq f\left(p(x_N)\, x_N + (1 - p(x_N)) \sum_{i=1}^{N-1} p'(x_i)\, x_i\right) \quad \text{(convexity)} \\
&= f\left(\sum_{i=1}^{N} p(x_i)\, x_i\right) \quad ,
\end{aligned}$$

where $p'(x_i) = \dfrac{p(x_i)}{1 - p(x_N)}$. $\square$

Gibbs' inequality



W. Gibbs, 1839–1903

Inqualities: Gibbs

Gibbs' inequality

For any two discrete probability distributions p and q, we have

$$\sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \geq \sum_{x \in \mathcal{X}} p(x) \log_2 q(x)$$

with equality if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

Proof (of the inequality part).
next slide...

Inequalities: Gibbs

Gibbs' inequality

$$\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) \geq \sum_{x\in\mathcal{X}} p(x)\log_2 q(x)$$

$$\begin{aligned}
\sum_{x\in\mathcal{X}} p(x)\log_2 q(x) - \sum_{x\in\mathcal{X}} p(x)\log_2 p(x) &= \sum_{x\in\mathcal{X}} p(x)\left(\log_2 q(x) - \log_2 p(x)\right) \\
&= \sum_{x\in\mathcal{X}} p(x)\log_2 \frac{q(x)}{p(x)} \qquad \boxed{\log_2 x - \log_2 y = \log_2 \frac{x}{y}} \\
&= E\left[\log_2 \frac{q(x)}{p(x)}\right] \leq \log_2 E\left[\frac{q(x)}{p(x)}\right] \qquad \boxed{\text{Jensen}} \\
&= \log_2 \sum_{x\in\mathcal{X}} p(x)\frac{q(x)}{p(x)} = \log_2 \sum_{x\in\mathcal{X}} q(x) = \log_2 1 = 0 \ . \qquad \square
\end{aligned}$$

What's next...

Basic concepts: entropy, mutual information, ...

Theory and applications:

- source coding theory
- noisy channel coding