Discrete Stochastics and Information Theory

Elaboration on Prof. Wöss' course

Lukas Prokop

March 18, 2015

Contents

1 Definitions and axioms of probability theory

1.1 [High] Explain the basic definitions of probability theory. What is Ω? What is σ?

Event space Ω
Defines the event space. A coin toss results in head or tail. So we could define an event space of $\{h, t\}$. Or we could consider sequences of coin tosses: $\{\text{hhh}, \text{hht}, \text{hth}, \text{thh}, \text{htt}, \text{tth}, \text{tht}, \text{ttt}\}$. Ω contains all possible outcomes. Events are (reasonable) subsets.

Subset $\mathcal{A}$
Defines a subset of possible outcomes of Ω with $\mathcal{A} \subset \mathbb{P}(\Omega)$. An example is a throw of the dice with result "even number" ($\mathcal{A} = \{2, 4, 6\}$) or "number 6" ($\mathcal{A} = \{6\}$).

σ algebra on a set $\mathcal{A}$
A set closed under complement, union and intersection. The following axioms hold:

1. $\varnothing \in \mathcal{A}$
2. $A \in \mathcal{A} \Rightarrow A^C \in \mathcal{A}$
3. If $A_n \in \mathcal{A}$ with $n = 1, 2, \ldots$ then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}$

Probability measure
Assigns a real number between 0 and 1 (inclusive) to every event. $P : \mathcal{A} \to [0, 1]$. It satisfies $P(\varnothing) = 0$ and $P(\Omega) = 1$ where P with a set parameter means "one of".

Probability space $(\Omega, \mathcal{A}, P)$
Defines a probability space $(\Omega, \mathcal{A}, P)$ consisting of an event space, the possible outcomes $\mathcal{A}$ and some probability measure.

1.2 [Low] What does σ additivity mean?

$$A_n \in \mathcal{A} \wedge A_n \cap A_m = \varnothing \quad \forall\, n \in \mathbb{N}^+, n \neq m$$
$$\Rightarrow \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

1.3 [Low] What does the "Law of continuity" tell?

The law can be derived from the definition of σ algebras:

$$A_n \in \mathcal{A}, A_1 \subset A_2 \subset A_3 \subset \ldots \Rightarrow \lim_{n\to\infty} \mathbb{P}[A]_n = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right)$$
$$B_n \in \mathcal{B}, B_1 \supset B_2 \supset B_3 \supset \ldots \Rightarrow \lim_{n\to\infty} \mathbb{P}[B]_n = \mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right)$$

1.4 [High] What is a random variable?

A random variable X is a function associating an event with a real value. It maps the probability space to real world values.

$$X : \Omega \to \mathbb{R}$$

1.5 [High] What is distribution, density, expected value, absolute convergence, continuity and monotonicity?

A *probability distribution* assigns a probability to each subset of possible outcomes of a random experiment. It can be defined by a probability mass function, probability density function, cumulative distribution function, survival function, hazard function, characteristic function or a rule to create a new random variable with a known joint probability distribution.

The *probability density function* (PDF) of a continuous random variable $X : \Omega \to \mathbb{R}$ is a function that describes the relative likelihood for this random variable to take on a given value. So there exists some $f : \mathbb{R} \to [0, \infty)$ such that

$$\mathbb{P}\left[X \in B\right] = \mathcal{P}^X(B) = \int_B f(x)\,\mathrm{d}x$$

Let X be a random variable defined on a probability space $(\Omega, \Sigma, \mathcal{P})$ then the *expected value* is defined as

$$\mathbb{E}(X) = \int_\Omega X\,\mathrm{d}\mathcal{P}$$

Let P^X operate on $(\mathbb{R}, \mathcal{B})$ which is the smallest σ algebra containing all intervals. $P^X(B)$ is the probability that B occurs with respect to a random variable X defined as

$$P^X(B) = \mathbb{P}\left[X \in \mathcal{B}\right] = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in \mathcal{B}\}\right)$$

Convergence is a property of a sequence of random variables. *Absolute convergence* means that even the sum of absolute values of the series converges as well.

A *continuous probability distribution* associates a probability to a continuous range of values. Only continuous probability distributions have a probability density function. The PDF can be *monotonic* meaning that it is a function between ordered sets that preserves the given order.

1.6 [High] What is conditional probability and independence of events?

Conditional probability is defined as division of joint probability divided by marginal probability:

$$\mathbb{P}(A \mid B) = \begin{cases} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} & \mathbb{P}(B) > 0 \\ 0 & \mathbb{P}(B) = 0 \end{cases} \tag{1}$$

$\mathbb{P}(A \mid B)$ semantically asks "assuming B to happen with certainty, what is the probability that A will happen". A and B are independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. Or in general $A_1, A_2, \ldots, A_n \in \mathcal{A}$ is independent if $\forall\, 1 \leq i_1 < i_2 < \ldots < i_k \leq n$

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\,\mathbb{P}(A_{i_2}) \cdot \ldots \cdot \mathbb{P}(A_{i_k})$$

1.7 [High] What is marginal/joint distribution? What's their relation?

Let A and B be two events occuring with *marginal* probability $\mathbb{P}(A)$ and $\mathbb{P}(B)$. The *joint probability* is intuitively defined as probability that A **and** B will happen. Formally $\mathbb{P}(A, B)$, defined as the intersection $\mathbb{P}(A \cap B)$.

$$\begin{aligned} \text{marginal probability:} \quad & p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y] \\ \text{joint probability:} \quad & p_X(x_1, x_2, \ldots, x_n) = \mathbb{P}[X = (x_1, x_2, \ldots, x_n)] \end{aligned}$$

Then the following relation holds:

$$\mathbb{P}(A \mid B) = \begin{cases} \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} & \text{if } \mathbb{P}(B) > 0 \\ 0 & \text{if } \mathbb{P}(B) = 0 \end{cases}$$

Bayes' Theorem is given by

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A) \cdot \mathbb{P}[A]}{\mathbb{P}[B]}$$

1.8 [High] Prove $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$ for independent variables in the discrete case

$$
\begin{aligned}
\mathbb{E}(X) &= \sum_x x \cdot p_X(x) \\
\mathbb{E}(Y) &= \sum_y y \cdot p_Y(y) \\
\mathbb{E}(X \cdot Y) &= \sum_{x,y} xy \cdot p_{X,Y}(x,y) \\
&= \sum_x \sum_y xy \cdot p_X(x) \cdot p_Y(y) \\
&= \left(\sum_x x \cdot p_X(x) \right) \cdot \left(\sum_y y \cdot p_Y(y) \right) \\
&= \mathbb{E}(X) \cdot \mathbb{E}(Y)
\end{aligned}
$$

1.9 [High] How is variance and covariance defined?

The expected value $\mathbb{E}(X)$ of a random variable X is also known as mean μ.

Covariance is a measure of how much two random variables change together. If the greater values of one variable primarily correspond with the greater values of the other variable (and vice versa for small values), the covariance is positive. If they are linearly disproportional, it is negative.

$$\sigma(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

Variance is a special case of covariance with only one parameter. Variance is the covariance of a variable with itself: $\sigma^2(X) = \sigma(X,X)$. It measures how far a set of numbers is spread out. A variance of zero indicates that all the values are identical. Variance is always non-negative.

$$\sigma^2(X) = \mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

$$
\begin{aligned}
&\text{discrete:} && \mathbb{V}(X) = \sum_{i=1}^{n} (p_i \cdot x_i^2) - \mu^2 \\
&\text{continuous:} && \mathbb{V}(X) = \int x^2 p(x)\,\mathrm{d}x - \mu^2
\end{aligned}
$$

Covariance is symmetrical: $\sigma(X,Y) = \sigma(Y,X)$. If X and Y are independent, then $\sigma(X,Y) = 0$.

1.10 [Low] Derive Bienaymé's equation for $\mathbb{E}(\overline{X_n})$

$$
\begin{aligned}
\overline{X_n} &= \frac{X_1 + \ldots + X_n}{n} \\
\mathbb{V}\left(\overline{X_n}\right) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}(X_i) \\
&= \frac{1}{n^2} n\, \mathbb{V}(X) \\
&= \frac{1}{n}\, \mathbb{V}(X) \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

1.11 [High] Distinguish between convergence "almost surely" and "in probability"

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables.

$\lim_{n\to\infty} X_n = X$ almost surely if $\mathbb{P}[\exists \lim X_n \wedge \lim X_n = X] = 1$

$\lim_{n\to\infty} X_n = X$ in probability if $\forall\, a > 0$, $\lim_{n\to\infty} \mathbb{P}[|X_n - X| \geq a] = 0$

Intuitively almost surely means convergence is given until infinity. In probability means n converges until a fixed constant (not necessarily infinity).

1.12 [Medium] In our notation, what is the difference between x, X, X_n and $\mathcal{X}$?

x

A specific value (realization) of a random value. For example we iterate over all values of the discrete domain. The x is typically used as iterator $x \in X$.

X

X denotes a random variable used in an equation. Often Y accompanies X as second random variable to discuss joint probability.

X_n

X_n denotes a sequence of random variables.

$\mathcal{X}$

$\mathcal{X}$ is used with two different meanings:

- In the context of Markov chains or information channels, we used $\mathcal{X}$ as the given input values for the Markov process.
- In the context of probability theory, $\mathcal{X}$ denotes the domain of a random variable X or the sequence X_n (if they all use the same domain)

1.13 [High] Define and prove Markov's inequality

Let $X \geq 0$ be a random variable with $0 < \mathbb{E}(X) < \infty$, then

$$\mathbb{P}\left[X > a \cdot \mathbb{E}(X)\right] \leq \frac{1}{a} \quad \forall\, a > 0$$

$$\begin{aligned}
A &= [X \geq a\,\mathbb{E}(X)] \\
&= \{w \in \Omega : X(w) \geq a \cdot \mathbb{E}(X)\} \\
X &\geq X\mathbb{1}_A \\
&\geq \underbrace{a\,\mathbb{E}(X)}_{\text{constant } c} \cdot \mathbb{1}_A \Leftrightarrow \\
X(w) &\geq X(W)\mathbb{1}_A(w) \\
&\geq c \cdot \mathbb{1}_A(w) \Rightarrow \\
\mathbb{E}(X) &\geq \mathbb{E}(c \cdot \mathbb{1}_A) \\
&= a\,\mathbb{E}(X)\mathbb{P}\left[A\right] \Rightarrow \frac{\mathbb{E}(X)}{\mathbb{E}(X)} \geq a\mathbb{P}\left[A\right] \Rightarrow \\
\mathbb{P}\left[A\right] &\leq \frac{1}{a}
\end{aligned}$$

1.14 [High] Define and prove Chebyshev's inequality

Let X be a random variable with finite $\mathbb{E}(X)$ and $\mathbb{V}(X)$, then $\forall\, a > 0$

$$\mathbb{P}\left[|X - \mathbb{E}(X)| \geq a\right] \leq \frac{\mathbb{V}(X)}{a^2}$$

$$\begin{aligned}
Y &= (X - \mathbb{E}(X))^2 \geq 0 \\
\mathbb{E}(Y) &= \mathbb{V}(X) \\
\mathbb{P}\left[|X - \mathbb{E}(X)| \geq a\right] &= \mathbb{P}\left[Y > a^2\right] \\
&= \mathbb{P}\left[Y \geq \frac{a^2}{\mathbb{V}(X)}\,\mathbb{E}(Y)\right] \\
&\leq \frac{1}{\frac{a^2}{\mathbb{V}(X)}} && \text{[apply Markov inequality]} \\
&= \frac{\mathbb{V}(X)}{a^2}
\end{aligned}$$

1.15 [Medium] Distinguish between the Weak and Strong Law of Large Numbers

The Weak Law of Large Numbers derives from Chebyshev's inequality and states,

> Given X_n as sequence of iid random variables, $\mu = \mathbb{E}(X_n)$ and $\sigma^2 = \mathbb{V}(X_n) < \infty$, then
>
> $$\overline{X_n} \to \mu \text{ in probability}$$

The Strong Law of Large Numbers states,

> Given X_n as sequence of iid random variables then there exists a finite random variable Y such that
>
> $$\overline{X_n} \to Y \text{ almost surely}$$

The Strong Law of Large Numbers can be equivalently defined as $\mathbb{E}(X_n)$ being finite and in this case $Y = \mathbb{E}(X_n)$ almost surely.

1.16 [High] Prove the Weak Law of Large Numbers

$$\begin{aligned}\mathbb{P}\left[\left|\overline{X_n} - \mu\right| \geq a\right] &\leq \frac{\mathbb{V}(\overline{X_n})}{a^2} && \text{[Chebyshev inequality must hold]} \\ &= \frac{\sigma^2}{na^2} \\ &\to 0\end{aligned}$$

2 Entropy

2.1 [High] Give Hartley's definition of information value

> The answer to a question that can assume the two values "yes" and "no" (without taking into account the meaning of the question) contains one unit of information.

Let U_N be a set of N elements of uniform probability; hence the information amount to identify an element is

$$\mathbb{H}(U_N) = \log_2(N) \in \mathbb{R}$$

It holds that

- $\mathbb{H}(U_2) = 1$
- $\mathbb{H}(U_N) \leq \mathbb{H}(U_{N+1})$
- $\mathbb{H}(U_{M \cdot N}) = \mathbb{H}(U_M) + \mathbb{H}(U_N)$

2.2 [Medium] Derive Shannon's entropy definition from Hartley's information value

As can be seen by the third corollary of Hartley, we can group information into (not necessarily equally large) groups of values:

$$U_N = U_{N_1} \uplus U_{N_2} \uplus \ldots \uplus U_{N_n}$$

If we want to identify the value in those groups, we need to ask two questions:

1. In which group?
2. Which element in the group?

If the value can be found in group k, then $\log_2 N_k$ questions are needed to identify the value. The average number of questions for the second question is:

$$H_2 = \sum_{k=1}^{n} \frac{N_k}{N} \log_2 N_k$$

What about H_1?

$$\begin{aligned}
\mathbb{H}(U_N) &= H_1 + H_2 \\
\log_2 N &= H_1 + \sum_{k=1}^{n} \frac{N_k}{N} \log_2 N_k \\
H_1 &= \log_2 N - \sum_{k=1}^{n} \frac{N_k}{N} \log_2 N_k \\
&= \sum_{k=1}^{n} \frac{N_k}{N} \log_2 N - \sum_{k=1}^{n} \frac{N_k}{N} \log_2 N_k \\
&= -\sum_{k=1}^{n} \frac{N_k}{N} \log_2 \frac{N_k}{N} \\
&= -\sum_{k=1}^{n} p_k \log_2 p_k \qquad \text{with } p_k = \frac{N_k}{N}
\end{aligned}$$

2.3 [High] Give Shannon's definition of *entropy* and provide examples for small/large entropy

Entropy is a measure of uncertainty of a random variable. Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p_X(x) = \mathbb{P}[X = x]$ for $x \in \mathcal{X}$. The entropy $\mathbb{H}(X)$ of a discrete random variable X is defined by

$$\mathbb{H}(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x) = -\mathbb{E}(\log_2 p_X(x))$$

$\mathbb{H}(X) \geq 0$ holds for all X. Furthermore we define

$$0 \log 0 = 0 \quad \forall\, b \geq 0 \qquad a \log \frac{a}{0} = \infty \quad \forall\, a > 0$$

For example a uniform distribution of two discrete values gives

$$\mathbb{H}(X) = -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2}\right) = 1$$

Uniform distribution always provides a *large* value of entropy. Certain events provide a *small* value. For example consider two events, where only event A always happens:

$$\mathbb{H}(X) = -\left(0 \cdot \log_2 0 + 1 \cdot \log_2 1\right) = (1 \cdot 0) = 0$$

The following holds:

- $\mathbb{H}(X)$ depends only on the probabilities; not on the specific interpretation.
$$f : \mathcal{X} \to \mathcal{X}' \wedge X' = f(X) \Rightarrow \mathbb{H}(X') = \mathbb{H}(X)$$
- $p_0 \mapsto \mathbb{H}(p_0, 1 - p_0)$ is continuous. The maximum is attained at $\frac{1}{2}$:
$$\mathbb{H}(p_0, 1 - p_0) = -p_0 \log_2(p_0) - (1 - p_0) \log_2(1 - p_0)$$
$$-\frac{1}{2} \log_2 \left(\frac{1}{2}\right) - \left(1 - \frac{1}{2}\right) \log_2 \left(1 - \frac{1}{2}\right) = +\frac{1}{2} - \left(-\frac{1}{2}\right) = 1$$
- For some fixed n,
$$\mathbb{H}(p_1, \ldots, p_n) \leq \mathbb{H}\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_2(n)$$

2.4 [High] What is joint entropy? What is the entropy of conditional distribution?

Given a pair of discrete random variable (X, Y). The joint entropy of X and Y is defined as:

$$\begin{aligned}\mathbb{H}(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) \\ &= -\mathbb{E}\left(\log_2 p(X, Y)\right)\end{aligned}$$

Entropy of conditional distribution:

$$\begin{aligned}\mathbb{H}(Y \mid X) &= -\sum_{y} p(y \mid x) \log_2 p(y \mid x) \\ &= \sum_{x} p_X(x) \mathbb{H}(Y \mid X = x)\end{aligned}$$

2.5 [Medium] Prove that $\mathbb{H}(Y \mid X) = \mathbb{H}(X, Y) - \mathbb{H}(X)$

$$\begin{aligned}\mathbb{H}(Y \mid X) &= -\sum_{x}\sum_{y} p_X(x) p_{X,Y}(y \mid x) \log_2 p_{X,Y}(y, x) \\ &= -\sum_{x}\sum_{y} p_{X,Y}(x, y) \left[\log_2 p_{X,Y}(x, y) - \log_2 p_X(x)\right] \\ &= \mathbb{H}(X, Y) + \sum_{x} \log_2 p_X(x) \sum_{y} p_{X,Y}(x, y) \\ &= \mathbb{H}(X, Y) - \mathbb{H}(X)\end{aligned}$$

2.6 [High] What is the Kullback-Leibler distance? Are the parameters symmetrical?

Let p and q be two probability distributions on the finite set $\mathcal{X}$. X is a random variable with distance probability p. The *relative entropy* or *Kullback-Leibler distance* of p with respect to q is

$$\mathbb{D}(p \parallel q) = \sum_{x} p(x) \log_2 \frac{p(x)}{q(x)} = \mathbb{E}\left(\log_2 \frac{p(x)}{q(x)}\right)$$

- Be aware that for $p(x) > 0$ and $q(x) = 0$: $\mathbb{D}(p \parallel q) = \infty$.
- $\mathbb{D}(p \parallel q)$ is always non-negative.
- From $\mathbb{D}(p \parallel q) = 0$ it follows that $p(x) = q(x)\ \forall x \in \mathcal{X}$.
- $\mathbb{D}(p \parallel q) \neq \mathbb{D}(q \parallel p)$
- Intuitively it is a measure of the information lost when q is used to approximate p.

2.7 [High] What is mutual information? What is conditional mutual information? What is the chain rule?

Let X and Y be a pair of random variables. *Condition mutual information* is defined as,

$$
\begin{aligned}
\mathbb{I}(X,Y) &= \mathbb{D}\left(p_{(X,Y)} \;\middle\|\; p_X \bigotimes p_Y\right) \\
&= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p_X(x) \cdot p_Y(y)} \\
&= \mathbb{H}(X) - \mathbb{H}(X \mid Y) = \mathbb{H}(Y) - \mathbb{H}(Y \mid X)
\end{aligned}
$$

Intuitively, mutual information measures the information that X and Y share. It measures how much knowing one of these variables reduces uncertainty about the other.

Specifically it holds that $\mathbb{I}(X,X) = \mathbb{H}(X)$. The *chain rule* states that

$$
\mathbb{H}(X_1, \ldots, X_n) = \sum_{k=1}^{n} \mathbb{H}(X_k \mid X_{k-1}, \ldots, X_1)
$$

The chain rule of mutual information is defined as,

$$
\mathbb{I}((X_1, \ldots, X_k); Y) = \sum_{k=1}^{n} \mathbb{I}(X_k; Y \mid X_{k-1}, \ldots, X_1)
$$

Conditional mutual information of X and Y given Z:

$$
\begin{aligned}
\mathbb{I}(X; Y \mid Z) &= \sum_{z} p_Z(z) \mathbb{I}(X; Y \mid Z = z) \\
&= \begin{cases} \mathbb{H}(Y \mid Z) - \mathbb{H}(Y \mid X, Z) \\ \mathbb{H}(X \mid Z) - \mathbb{H}(X \mid Y, Z) \end{cases}
\end{aligned}
$$

2.8 [Medium] What is Jensen's inequality?

Given a continuous, convex function $f : I \to \mathbb{R}$ (with I as open interval) and X as random variable,

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

2.9 [Medium] What is information inequality?

Let $p(x)$ and $q(x)$ be a pair of probability distributions with $x \in \mathcal{X}$, then $\mathbb{D}(p \parallel q) \geq 0$.

Equality holds if and only if $p(x) = q(x) \quad \forall\, x$.

Corollaries:

- $\mathbb{I}(X;Y) \geq 0$

- $\mathbb{H}(X \mid Y) \leq \mathbb{H}(X)$
- $\mathbb{I}(X;Y \mid Z) \geq 0$
- $\mathbb{H}(X_1, \ldots, X_n) \leq \sum_{k=1}^{n} \mathbb{H}(X_k)$

2.10 [Medium] What is the log-sum inequality?

Given $a_1, \ldots, a_n$ and $b_1, \ldots, b_n \geq 0$, then

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

Given $p^{(1)}(\cdot), p^{(2)}(\cdot)$ and $q^{(1)}(\cdot), q^{(2)}(\cdot)$ as probability densities on $\mathcal{X}$ with $0 \leq \lambda \leq 1$

$$\Rightarrow \mathbb{D}\left(\lambda \cdot p^{(1)} + (1-\lambda)p^{(2)} \;\middle\|\; \lambda \cdot q^{(1)} + (1-\lambda) \cdot q^{(2)}\right) \leq \lambda \mathbb{D}\left(p^{(1)} \;\middle\|\; q^{(1)}\right) + (1-\lambda)\mathbb{D}\left(p^{(2)} \;\middle\|\; q^{(2)}\right)$$

2.11 [Medium] What is a Markov chain?

A *Markov chain* is a random process satisfying the Markov property which means transitions happen memoryless. Given a state space, transitions can happen between any two states with associated probabilities. Transition probabilities are stored in the so-called "stochastic matrix".

Hence a Markov chain is a sequence of random variables $X_1, X_2, X_3, \ldots$ with the Markov property, namely that, given the present state, the future and past states are independent. Formally $\mathbb{P}(X_{n+1} \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = x \mid X_n = x_n)$.

A Markov chain is ergodic iff every transition probability is greater zero. A Markov triple is a Markov chain of three states.

2.12 [High] Define Data Processing inequality

Let $X \to Y \to Z$ be a Markov triple **in that order** meaning X and Z are independent but conditional on Y. Hence

$$\mathbb{P}[X = x, Z = z \mid Y = y] = \mathbb{P}[X = x \mid Y = y] \cdot \mathbb{P}[Z = z \mid Y = y]$$

It holds that

$$\mathbb{I}(X;Y) \geq \mathbb{I}(X;Z)$$

The information that Y knows about X cannot be increased by manipulating (processing) Y deterministically or at random.

2.13 [Medium] Prove Data Processing inequality

The chain rule tells

$$\mathbb{I}((Y,Z);X) = \mathbb{I}(Y;X) + \mathbb{I}(Z;X \mid Y)$$

So,

$$\begin{gathered}\mathbb{I}(X;Y,Z) = \mathbb{I}(X;Y,Z) \\ \mathbb{I}(X;Y) + \underbrace{\mathbb{I}(X;Z \mid Y)}_{\geq 0} = \mathbb{I}(X;Z) + \underbrace{\mathbb{I}(X;Y \mid Z)}_{=0} \\ \mathbb{I}(X;Y) \geq \mathbb{I}(X;Z)\end{gathered}$$

2.14 [High] Define and prove Fano's inequality

Define a Markov triple $X \to Y \to \hat{X}$ with $\mathbb{P}_{\text{err}}$ as the error that $X \neq \hat{X}$.

$$H(\mathbb{P}_{\text{err}}) + \mathbb{P}_{\text{err}} \log |\mathcal{X}| \geq \mathbb{H}\left(X \;\middle|\; \hat{X}\right) \geq \mathbb{H}(X \mid Y)$$

Equivalently

$$\begin{gathered}1 + \mathbb{P}_{\text{err}} \log |\mathcal{X}| \geq \mathbb{H}(X \mid Y) \\ \mathbb{P}_{\text{err}} \geq \frac{\mathbb{H}(X \mid Y) - 1}{\log_2 |\mathcal{X}|}\end{gathered}$$

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases}$$

$$\begin{gathered}\mathbb{H}\left(E, X \;\middle|\; \hat{X}\right) = \mathbb{H}\left(E, X \;\middle|\; \hat{X}\right) \\ \underbrace{\mathbb{H}\left(E \;\middle|\; \hat{X}\right)}_{\leq \mathbb{H}(\mathbb{P}_{\text{err}})} + \underbrace{\mathbb{H}\left(X \;\middle|\; E, \hat{X}\right)}_{\leq \mathbb{P}_{\text{err}} \log |\mathcal{X}|} = \mathbb{H}\left(X \;\middle|\; \hat{X}\right) + \underbrace{\mathbb{H}\left(E \;\middle|\; X, \hat{X}\right)}_{=0}\end{gathered}$$

$\mathbb{H}\left(X \;\middle|\; E, \hat{X}\right)$ is bounded by

$$\begin{aligned}\mathbb{H}\left(X \;\middle|\; E, \hat{X}\right) &= \mathbb{P}[E=0]\,\mathbb{H}\left(X \;\middle|\; \hat{X}, E=0\right) + \mathbb{P}[E=1]\,\mathbb{H}\left(X \;\middle|\; \hat{X}, E=1\right) \\ &\leq (1 - \mathbb{P}_{\text{err}}) \cdot 0 + \mathbb{P}_{\text{err}} \log |\mathcal{X}|\end{aligned}$$

Hence, we obtain Fano's inequality. Intuitively it relates the probability of error in guessing the random variable X to its conditional entropy $\mathbb{H}(X \mid Y)$. Given X as a function of Y. Iff X can estimate Y with zero probability of error, then $\mathbb{H}(X \mid Y) = 0$. Fano's inequality uses this fact: Estimate X with a low probability of error if the conditional entropy $\mathbb{H}(X \mid Y)$ is small.

3 Asymptotic entropy

3.1 [Low] What does iid mean?

iid is a property of a set of random variables. A set of random variables can be "independent, identically distributed" meaning that they all utilize the same probability distribution, but are independent in every possible form.

3.2 [High] What is a stochastic process?

A stochastic process (in discrete time) is a sequence of random variables

$$(X_n)_{n\geq 1}$$

3.3 [High] What is asymptotic entropy?

Given X_n as discrete random variables and its values in some finite/countable set $\mathcal{X}$. The *asymptotic entropy* or *asymptotic rate* h of the stochastic process is defined as

$$h = \lim_{n\to\infty} \frac{1}{n}\mathbb{H}(X_1, \ldots, X_n)$$

Furthermore

$$h' = \lim_{n\to\infty} \mathbb{H}(X_n \mid X_{n-1}, \ldots, X_1)$$

if the limit exists. If h' exists, then h exists and $h = h'$.

Convergence is interpreted as

$$\lim_{n\to\infty} a_n = a \Leftrightarrow \forall\, \varepsilon > 0\, \exists\, N_\varepsilon : \forall\, n > N_\varepsilon : |a_n - a| < \varepsilon$$

3.4 [High] What is a stationary distribution?

A distribution $(X_n)_{n\geq 1}$ is stationary if $\forall n, l \in \mathbb{N} : (X_1, \ldots, X_n)$ and $(X_{l+1}, \ldots, X_{l+n})$ have the same joint distribution.

$$\forall\, x_1, \ldots, x_n \in \mathcal{X} : \mathbb{P}[X_1 = x_1, \ldots, X_n = x_n] = \mathbb{P}[X_{l+1} = x_1, \ldots, X_{l+n} = x_n]$$

If $(X_n)_{n\geq 1}$ is stationary, h' exists.

3.5 [High] What is time homogeneity? What is a stochastic matrix?

A Markov chain is called time-homogeneous if

$$p_n(y \mid x) = p_m(y \mid x) \qquad \forall\, m, n$$

Hence, transition probability do not depend on time.

3.6 [High] Prove that at least one stationary distribution exists for every stochastic matrix.

Formally, a distribution at time n is given as

$$\mathbb{P}[X_n = y] = \sum_{x \in \mathcal{X}} \mathbb{P}[X_n = y \mid X_0 = x] \cdot \mathbb{P}[X_0 = x] = \sum_x p_{x,y}^{(n)} \mu_x = (\mu P^n)_x$$

Now for an arbitrary stochastic matrix $P = (p_{x,y})_{x,y \in \mathcal{X}}$ at least one stationary distribution ν exists.

Consider an initial distribution $\mu = \mu_0$.

$$\left(\mu + \mu P + \mu P^2 + \cdots + \mu P^{n-1}\right) \frac{1}{n} = \mu_n$$

μ_n is a probability vector on $\mathcal{X}$ for various n. A subsequence $(\mu_{n_k})_{k \in \mathbb{N}}$ exists which converges to some vector ν.

$$\begin{aligned}
\sum_{x \in \mathcal{X}} \nu_X &= \sum_{x \in \mathcal{X}} \lim \mu_{n_k} \\
&= \lim \sum_{x \in \mathcal{X}} \mu_{n_k} && \text{[because } \mathcal{X} \text{ is finite]} \\
&= 1 \\
\mu_n^P - \mu_n &= \frac{1}{n}\left(\mu P + \mu P^2 + \cdots + \mu P^n\right) \\
&= \frac{1}{n}\left(\mu + \mu P + \cdots + \mu P^{n+1}\right) \\
&= \frac{1}{n} \underbrace{\left(\mu - \mu P^n\right)}_{\text{bounded by 1}} \\
&\to 0
\end{aligned}$$

$$\begin{array}{ccccc}
\mu_{n_k} P & - & \mu_{n_k} & \to & 0 \\
\downarrow & & \downarrow & & \\
\nu P & - & \nu & = & 0
\end{array}$$

3.7 [Medium] What does irreducibility of $(\mathcal{X}, P)$ mean?

$(\mathcal{X}, P)$ is called irreducible[1] if $\forall\, x, y \in \mathcal{X} : \exists n = n_{x,y} : p_{x,y}^n > 0$. Graph theoretically irreducibility means strongly connectedness.

[1] for finite $\mathcal{X}$ it is called ergodic

If $(\mathcal{X}, P)$ is finite and irreducible then there is a unique stationary distribution μ and

$$\nu_x = \frac{1}{\mathbb{E}_X(t^x)}$$

where $\mathbb{E}_X(t^x)$ denotes the expected return time to x with $t^x_{(w)} = \inf\left\{n \geq 1 : X_n^{(w)} = x\right\}$. t^x is almost surely finite: $\mathbb{P}\left[t^x < \infty\right] = 1$ and $\mathbb{E}_X(t^x) < \infty$.

3.8 [Low] Discuss a random walk.

3.9 [Medium] Under which circumstances does a unique stationary initial distribution exist? Prove it.

A unique stationary initial distribution ν exists, if $(\mathcal{X}, P)$ is irreducible.

Proof: missing

4 Asymptotic equipartition

4.1 [High] What is asymptotic equipartition?

Suppose that $(X_n)_{n\geq 1}$ is an $\mathcal{X}$-valued stochastic process with entropy rate h. Then the process is said to have asymptotic equipartition property (AEP), if

$$-\frac{1}{n}\log_2 p_n(X_1, \ldots, X_n) \to h \text{ almost surely}$$

(Sometimes convergence in probability suffices)

4.2 [High] Under which conditions does the asymptotic equipartition property hold? What is it good for?

$$-\frac{1}{n}\log_2 p_n(X_1, \ldots, X_n) \to h \text{ almost surely}$$

must hold. Furthermore the AEP is good for data compression.

4.3 [Medium] What is a typical set?

The typical set is defined as

$$A_\varepsilon^{(n)} = \left\{(x_1, \ldots, x_n) \in \mathcal{X}^n : \left|-\frac{1}{n}\log_2 p_n(x_1, \ldots, x_n) - h\right| < \varepsilon\right\}$$

where $\varepsilon > 0$. The following theorems hold for typical sets

- $2^{-n(h+\varepsilon)} < p_n(x_1, \dots, x_n) < 2^{-n(h-\varepsilon)} \quad \forall\, (x_1, \dots, x_n) \in A_\varepsilon^{(n)}$
- $\mathbb{P}\left[(X_1, \dots, X_n) \in A_\varepsilon^{(n)}\right] > 1 - \varepsilon \quad \forall\, n \geq N_\varepsilon$
- $(1-\varepsilon)2^{n(h-\varepsilon)} \leq \left|A_\varepsilon^{(n)}\right| \leq 2^{n(h+\varepsilon)}$

Those theorems follow from the definition of typical sets.

5 Coding and compression

5.1 [High] How is expected and average code length defined?

Expected code length
$\sum_{(x_1,\dots,x_n)} l(C(x_1, \dots, x_n))p_n(x_1, \dots, x_n)$

Average code length
$\frac{1}{n}\,\mathbb{E}\,(l(x_1, \dots, x_n) \leq h + \varepsilon' \qquad \forall\, n > N_\varepsilon$

5.2 [Medium] What does the Ergodic theorem state?

For finite irreducible Markov chains $(\mathcal{X}, P)$ it holds that for any $f : \mathcal{X} \to \mathbb{R}$

$$\begin{aligned} \frac{1}{n}\left(f(X_0) + f(X_1) + \cdots + f(X_{n-1})\right) &= \sum_{x \in \mathcal{X}} f(X)\nu_x \text{ almost surely} && \text{[discrete case]} \\ &= \int_{\mathcal{X}} f \,\mathrm{d}\nu && \text{[continuous case]} \end{aligned}$$

6 Codes

6.1 [High] Define the properties non-singularity, unique decodability and prefix-freedom.

nonsingular $C : \mathcal{X} \to \Sigma^+$ is injective

uniquely decodable $C : \mathcal{X}^* \to \Sigma^*$ is injective

instantaneous or prefix-free $\forall\, x, y \in \mathcal{X} : C(x)$ is not a prefix of $C(y)$

6.2 [High] What is the Theorem Kraft Inequality?

Let $D = |\Sigma|$ and $C : \mathcal{X} \to \Sigma$ is a prefix-free code. It holds that

$$\sum_{x \in \mathcal{X}} D^{-l(C(x))} \leq 1$$

Conversely if $l_X \in \mathbb{N}$ $(x \in \mathcal{X})$ are such that $\sum_{x \in \mathcal{X}} D^{-l_x} \leq 1$ then there is a prefix-free code C with $C(L(x)) = L_x \, \forall \, x \in \mathcal{X}$.

6.3 [Low] Define a general integer optimization problem.

6.4 [Low] What are Lagrange multipliers and what are they used for?

6.5 [Medium] Which theorem regarding code length holds for every prefix-free code?

For every prefix-free code $C : \mathcal{X} \to \Sigma^+$ the expected code length satisfies $L_C \geq H_D(X)$. If equality holds, then $p(x) = D^{-l(C(x))}$. Remember that $D = |\Sigma|$.

It follows that if p is D-adic[2] then the minimum value $H_D(X)$ is attained. We can set $l_X = -\log_D p(x)$ satisfy Kraft equality.

Otherwise minimise $\log_D B + \frac{1}{\log_2 D} \mathbb{D}\left(p \parallel r\right)$.

6.6 [High] Which theorem holds for optimal code lengths?

Optimal code lengths l_x^* with $x \in \mathcal{X}$ are such that the expected code length satisfies $H_D(X) \leq L^* < H_D(X) + 1$.

6.7 [High] How does the Huffman algorithm work?

6.8 [High] Prove that Huffman codes are optimal.

6.9 [Medium] Which codes are canonical? Are Huffman codes canonical?

Canonical codes satisfy the following properties:

1. if $p(x) > p(y)$ then $l(C(x)) \leq l(C(y))$
2. If $v, w \in C(\mathcal{X})$ are longest code words, then $l(v) = l(w)$

[2] meaning $p(x) \in \{D^{-n} : n \in \mathbb{N}\}$

3. C can be modified into another optimal prefix-free code C'. C' satisfies: Let v, w be the least likely symbols
   - v and w are siblings in the coding tree
   - $\forall\, z \in \mathcal{X} \setminus \{x, y\} : p(z) \geq p(x) \geq p(y)$

Huffman codes are canonical.

7 Information channels

7.1 [High] Define discrete channels and the n-th extension of $\mathcal{C}$ without feedback.

A *discrete memoryless channel* $\mathcal{C} = (\mathcal{X}, P, Y)$ with $P = (p(y \mid x))_{x \in \mathcal{X}}$ with $p(y \mid x)$ is the probability that the outcome is y given that the input is x.

The *n-th extension of* $\mathcal{C}$ *without feedback* is the channel

$$\mathcal{C}^n = (\mathcal{X}^n, P_n, Y^n)$$

7.2 [Medium] Give examples for channels.

- Noise-less binary channel
- Channel with non-overlapping outputs
- Noisy typewriter
- Binary symmetric channel
- Binary erasure channel

7.3 [High] How is capacity of a channel defined?

$$\max \{\mathbb{I}(X; Y) : p_X \in M(\mathcal{X})\}$$

where $M(\mathcal{X}) = \{p(\cdot)$ probability distributions on $\mathcal{X}\}$ given $P = (p(y \mid x))_{x \in \mathcal{X}}$ and p_X.

7.4 [Medium] Derive the capacity of a binary symmetric channel.

$$\mathrm{Cap}(\mathcal{C}) = 1 - \mathbb{H}(p, 1-p)$$

7.5 [Medium] What is the mutual information of a binary erasure channel?

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X \mid Y)$$

7.6 [Medium] When is P of an information channel called weakly symmetric? Which associated theorem exists?

$P = (p(y \mid x))_{x \in \mathcal{X}, y \in \mathcal{Y}}$ is called weakly symmetric if

1. any two rows $p(\cdot \mid x), p(\cdot \mid x')$ are permutations of each other
2. all column sums are the same

$$\exists c : \sum_x p(y \mid x) = c \qquad \forall\, y \in Y$$

If P is weakly symmetric, then $\text{Cap}(\mathcal{C}) = \log_2 |Y| - \mathbb{H}(p(\cdot \mid x))$ and the maximum is achieved when X is uniform on $\mathcal{X}$.

7.7 [Medium] Prove $\text{Cap}(\mathcal{C}^n) = n \cdot \text{Cap}(\mathcal{C})$

7.8 [High] What is a (M, n) code? What is the average and maximum error of it?

An (M, n) code for the channel $C = (\mathcal{X}, P, Y)$ consists of the following:

1. An input set W with $|W| = M$ or wlog. $W = \{1, \ldots, M\}$
2. The codebook B
3. A decoding function $g : Y^n \to \hat{W}$

The average error is defined as

$$\begin{aligned}\lambda_w^{(n)} &= \mathbb{P}\left[g(Y^{(n)}) \neq w \;\middle|\; X^{(n)} = x^{(n)}(w)\right] \\ &= \sum_{y \in Y} p(y \mid x(w))\end{aligned}$$

The maximum probability of the error is $\lambda^{(n)} = \max_w \lambda_w^{(n)}$.

7.9 [High] What is the rate of a code?

$$R = \frac{\log_2 M}{n}$$

7.10 [Medium] What is an achievable rate? What is an achievable capacity?

A rate $R > 0$ is achievable if there is a sequence of (M_n, n) codes such that

$$R_n = \frac{\log_2 M_n}{n} \to R$$

and

$$\lambda^{(n)} \to 0$$

The achievable capacity R^* of a channel is the supremum of all achievable rates:

$$\forall\, \varepsilon > 0 \qquad \exists (M, n) \text{ code with rate } R > R^* - \varepsilon \wedge \lambda^{(n)} < \varepsilon$$

7.11 [Medium] What is the set of jointly typical sequences?

7.12 [High] Define Shannon's 2nd theorem and give an overview over the proof.

$$R^* = \mathrm{Cap}(\mathcal{C})$$

8 List of theorems

See also

- Markov inequality
- Chebyshev inequality
- Weak and Strong Law of Large Numbers

Theorem 1. If X, Y are independent random variables with finite $\mathbb{E}(X)$ and $\mathbb{E}(Y)$, then

$$\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

Theorem 2. If $X_n \to X$ almost surely then $X_n \to X$ in probability.

Theorem 3. $X_n \to X$ almost surely $\Leftrightarrow$ $U_k \to 0$ in probability with $U_k = \sup\{|X_n - X| : n \geq k\}$

Theorem 4. If $\lim X_n = X$ almost surely and $\lim X_n = X'$ almost surely, then $X = X'$ almost surely. If $\lim X_n = X$ in probability and $\lim X_n = X'$ in probability, then $X = X'$ almost surely.

Theorem 5. The function $N \mapsto H(U_N) = \log_2 N$ is the **unique** function satisfying the associated 3 axioms to Hartley formula:

- $H(U_2) = 1$

- $H(U_N) \leq H(U_{N+1})$ ("monotonicity")
- $H(U_{M \cdot N}) = H(U_M) + H(U_N)$

It is also the **unique** function satisfying

- $H(U_{N+1}) - H(U_N) \xrightarrow{N \to \infty} 0$

Theorem 6.

$$\begin{aligned} \mathbb{H}(Y \mid X) &= \mathbb{H}(X, Y) - \mathbb{H}(X) \\ \Leftrightarrow \mathbb{H}(X, Y) &= \mathbb{H}(X) + \mathbb{H}(Y \mid X) \end{aligned}$$