Information Theory

> Reference: section 3, *Quantum Cryptography and Secret-Key Distillation*
> See also: https://en.wikipedia.org/wiki/Information_theory

Source Coding

Shannon Entropy

> See also: https://en.wikipedia.org/wiki/Entropy_(information_theory)

The *Shannon Entropy* of a discrete random variable X is denoted by $H(X)$ and is defined as follow, all logarithms are in base 2.

$H(X) = -\sum_x P_X(x) \log P_X(x)$

The *entropy* of a binary random variable with distribution $\{p, 1-p\}$ for $0 \leq p \leq 1$ is denoted as

$h(p) = -p \log p - (1-p) \log(1-p).$

The *Shannon Entropy* is a property of the distribution $P_X(x)$, but not the symbol set $\mathcal{X}$.

For a particular encoding method, the entropy of a random variable tells us precisely about its compressibility. More on *Huffman Coding* later.

Properties of source codes

> See also: https://en.wikipedia.org/wiki/Shannon's_source_coding_theorem

The source coding theorem establishes a limit of data compression.

Consider a source code α, for a random variable X is a **mapping** from $\mathcal{X}$ to the *codewords* $\{0,1\}^*$, the set of finite binary strings, and the start denotes the concatenation of zero, one or any finite number of symbols.

We consider only binary codes. The codeword associated with $x \in \mathcal{X}$ is written as $\alpha(x)$.

We can define the avarage length of a source code as $L(\alpha) = \sum_{x \in \mathcal{X}} P_X(x)|\alpha(x)|$.

A code α is *non-singular* if every $x \in \mathcal{X}$ is encoded into a different codeword $\alpha(x)$. This ensures that we can decode x given its codeword $\alpha(x)$.

We say that a code is *uniquely decodable* if for all strings $\bar{a} \in \mathcal{X}$, the resulting codeword $\alpha(\bar{x})$ is different.

The binary string s_1 is said to be a *prefix* of s_2 if the $|s_1|$ first bits of s_2 are equal to those of s_1. And if $s1 \neq s_2$, we say $s1$ is a *proper prefix* of s_2.

A code is said to be *instantaneous* or *prefix-free* if no codeword is a prefix of another codeword.

Huffman coding

Huffman codes are an example of prefix-free codes. See https://en.wikipedia.org/wiki/Huffman_coding for details.

Arithmetic coding

Arithmetic coding is an alternative of Huffman coding. See https://en.wikipedia.org/wiki/Arithmetic_coding for details.

Channel Coding

> See also: https://en.wikipedia.org/wiki/Coding_theory#Channel_coding

Channel coding is the most important question addressed by information theory. It consists in finding the most efficient way to transmit information over a potentially noisy channel.

A channel is characterized by an input alphabet $\mathcal{X}$, the symbols that the **sender** can transmit, and an output alphabet $\mathcal{Y}$, the symbols that the **receiver** gets.

If we do not consider other condition, we can model the channel by a probability transition matrix $p(y|x)$, which expresses the *probability of observing the output symbol y given that the input symbol x*. This matrix already considers all of the events may happen during channel transmission.

Measuring quality of channel transmission: *mutual information* between two random variables. The mutual information between X and Y is written as $I(X;Y)$ and denoted as

$I(X;Y) = H(X) + H(Y) - H(X,Y) = I(Y;X)$.

It satisfies $I(X;Y) \geq 0$ with equality X and Y are independent.

The fundamental upper bound on the rate of transmission is $I(X;Y)$ bits per channel use. The maximization defines as *channel capacity* on the input distribution $P_X(x)$,

$C = \max_{P_X(x)} I(X;Y)$.

Error-correcting codes

> See also: https://en.wikipedia.org/wiki/Error_correction_code

Error-correcting codes are methods to encode information in such a way that they are made resistant against errors caused by the channel over which they are transmitted.

There are well-know error-correcting codes call *linear codes*.

Markov chains

> See also: https://en.wikipedia.org/wiki/Markov_chain

The three random variables $X \to Y \to Z$ are said to form a *Markov chain* (in that order) if the joint probability distribution can be written as $P_{XYZ}(x, y, z) = P_X(x) P_{Y|X}(y|x) P_{Z|Y}(z|y)$.

Rényi Entropies

> See also: https://en.wikipedia.org/wiki/Rényi_entropy

The Rényi entropies form a family of functions on the probability distributions, which generalize (and include) the Shannon entropy.

The *Rényi entropy* of order r, with $0 < r < \infty$ and $r \neq 1$, of X is defined as,

$H_r(X) = \frac{1}{1-r} \log \sum_x (P_X(x))^r$.

For $r = 0, 1, \infty$, we conventionally define,

$H_0(X) = \log |\{x \in \mathcal{X} : P_X(x) > 0\}|$,

the logarithm of support size of X;

$H_1(X) = H(X)$,

the regular Shannon entropy; and

$H_\infty(X) = -\log \max_x P_X(x)$, the negative logarithm of the largest symbol probability.

An important particular case is the order-2 Rényi entropy $H_2(X) = -\log \sum P_X^2(x)$, which is in fact the negative logarithm of the *collision probability*.

> The *collision probability* $\sum P_X^2(x)$ is the probability that two independent realizations of the random variable X are equal.

For a random variable U with uniform distribution, the order-2 Rényi and Shannon entropies match. For any other random variable, the order-2 Rényi entropy is smaller than its Shannon entropy.

The joint Rényi entropy of multiple random variables is calculated over their joint probability distribution, and satisfy $H_r(X,Y) \leq H_r(X) + H_r(Y)$.

The conditional Rényi entropy can be defined as $H_r(X|Y) = \sum_{y \in y} P_Y(y) H_r(X|Y=y)$.

Continuous Variables

In this section, we will treat on continuous variables rather than discrete variables.

Differential entropy

> See also: https://en.wikipedia.org/wiki/Differential_entropy

The *differential entropy* of a continuous random variable X is defined as,

$H(X) = -\int_x dx p_X(x) \log p_X(x)$.

The fundamental difference between differential entropy and Shannon entropy: The differential entropy is sensitive to an invertible transformation of the symbols, while the Shannon entropy is not. So, in differential entropy, $H(X) < 0$ can happen.

The *conditional differential entropy* is usually defined as $H(X|Y) = H(X,Y) - H(Y)$ for continuous random variable X and Y.

The mutual information is $I(X;Y) = H(X) + H(Y) - H(X,Y)$, and $I(X;Y) \geq 0$.

Gaussian variables and Gaussian channel

> See also:
> https://en.wikipedia.org/wiki/Differential_entropy#Maximization_in_the_normal_distribution

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian variable, with the mean μ and standard deviation Σ, i.e.

$p_X(x) = \frac{1}{\Sigma\sqrt{2\pi}} e^{-\frac{(x-y)^2}{2\Sigma^2}}$

The differential entropy of X is $H(X) = 2^{-1} \log(2\pi e \Sigma^2)$.

Let X be transmitted through a *Gaussian channel*, then we can get,

> Gaussian channel: a channel which adds a Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma)$ of standard deviation σ on the signal, giving $Y = X + \epsilon$ as the output.

- The entropy of output Y on input X is distributed as a Gaussian with standard deviation σ.
- The entropy of Y is conditional on X, $H(Y|X) = 2^{-1} \log(2\pi e \sigma^2)$ bits.

- The distribution of Y is Gaussian with variance $\Sigma^2 + \sigma^2$, then $H(Y) = 2^{-1}\log(2\pi e(\Sigma^2 + \sigma^2))$ bits.
- The mutual information on this transmission is $I(X;Y) = H(Y) - H(Y|X) = \frac{1}{2}\log(1 + \frac{\Sigma^2}{\sigma^2})$.
- $\frac{\Sigma^2}{\sigma^2}$ is called the *signal-to-noise ratio* (snr).

By theory, a Gaussian channel can transmit an arbitrarily high number of bits if the input distribution has a sufficiently high standard deviation Σ.

Gaussian distribution yields the *best* rate for a given variance. The capacity of a Gaussian channel can be written as $\Sigma = \Sigma_{\max}$.