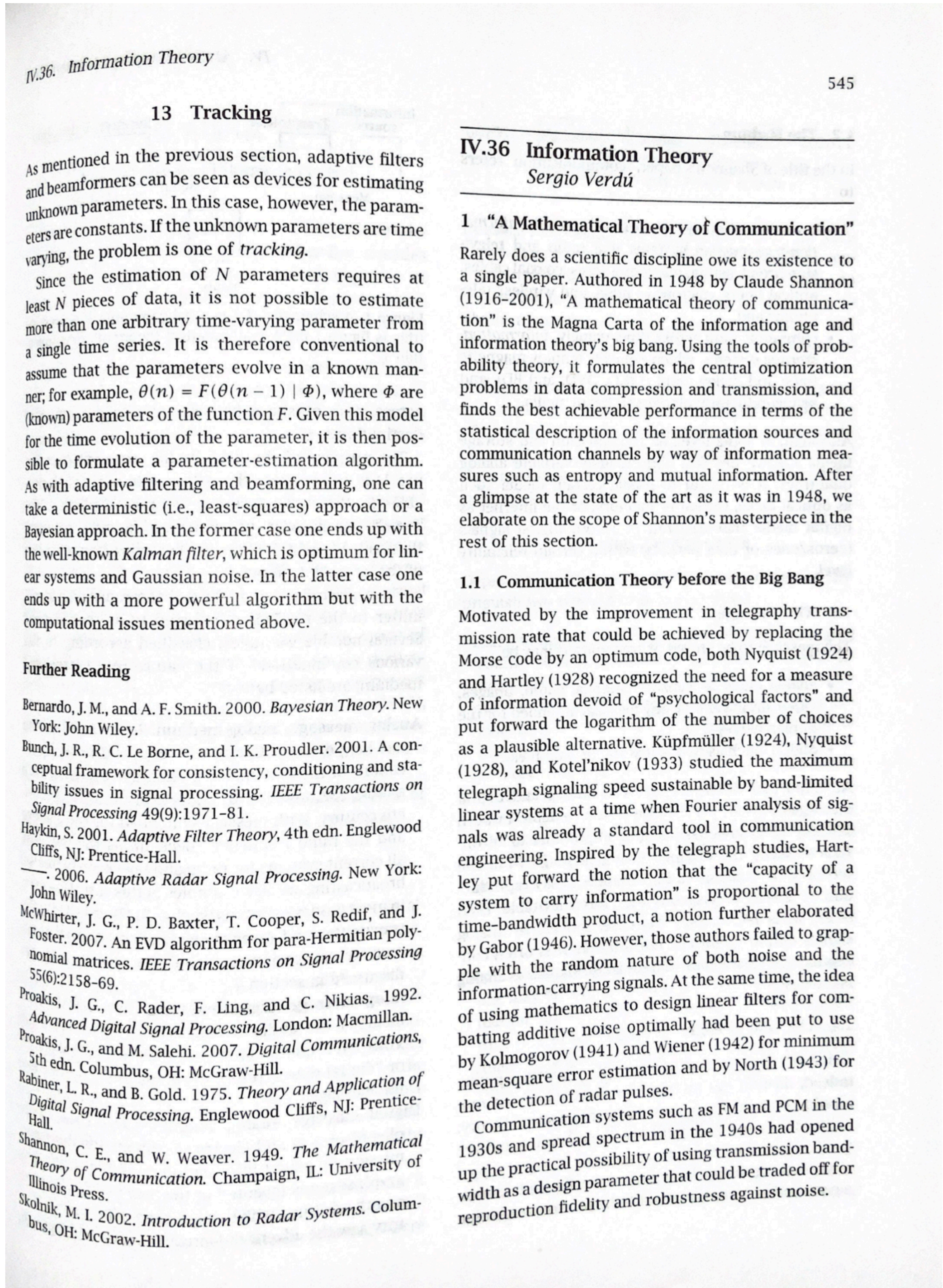
Topics in Information-Theoretic Cryptography

Lecture 1 - Classical Information Theory

Yanina Shkel, September 30, 2021

Today’s Reading

- A gentle overview of Information Theory as a branch of Applied Mathematics
- Introduces basic information measures:
  - Entropy
  - Relative Entropy
  - Mutual Information



13 Tracking

As mentioned in the previous section, adaptive filters and beamformers can be seen as devices for estimating unknown parameters. In this case, however, the parameters are constants. If the unknown parameters are time varying, the problem is one of *tracking*.

Since the estimation of N parameters requires at least N pieces of data, it is not possible to estimate more than one arbitrary time-varying parameter from a single time series. It is therefore conventional to assume that the parameters evolve in a known manner; for example, $\theta(n) = F(\theta(n-1) \mid \Phi)$, where Φ are (known) parameters of the function F. Given this model for the time evolution of the parameter, it is then possible to formulate a parameter-estimation algorithm. As with adaptive filtering and beamforming, one can take a deterministic (i.e., least-squares) approach or a Bayesian approach. In the former case one ends up with the well-known *Kalman filter*, which is optimum for linear systems and Gaussian noise. In the latter case one ends up with a more powerful algorithm but with the computational issues mentioned above.

Further Reading

Bernardo, J. M., and A. F. Smith. 2000. *Bayesian Theory*. New York: John Wiley.

Bunch, J. R., R. C. Le Borne, and I. K. Proudler. 2001. A conceptual framework for consistency, conditioning and stability issues in signal processing. *IEEE Transactions on Signal Processing* 49(9):1971–81.

Haykin, S. 2001. *Adaptive Filter Theory*, 4th edn. Englewood Cliffs, NJ: Prentice-Hall.

——. 2006. *Adaptive Radar Signal Processing*. New York: John Wiley.

McWhirter, J. G., P. D. Baxter, T. Cooper, S. Redif, and J. Foster. 2007. An EVD algorithm for para-Hermitian polynomial matrices. *IEEE Transactions on Signal Processing* 55(6):2158–69.

Proakis, J. G., C. Rader, F. Ling, and C. Nikias. 1992. *Advanced Digital Signal Processing*. London: Macmillan.

Proakis, J. G., and M. Salehi. 2007. *Digital Communications*, 5th edn. Columbus, OH: McGraw-Hill.

Rabiner, L. R., and B. Gold. 1975. *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Shannon, C. E., and W. Weaver. 1949. *The Mathematical Theory of Communication*. Champaign, IL: University of Illinois Press.

Skolnik, M. I. 2002. *Introduction to Radar Systems*. Columbus, OH: McGraw-Hill.

IV.36 Information Theory

Sergio Verdú

1 “A Mathematical Theory of Communication”

Rarely does a scientific discipline owe its existence to a single paper. Authored in 1948 by Claude Shannon (1916–2001), “A mathematical theory of communication” is the Magna Carta of the information age and information theory’s big bang. Using the tools of probability theory, it formulates the central optimization problems in data compression and transmission, and finds the best achievable performance in terms of the statistical description of the information sources and communication channels by way of information measures such as entropy and mutual information. After a glimpse at the state of the art as it was in 1948, we elaborate on the scope of Shannon’s masterpiece in the rest of this section.

1.1 Communication Theory before the Big Bang

Motivated by the improvement in telegraphy transmission rate that could be achieved by replacing the Morse code by an optimum code, both Nyquist (1924) and Hartley (1928) recognized the need for a measure of information devoid of “psychological factors” and put forward the logarithm of the number of choices as a plausible alternative. Küpfmüller (1924), Nyquist (1928), and Kotel’nikov (1933) studied the maximum telegraph signaling speed sustainable by band-limited linear systems at a time when Fourier analysis of signals was already a standard tool in communication engineering. Inspired by the telegraph studies, Hartley put forward the notion that the “capacity of a system to carry information” is proportional to the time–bandwidth product, a notion further elaborated by Gabor (1946). However, those authors failed to grapple with the random nature of both noise and the information-carrying signals. At the same time, the idea of using mathematics to design linear filters for combatting additive noise optimally had been put to use by Kolmogorov (1941) and Wiener (1942) for minimum mean-square error estimation and by North (1943) for the detection of radar pulses.

Communication systems such as FM and PCM in the 1930s and spread spectrum in the 1940s had opened up the practical possibility of using transmission bandwidth as a design parameter that could be traded off for reproduction fidelity and robustness against noise.

Today’s Reading

- A gentle overview of Information Theory as a branch of Applied Mathematics
- Introduces basic information measures:
  - Entropy
  - Relative Entropy
  - Mutual Information

of a coin with bias p can be compressed losslessly at any rate exceeding $h(p)$ bits per coin flip with

$$h(p) = p \log \frac{1}{p} + (1-p) \log \left(\frac{1}{1-p}\right),$$

which is the entropy of the biased coin source. The ubiquitous linear-time Lempel–Ziv universal data-compression algorithms are able to achieve, asymptotically, the entropy rate of ergodic stationary sources. Therefore, at least in the long run, universality incurs no penalty.

Relative entropy: a measure of the dissimilarity between two distributions P and Q defined on the same measurable space $(\mathcal{A}, \mathcal{F})$, defined as

$$D(P \| Q) = \int \log \left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P.$$

Relative entropy plays a central role not only in information theory but also in the analysis of the ability to discriminate between data models, and in particular in large-deviation results, which explore the exponential decrease (in the number of observations) of the probability of very unlikely events. Specifically, if n independent data samples are generated with probability distribution Q, the probability that they will appear to be generated from a distribution in some class $\mathcal{P}$ behaves as

$$\exp\left(-n \inf_{P \in \mathcal{P}} D(P \| Q)\right).$$

Relative entropy was introduced by Kullback and Leibler in 1951 with the primary goal of extending Shannon's measure of information to nondiscrete cases.

Mutual information: a measure of the dependence between two (not necessarily discrete) random variables X and Y given by the relative entropy between the joint measure and the product of the marginal measures:

$$I(X;Y) = D(P_{XY} \| P_X \times P_Y).$$

Note that $I(X;X) = H(X)$ if X is discrete.

For stationary channels that behave ergodically, the channel capacity is given by

$$C = \lim_{n \to \infty} \frac{1}{n} \max I(X_1, \dots, X_n; Y_1, \dots, Y_n),$$

where the maximum is over all joint distributions of $(X_1, \dots, X_n)$, and $(Y_1, \dots, Y_n)$ are the channel responses to $(X_1, \dots, X_n)$. If the channel is stationary memoryless, then the formula boils down to

$$C = \max I(X;Y).$$

The capacity of a channel that erases a fraction δ of the codeword symbols (drawn from an alphabet $\mathcal{A}$) is

$$C = (1-\delta) \log |\mathcal{A}|,$$

as long as the location of the erased symbols is known to the decoder and the nonerased symbols are received error free. In the case of a binary channel that introduces errors independently with probability δ, the capacity is given by

$$C = 1 - h(\delta),$$

while in the case of a continuous-time additive Gaussian noise channel with bandwidth B, transmission power P, and noise strength N, the capacity is

$$C = B \log \left(1 + \frac{P}{BN}\right) \text{ bits per second},$$

a formula that dispels the pre-1948 notion that the information-carrying capacity of a communication channel is proportional to its bandwidth and that is reminiscent of the fact that in a cellular phone the stronger the received signal the faster the download. In lossy data compression of a stationary ergodic source $(X_1, X_2, \dots)$, the rate compatible with a given per-sample distortion level d under a distortion measure $d \colon \mathcal{A}^2 \to [0, \infty]$ is given by

$$R(d) = \lim_{n \to \infty} \frac{1}{n} \min I(X_1, \dots, X_n; Y_1, \dots, Y_n),$$

where the minimum is taken over the joint distribution of source X^n and reproduction Y^n, with given P_{X^n}, and such that

$$\frac{1}{n} \sum_{i=1}^{n} d(X_i, Y_i) \leqslant d.$$

For stationary memoryless sources, just as for capacity we obtain a "single-letter" expression $R(d) = \min I(X;Y)$.

It should be emphasized that the central concern of information theory is not the definition of information measures but the theorems that use them to describe the fundamental limits of compression and transmission. However, it is rewarding that entropy, mutual information, and relative information, as well as other related measures, have found applications in many fields beyond communication theory, including probability theory, statistical inference, ergodic theory, computer science, physics, economics, life sciences, and linguistics.

ENTROPY

Entropy

Definition

Def$\underline{\text{n}}$: The entropy $H(X)$ of discrete random variable X is defined by

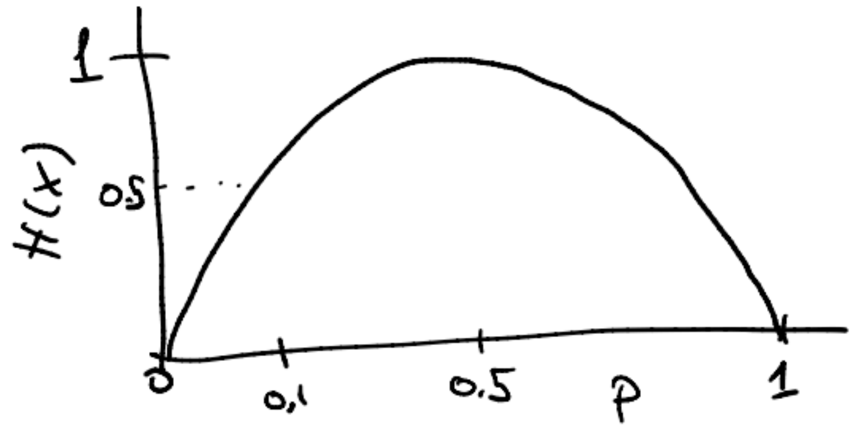
$$H(X) = \mathbb{E}_X\left[\log \frac{1}{P_X(X)}\right]$$

note, may also write $H(P_X)$.

Example: Let X be Bernoulli, i.e.

$$X = \begin{cases} 1 & \text{w/ prob } p. \\ 0 & \text{w/ prob } 1-p. \end{cases}$$

then $H(X) = -p \log p - (1-p)\log(1-p)$



Properties: $0 \le H(X) \le \log |\mathcal{X}|$

support of X

equality iff X is deterministic

equality iff X is equiprobable on its support

Entropy

Joint and Conditional Entropy

Definition: the joint entropy $H(X,Y)$ of a pair of discrete random variables (X,Y) w/ joint distribution P_{XY} is defined as

$$H(X,Y) = \mathbb{E}_{XY}\left[\log \frac{1}{P_{XY}(X,Y)}\right]$$

Definition: If $(X,Y) \sim P_{XY}$, the conditional entropy $H(Y|X)$ is defined as

$$H(Y|X) = \mathbb{E}_{XY}\left[\log \frac{1}{P_{Y|X}(Y|X)}\right]$$

need not be descrete

Entropy

Chain Rule

Properties: Chain Rule $H(X,Y) = H(X) + H(Y|X)$

$= H(Y) + H(X|Y)$

More generally,

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1)$$

jointly distr. w/ $P_{X_1 X_2 \ldots X_n}$

Entropy
Properties

Conditioning reduces entropy:

$$H(X|Y) \leq H(X)$$

equality iff X & Y are independent

However, $H(X|Y=y)$ may be greater than, or less than, or equal to $H(X)$

Functions: Let $f: \mathcal{X} \to \mathcal{Y}$ be a deterministic function,

$$H(f(X)) \leq H(X)$$

equal iff f is 1-1

RELATIVE ENTROPY

Relative Entropy (KL-Divergence)

Definition

Definition: The relative entropy between two probability distributions P_X and Q_X is defined as

$$D(P_X \| Q_X) = \mathbb{E}_{P_X}\left[\log \frac{P_X(x)}{Q_X(x)}\right]$$

$$= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{Q_X(x)}$$

Properties: $D(P_X \| Q_X) \geq 0$

Equality iff $P_X(x) = Q_X(x)$ for $x \in \mathcal{X}$

MUTUAL INFORMATION

Mutual Information

Definition

Definition: Consider two random variables X & Y, w/ joint distribution P_{XY}, & marginal distributions P_X & P_Y, then the mutual information is defined by

$$I(X;Y) = \mathbb{E}_{XY}\left[\log \frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)}\right]$$

Mutual Information
Properties

Properties: $I(X;Y) = I(Y;X)$

$$I(X;Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$

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

$$I(X;X) = H(X)$$

Finally,

$$0 \le I(X;Y) \le \min\{H(X), H(Y)\}$$

equality iff X & Y independent

equality iff $X = f(Y)$ (or $Y = f(X)$) for some deterministic func. f.

Mutual Information and Relative Entropy: $I(X;Y) = D(P_{XY} \| P_X P_Y)$

Mutual Information

Conditional Mutual Information

Definition: the conditional mutual information of X & Y given Z is defined by

$$I(X;Y|Z) = \mathbb{E}_{XYZ}\left[\log \frac{P_{XY|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}\right]$$

$$= H(X|Z) - H(X|Y,Z)$$

Chain rule: $I(X_1, X_2; Y) = I(X_1;Y) + I(X_2;Y|X_1)$

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \ldots, X_1)$$

Also, $I(X;Y|Z) \geq 0$

equality holds iff X & Y are independent given Z.

That is $X - Z - Y$ form a Markov chain.

Mutual Information

Data Processing Inequality (DPI)

Data Processing Inequality:

If $X - Y - Z$ form a Markov chain then

$$I(X;Y) \geq I(X;Z)$$

In general $\cancel{I(X;Y|Z) \leq I(X;Y)}$

If $X - Y - Z$ then $I(X;Y|Z) \leq I(X;Y)$

If $X - Z - Y$ the DPI hold w/ equality