Lecture 4. Classical entropy and information

Javier R. Fonollosa

Universitat Politècnica de Catalunya

javier.fonollosa@upc.edu

October 19, 2023

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona

Overview

Introduction

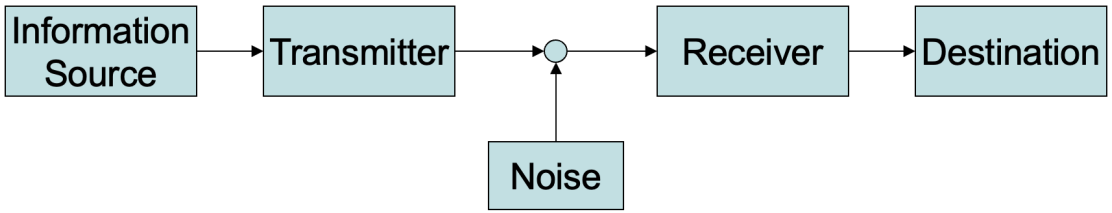
Introduction

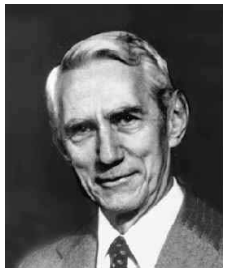
- In this Lecture we define *classical* entropy and mutual information providing a review of their basic properties and an interpretation as a measure of information and information transfer respectively.
- We also introduce some basic inequalities used in classical information theory, like Fano's inequality and the data processing inequality.
- We will finish by stating Shannon's fundamental theorem for point-to-point communication.

Shannon and the emergence of Information Theory (I)

- Information theory answers two fundamental questions in communications:
  - What is the ultimate data compression? (The entropy H)
  - What is the ultimate transmission rate of communication? (Channel capacity C).
- Increasing the transmission rate *does not imply* increasing error rate. Asymptotically error free communication *is possible*.
- Shannon. "The Mathematical Theory of Communication. 1948" invents the source-encoder-channel-decoder-destination model and gives a general solution: $H < R < C$, where R is the transmission Rate.

Shannon and the emergence of Information Theory (and II)





Claude Shannon
1916-2001

- Shannon proved that:
  - We can obtain reliable transmission of information at any rate below channel capacity (C).
  - This capacity depends on the noise characteristics of the channel.
  - Random processes have a irreducible compression rate, that he named entropy (H).
- Asymptotically error free communication can be achieved if $H < C$.
- Shannon introduced the notion of entropy H of a random variable, the mutual information I, between two random variables and a calculus that relates these quantities.

Entropy

Entropy

Entropy

Let X be a discrete random variable with pmf $p(x)$, the information revealed once X is disclosed is measured by the entropy:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = -\mathsf{E}_X(\log p(X)) = \mathsf{E}_X\left(\log \frac{1}{p(X)}\right)$$

For X a Bernoulli random variable with $p = \Pr\{X = 1\}$, i.e, $X \sim \mathrm{Bern}(p)$,

$$H(X) = -p \log p - (1-p) \log(1-p) = -p \log p - \bar{p} \log \bar{p} \triangleq H(p)$$

The entropy is a non negative and concave function in $\frac{1}{p(X)}$ which means:

$$H(X) = \mathsf{E}_X\left(\log \frac{1}{p(X)}\right) \leq \log \mathsf{E}_X\left(\frac{1}{p(X)}\right) = \log \sum_{x \in \mathcal{X}} \frac{p(x)}{p(x)} = \log |\mathcal{X}|$$

by Jensen's inequality.

Conditional entropy

Conditional entropy

Let X and Y be discrete random variables and $Y|\{X = x\} \sim p(y|x)$, the **conditional entropy** $H(Y|X)$ (uncertainty that remains in Y once X is revealed) is:

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} H(Y|X = x)p(x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(y|x) = -\mathsf{E}_{X,Y}(\log p(Y|X)). \end{aligned}$$

Also, by Jensen's inequality:

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

with **equality** if X and Y are independent.

Joint entropy

The definition of entropy can be extended to any pair of discrete random variables.

Joint entropy

Let $(X, Y) \sim p(x, y)$, the joint entropy is defined as the entropy of the combined random variable:

$$H(X, Y) = -\mathrm{E}_{X,Y}(\log p(X, Y))$$

The relations $p(x, y) = p(x)p(y|x) = p(y)p(x|y)$ imply:

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

which means, using $H(X|Y) \leq H(X)$, that:

$$H(X, Y) \leq H(X) + H(Y)$$

with equality if X and Y are independent.

Chain rule of entropy

The previous definitions admit a generalization for an arbitrary number of discrete random variables or discrete random vectors.

Chain rule of entropy

Let $X^n \sim p(x^n)$:

$$\begin{aligned} H(X^n) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \cdots, X_{n-1}) \\ &= \sum_{i=1}^{n} H(X_i|X_1, \cdots, X_{i-1}) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \end{aligned}$$

Using induction one can show that:

$$H(X^n) \leq \sum_{i=1}^{n} H(X_i)$$

with equality if all X_i are mutually independent.

Fano's inequality

Fano's inequality (I)

Fano's inequality

Let $(X, Y) \sim p(x, y)$ and $P_e = \Pr\{X \neq Y\}$. Then:

$$H(X|Y) \leq H(P_e) + P_e \log(|\mathcal{X}| - 1) \leq 1 + P_e \log |\mathcal{X}|$$

Robert Fano
1917-2016

Proof: Define the binary random variable E as 0 if $\{X = Y\}$ and 1 otherwise. Then $\mathrm{E}(E) = \Pr\{X \neq Y\} = P_e$ and :

$$H(E) = -P_e \log P_e - (1 - P_e) \log(1 - P_e) = H(P_e)$$

By the properties of joint and conditional entropy:

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

Fano's inequality (and II)

From the previous slide:

$$H(X|Y) = -H(E|X,Y) + H(E|Y) + H(X|E,Y)$$

But:

$$\begin{aligned} H(E|X,Y) &= 0 \\ H(E|Y) &\leq H(E) = H(P_e) \\ H(X|E,Y) &= P_e H(X|E=1,Y) + (1-P_e)H(X|E=0,Y) \\ &\leq P_e \log(|\mathcal{X}|-1) \leq P_e \log|\mathcal{X}| \end{aligned}$$

Thus

$$H(X|Y) \leq H(P_e) + P_e \log|\mathcal{X}| \leq 1 + P_e \log|\mathcal{X}| \quad \blacksquare$$

Mutual Information

Definition and basic properties

Mutual Information for discrete random variables

Let $(X, Y) \sim p(x, y)$, be a pair of discrete random variables. The information about X obtained from Y, defined as mutual information is:

$$I(X;Y) = \mathsf{E}_{X,Y}\left(\log \frac{p(x,y)}{p(x)p(y)}\right) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

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

- $I(X;Y)$ is concave in $p(x)$ for fixed $p(y|x)$ and convex in $p(y|x)$ for fixed $p(x)$. Proof, see Theorem 2.7.4 in [Cover, 2006].
- Also $I(X;Y) \geq 0$, with equality iff X and Y are independent.

Relative entropy

The mutual information is a special case of the relative entropy:

Relative entropy (Kullback-Leibler divergence)

Let $p(x)$ and $q(x)$ be two pmf defined over $x \in \mathcal{X}$,

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathsf{E}_{p(x)} \left(\log \frac{p(x)}{q(x)} \right) \geq 0.$$

If $p(x) > 0$ and $q(x) = 0$ for any $x \in \mathcal{X}$, then $D(p||q) = \infty$. Positivity follows from Jensen's inequality.

$$I(X;Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y)) \geq 0.$$

with equality iff X and Y are independent.

Conditional mutual information: definition

Conditional mutual information

Let $(X, Y)|\{Z = z\} \sim p(x, y|z)$, $Z \sim p(z)$, the mutual information between X and Y, given $\{Z = z\}$ is $I(X; Y|\{Z = z\})$ and the conditional mutual information between X and Y, given Z is defined as:

$$\begin{aligned} I(X; Y|Z) &= \sum_{z \in \mathcal{Z}} I(X; Y|Z = z)p(z) \\ &= H(X|Z) - H(X|Y, Z) \\ &= H(Y|Z) - H(Y|X, Z) \\ &= H(X|Z) + H(Y|Z) - H(X, Y|Z) \geq 0 \end{aligned}$$

Like the unconditional mutual information, the conditional mutual information between X and Y, given Z, $I(X; Y|Z)$ is nonnegative.

Conditional mutual information and Markov chains

Markov Chain

Random variables X, Y and Z (the order **is** important) form a Markov chain, denoted by $X \leftrightarrow Y \leftrightarrow Z$ iff:

$$p(z|x,y) = p(z|y), \text{which also implies } p(x,z|y) = p(x|y)p(z|y),$$

that is, iff X and Z are conditionally independent given Y, $I(X;Z|Y) = 0$.

Zero Conditional mutual information

The conditional mutual information between X and Z, given Y, $I(X;Z|Y)$ is equal to zero iff $X \leftrightarrow Y \leftrightarrow Z$ form a Markov chain.

$$I(X;Z|Y) = 0 \Leftrightarrow X \leftrightarrow Y \leftrightarrow Z$$

Conditional mutual information: Chain rule

Note that unlike the conditional entropy where $H(X,Y|Z) \leq H(X,Y)$, $I(X;Y|Z)$ is not necessarily smaller or greater than $I(X;Y)$, except for two particular cases:

- Independent X and Z, $p(x,y,z) = p(x)p(z)p(y|x,z)$,

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

- Conditional independence of Z and Y, given X, i.e., $Z \leftrightarrow X \leftrightarrow Y$ form a Markov chain, then:

$$I(X;Y|Z) \leq I(X;Y)$$

Chain Rule and data processing inequality

Chain rule for mutual information

Let $(X^n, Y) \sim F(x^n, y)$ then,

$$I(X^n; Y) = \sum_{i=1}^{n} I(X_i; Y|X^{i-1})$$

Data processing inequality

If $X \leftrightarrow Y \leftrightarrow Z$ form a Markov chain, then

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

which can be proved by the relations:

$$\begin{aligned} I(X; Y, Z) &= I(X; Y) + I(X; Z|Y) = I(X; Y) \\ &= I(X; Z) + I(X; Y|Z) \geq I(X; Z) \end{aligned}$$

Markov's inequality and the (weak) Law of Large Numbers (LLN)

Markov's inequality

Markov's inequality

For any non-negative random variable t and any positive a,

$$\mathsf{Pr}\{t \geq a\} \leq \frac{\mathsf{E}(t)}{a}.$$

This inequality is used in bounding the tails of a distribution.

Proof: Let $\mathbb{1}\{\cdot\}$ be the indicator function which is 1 if the argument is true and zero otherwise.

$$\mathsf{Pr}\{t \geq a\} = \mathsf{E}\left(\mathbb{1}\{t \geq a\}\right) \leq \mathsf{E}\left(\frac{t}{a}\right) = \frac{\mathsf{E}(t)}{a}$$

■

The weak law of large numbers (LLN)

Weak law of large numbers

Let $X_1, X_2, ...$ be a sequence of i.i.d. random variables with finite mean $\mathrm{E}(X)$ and variance $\mathrm{Var}(X)$, then for every $\epsilon > 0$,

$$\mathrm{P}\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}(X)| \geq \epsilon\} \leq \frac{\mathrm{Var}(X)}{n\epsilon^2},$$

$$\lim_{n\to\infty} \mathrm{P}\{|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}(X)| \geq \epsilon\} = 0.$$

which means that $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to $\mathrm{E}(X)$ in probability.

Proof:

Note that $|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}(X)| \geq \epsilon \iff |\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}(X)|^2 \geq \epsilon^2$. Take $t \equiv |\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}(X)|^2$ and apply Markov's inequality.

Classical communication theorem

Channel coding problem statement

Consider a discrete memoryless channel (DMC) defined by $(\mathcal{X}, p(y|x), \mathcal{Y})$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets and $p(y|x)$ are conditional pmfs.

Channel code

A $(2^{nR}, n)$ code $\mathcal{C}$ for the DMC $p(y|x)$ is:

- a message set $[1 : 2^{nR}]$.
- an encoding function $x^n : [1 : 2^{nR}] \to \mathcal{X}^n$ that assigns a codeword $x^n(m)$ to each message $m \in [1 : 2^{nR}]$,
- a decoding function $\hat{m} : \mathcal{Y}^n \to [1 : 2^{nR}] \cup \{e\}$ that assigns an estimate $\hat{m} \in [1 : 2^{nR}]$ or an error message e to each received sequence y^n. Note that the previous definition and the memoryless condition on the channel implies:

$$p_{Y^n|X^n,M}(y^n|x^n, m) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$$

Capacity definition

The performance of a code is measured by its error probability, i.e., let $\lambda_m(\mathcal{C}) = \mathsf{Pr}\{\hat{M} \neq m | M = m\}$ be the conditional probability of error of codebook $\mathcal{C}$ assuming message m is sent:

$$\mathsf{P}_e^{(n)}(\mathcal{C}) = \mathsf{Pr}\{\hat{M} \neq M\} = \frac{1}{2^{\lceil nR \rceil}} \sum_{m=1}^{2^{\lceil nR \rceil}} \lambda_m(\mathcal{C})$$

Capacity definition

A rate R is said to be achievable if there exists a sequence of $(2^{nR}, n)$ codes such that

$$\lim_{n \to \infty} \mathsf{P}_e^{(n)}(\mathcal{C}) = 0$$

The capacity is the supremum over all achievable rates.

We will assume 2^{nR} is integer.

Channel coding theorem

Channel coding theorem

The capacity of a discrete memoryless channel $p(y|x)$ is the information capacity:

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

Capacity of the BSC

Consider a Binary Symmetric channel with crossover probability p,BSC(p), i.e, binary channel input X and output Y and binary symbols are flipped with probability p. This channel is equivalent to consider $Y = X \oplus Z$ with $Z \sim \text{Bern}(p)$ independent of X. The capacity is:

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(Y) - H(Y|X)) \\ &= \max_{p(x)} (H(Y) - H(X \oplus Z|X)) \\ &= \max_{p(x)} (H(Y) - H(Z|X)) \\ &= \max_{p(x)} (H(Y) - H(Z)) \\ &= 1 - H(p), \end{aligned}$$

attained for $X \sim \text{Bern}(1/2)$ which implies $Y \sim \text{Bern}(1/2)$.

Capacity of the BEC

Consider a Binary Erasure channel with erasure probability p, BEC(p), i.e, binary channel input X and ternary output Y where binary symbols are mapped into a erasure symbol e with probability p. The capacity is:

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) \\ &= \max_{p(x)} (H(X) - H(X|Y)) \\ &= \max_{p(x)} (H(X) - pH(X)) \\ &= 1 - p. \end{aligned}$$

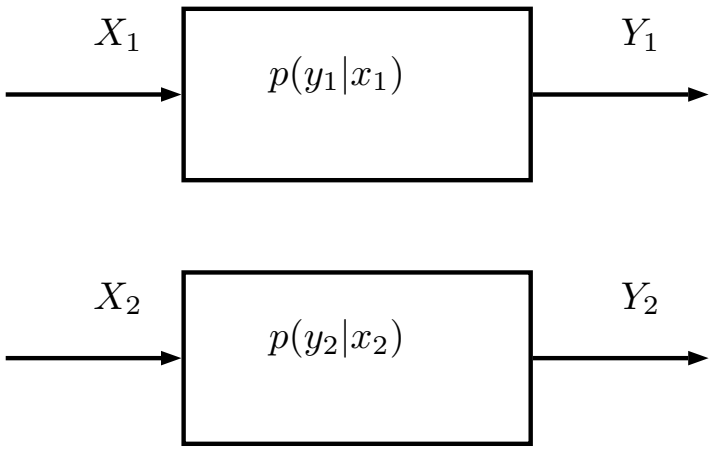
attained for $X \sim \mathrm{Bern}(1/2)$. Note that

$$\begin{aligned} H(X|Y) &= \mathsf{Pr}(Y=0)H(X|Y=0) + \mathsf{Pr}(Y=1)H(X|Y=1) \\ &\quad + \mathsf{Pr}(Y=e)H(X|Y=e) \\ &= \mathsf{Pr}(Y=e)H(X|Y=e) = pH(X) \end{aligned}$$

Product DMC

Consider two parallel DMCs, $p(y_1, y_2|x_1, x_2) = p(y_1|x_1)p(y_2|x_2)$ each of which is characterized by a given capacity,

$$C_1 = \max_{p(x_1)} I(X_1; Y_1) \text{ and } C_2 = \max_{p(x_2)} I(X_2; Y_2).$$



- $X_1 \leftrightarrow X_2 \leftrightarrow Y_2$ form a Markov chain since

$$\begin{aligned} p(x_1, x_2, y_2) &= p(x_1)p(x_2|x_1)p(y_2|x_1, x_2) \\ &= p(x_1)p(x_2|x_1)p(y_2|x_2) \end{aligned}$$

- Similarly $X_2 \leftrightarrow X_1 \leftrightarrow Y_1$
- Therefore $Y_1 \leftrightarrow X_1 \leftrightarrow X_2 \leftrightarrow Y_2$

Product DMC

Additivity of the capacity of the product DMC

The capacity of the product DMC is at least the sum of the capacities, $C \geq C_1 + C_2$ since independent coding is a particular case of joint coding. In order to proof that it is in fact *equal* to the sum we consider:

$$C = \max_{p(x_1,x_2)} I(X_1, X_2; Y_1, Y_2)$$

where, since $Y_1 \leftrightarrow X_1 \leftrightarrow X_2 \leftrightarrow Y_2$ form a Markov chain:

$$\begin{aligned} I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \\ &= H(Y_1, Y_2) - H(Y_1|X_1, X_2) - H(Y_2|Y_1, X_1, X_2) \\ &= H(Y_1, Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\ &\leq H(Y_1) + H(Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\ &= I(X_1; Y_1) + I(X_2; Y_2) \end{aligned}$$

Therefore $C \leq C_1 + C_2$ completing the proof. ■

Codebook generation

Achievability

For every rate $R < C = \max_{p(x)} I(X;Y)$ there exists a sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} \mathsf{P}_e^{(n)}(\mathcal{C}) = 0$

Random codebook generation

Fix the pmf $p(x)$ that attains the capacity. Randomly and independently generate $x^n(m)$ for $m \in [1 : 2^{nR}]$ according to $p(x^n) = \prod_{i=1}^n p_X(x_i)$. The generated sequences constitute the codebook. The random coding generation implies:

$$p(\mathcal{C}) = \prod_{m=1}^{2^{nR}} \prod_{i=1}^{n} p_X(x_i(m))$$

The codebook is revealed to the encoder and to the decoder prior to transmission.

Encoding and decoding

Encoding

To send message $m \in [1 : 2^{nR}]$ transmit $x^n(m)$.

Decoding

Use maximum likelihood decoding, that is, let y^n the received sequence, the receiver declares that $\hat{m} \in [1 : 2^{nR}]$ was transmitted if $\hat{m}$ is the message such that $x^n(\hat{m})$ maximizes $p(y^n|x^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$ (breaking ties arbitrarily).

Probability of error

Assuming message m was transmitted, the receiver makes an error if there is another message $m' \neq m$ such that $p(y^n|x^n(m')) \geq p(y^n|x^n(m))$. The probability of error averaged over all possible messages m and random codebooks $\mathcal{C}$ is:

$$\begin{aligned}
\Pr(\mathcal{E}) &= \mathrm{E}_{\mathcal{C}}(\mathrm{P}_e^{(n)}) \\
&= \mathrm{E}_{\mathcal{C}}\left(\frac{1}{2^{nR}}\sum_{m=1}^{2^{nR}}\lambda_m(\mathcal{C})\right) \\
&= \frac{1}{2^{nR}}\sum_{m=1}^{2^{nR}}\mathrm{E}_{\mathcal{C}}(\lambda_m(\mathcal{C})) \\
&= \mathrm{E}_{\mathcal{C}}(\lambda_1(\mathcal{C})) = \Pr(\mathcal{E}|M=1).
\end{aligned}$$

Pairwise error probability

Pairwise error probability

Now, fix $x^n(1)$ and y^n and denote by $\mathsf{Pr}(\mathcal{E}_m|M=1)$ the pairwise error probability. The probability of $m \neq 1$ being declared when $x^n(1)$ was transmitted is:

$$\mathsf{Pr}\{\mathcal{E}_m|M=1\} = \mathsf{Pr}\{p(y^n|x^n(m)) \geq p(y^n|x^n(1))\}.$$

Applying Markov's inequality:

$$\begin{aligned}
\mathsf{Pr}\{\mathcal{E}_m|M=1\} &\leq \frac{\mathsf{E}_{X^n}\left(p(y^n|x^n(m))\right)}{p(y^n|x^n(1))} \\
&= \frac{\sum_{x^n(m)\in\mathcal{X}^n} p(y^n|x^n(m))p(x^n(m))}{p(y^n|x^n(1))} \\
&= \frac{p(y^n)}{p(y^n|x^n(1))}, \text{ which is independent of } m.
\end{aligned}$$

Union bound

Union bound

By the union bound we get:

$$\mathsf{Pr}(\mathcal{E}|M=1) \leq \mathsf{Pr}\Big\{ \bigcup_{m=2}^{2^{nR}} \mathcal{E}_m | M=1 \Big\} \leq 2^{nR} \frac{p(y^n)}{p(y^n|x^n(1))}$$

Since the channel is memoryless, $p(y^n|x^n) = \prod_{i=1}^n p_{Y|X}(y_i|x_i)$ and $p(y^n) = \prod_{i=1}^n p_Y(y_i)$, i.e.

$$\frac{1}{n} \log \left(\frac{p(y^n)}{p(y^n|x^n(1))} \right) = \frac{1}{n} \sum_{i=1}^n \log \left(\frac{p_Y(y_i)}{p_{Y|X}(y_i|x_i(1))} \right), \text{ and by the LLN}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \log \left(\frac{p_Y(y_i)}{p_{Y|X}(y_i|x_i(1))} \right) = \mathsf{E}_{X,Y} \left(\log \frac{p_Y(y)}{p_{Y|X}(y|x)} \right) = -I(X;Y)$$

Probability of error

The previous result implies that for large enough n we have:

$$\frac{1}{n}\log\left(\frac{p(y^n)}{p(y^n|x^n(1))}\right) \leq -I(X;Y)+\delta$$

Probability of error

Introducing the last result back into the error probability expression we get:

$$\Pr(\mathcal{E}|M=1) \leq 2^{nR}2^{-n(I(X;Y)-\delta)} = 2^{-n(I(X;Y)-R-\delta)} = 2^{-n(C-R-\delta)},$$

which can be made arbitrarily small if $R < C = \max_{p(x)} I(X;Y)$, since δ can be arbitrarily small. ■

Proof of the converse (I)

Converse

For every sequence of $(2^{nR}, n)$ codes such that $\lim_{n\to\infty} \mathsf{P}_e^{(n)}(\mathcal{C}) = 0$, we must have $R < C = \max_{p(x)} I(X;Y)$.

Every $(2^{nR}, n)$ code induces a pmf on (M, X^n, Y^n)

$$p(m, x^n, y^n) = p(m)p(x^n|m)p(y^n|m, x^n) = 2^{-nR}p(x^n|m)\prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$$

By Fano's inequality:

$$H(M|\hat{M}) \leq 1 + \mathsf{P}_e \log |\mathcal{M}| = 1 + \mathsf{P}_e nR = n(1/n + \mathsf{P}_e R) = n\epsilon_n$$

By assumption $\lim_{n\to\infty} \mathsf{P}_e^{(n)} = 0$ which implies $\epsilon_n \to 0$.

Proof of the converse (and II)

We know that $M \leftrightarrow X^n \leftrightarrow Y^n \leftrightarrow \hat{M}$ form a Markov chain and $H(M|\hat{M}) \leq n\epsilon_n$. Now:

$$\begin{aligned} nR &= H(M) \\ &= I(M;\hat{M}) + H(M|\hat{M}) \\ &\leq I(M;\hat{M}) + n\epsilon_n \\ &\leq^{(a)} I(X^n;Y^n) + n\epsilon_n \\ &\leq^{(b)} nC + n\epsilon_n \end{aligned}$$

■

Where:

- (a) follows from the data processing inequality.
- (b) follows from the additivity of the capacity of the product DMC.

Remark

DMC with feedback

In a DMC with noiseless causal feedback, the encoder assigns a symbol $x_i(m, y^{i-1})$ to each message $m \in [1 : 2^{nR}]$ and past received sequence $y^{i-1} \in \mathcal{Y}^{i-1}$ for i $[1 : n]$, which means that every $(2^{nR}, n)$ code induces a pmf on (M, X^n, Y^n) as:

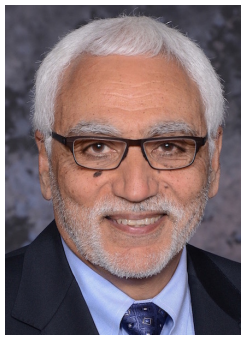
$$p(m, x^n, y^n) = 2^{-nR} \prod_{i=1}^{n} p(x_i|m, y^{i-1}) p_{X|Y}(y_i|x_i)$$

Nevertheless, it is easy to check that the inequalities that where used to prove the converse also hold (the additivity of the capacity of the product DMC) now and therefore the capacity of the DMC is not increased by feedback . . . although feedback can be very useful from an operational perspective!, e.g, in the BEC.

References

References

Thomas Cover 1938-2012



Abbas El Gamal 1950

Thomas Cover and Joy A. Thomas.

Elements of Information Theory, Second Edition.

Willey 2006.

Abbas El Gamal and Young-Han Kim.

Network Information Theory

Cambridge University Press 2011.