Brief Summary on Information Theory Basics

Petros Elia

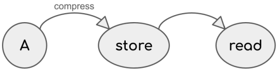
EURECOM Fall 2024

November 19, 2024

1 Outline

- Two main impactful applications
  - Source compression
  - Communications
- Also applies in
  - Distributed computing
  - Learning
  - Investment theory
  - Statistical physics

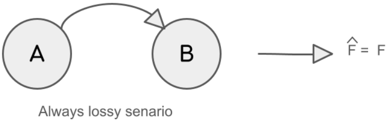
For compression:



- Lossless or Lossy compression
- Q: What is the minimum amount of space needed?

For communications:

- Q: How fast can we transmit such that $\hat{F} = F$?



- Depending on the physics, the fundamental limits change.
- Basic channel, BSC is defined by the probability of transition p.
- Multi-user scenario: both in compression and in communications

- Random variable (R.V.) $X =$ "Volleyball Winner at Paris Olympics "
- Alphabet $\mathcal{X} = \{$Argentina, France, Italy, Brazil$\}$
  - with probability p_1, p_2, p_3, p_4.
  - Random event $A \rightarrow p$ probability
- What is the <u>Information content</u> of an event (properties to be considered)

$$
\begin{array}{ll}
I(p=1)=0 & I(p\downarrow)\uparrow \\
I(p)\geq 0 & I(p\rightarrow 0)\rightarrow\infty
\end{array}
$$

- Also, if A, B independent; we should have $\rightarrow I(A,B) = I(A) + I(B)$
  - Also nicely reflecting $p(A,B) = p_A p_B$
- Shannon information:
  - $I(p) = -\log_2(p)$ in bits (assume this)
  - $I(p) = -\log_e(p)$ in nats
- Ex: Rain $= 99\%$, Sunny $= 1\%$.
  - $I(\text{'Today is raining'}) = -\log(0.99) = 0.014$ *bits*
  - $I(\text{'Today is sunny'}) = -\log(0.99) = 6.64$ *bits*

EURECOM
Sophia Antipolis

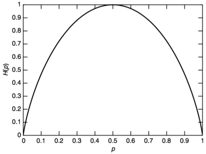
- Definition 1: **Entropy**

$$H(X) = -\sum_{i=1}^{|\mathcal{X}|} p_i log(p_i) \geq 0 \quad \text{(average information/\underline{UNCERTAINTY})}$$

  - Alphabet cardinality $|\mathcal{X}|$
- **Binary Entropy** for binary X (e.g. Rain vs. Sunny)
  - $\mathcal{X} = \{0, 1\}$: $p(X=0) = p$, $p(X=1) = 1-p$:

$$H(X) = -plogp - (1-p)log(1-p)$$

  - Maximized at $p = 0.5$.



2 Entropy - Example

Calculate: $H(X) = -\sum_{i}^{|\mathcal{X}|} p_i \log(p_i) \geq 0$

- For $\mathcal{X} = \{(X^{(1)}, \frac{1}{2}), (X^{(2)}, \frac{1}{4}), (X^{(3)}, \frac{1}{8}), (X^{(4)}, \frac{1}{8})\}$
  - {Volleyball: Argentina, France, Italy, Brazil}

$$
\begin{aligned}
H(X) &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{8}\log_2\frac{1}{8} \\
&= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 2^2 + \frac{2}{8}\log_2 2^3 \\
&= \frac{1}{2} + \frac{2}{4} + \frac{2}{8}\times 3 = \frac{7}{4}\ \textit{bits}
\end{aligned}
$$

- Vs. 100m Sprint:
  $\mathcal{X} = \{(X^{(1)}, \frac{9999}{10000}), (X^{(2)}, \frac{1}{30000}), (X^{(3)}, \frac{1}{30000}), (X^{(4)}, \frac{1}{30000})\}$
  - {Usain Bolt, Me, You, Obama}

$$
H(X) = -\frac{9999}{10000}\log_2\frac{9999}{10000} - \frac{3}{30000}\log_2\frac{1}{30000} = 0.0016\ \textit{bits}
$$

Definition 2: **Joint Entropy** of two R.Vs. X, Y

- Let X, Y be two discrete R.V (discrete alphabets)
- $X \in \mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots\}, \quad Y \in \mathcal{Y} = \{y^{(1)}, y^{(2)}, \ldots\}$
- $p(X), p(Y), p(X, Y)$

- Consider $Z = (X, Y), \quad Z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
- Thus

$$H(X, Y) = -E[\log p(X, Y)] = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y)$$

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

EURECOM Sophia Antipolis

2 Conditional Entropy

Definition 3: **Conditional Entropy:** $H(X|Y)$

$$\begin{aligned} H(X|Y) &\triangleq E_y[H(X|Y=y)] \\ &= -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x|Y=y)p(y)\log p(x|Y=y) \\ &\overset{(BR)}{=} -\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log\frac{p(x,y)}{p(y)} \end{aligned}$$

- Naturally **Conditioning Reduces Entropy**

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

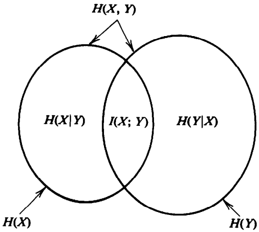
  - with equality if X and Y are independent.

1

[1] Bayer Rule (BR): $p(X,Y) = p(X)p(Y|X) = p(Y)p(X|Y)$

Definition 4: **Mutual information:**

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



- $I(X;Y) = 0$ when X and Y are independent.

Definition 5: **Chain Rule:**

- X, Y pair of discrete R.V
- Then

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

- Example

$$H(X_1, X_2, ..., X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + ... + H(X_n|X_1, X_2, ..., X_{n-1})$$

2 Chain Rule and Conditional Mutual Information

- We have two chain rules
  - Recall: Entropy chain rule

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

  - Information chain rule

$$I((X,Y);Z) = I(X;Z) + \underbrace{I(Y;Z|X)}_{*}$$

- ***Conditional mutual information**

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

Definition 6: **Relative entropy** (Kullback-Leibler distance/divergence)

- Given RV X and two distributions $p(X)$ and $q(X)$, then what is the distance between them?

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \geq 0$$

  > Takes form of Informational Distance $D(p||q) = E_{p(X)}[\log p(X) - \log q(X)]$

- Can measure the degree of independence between X and Y, i.e:

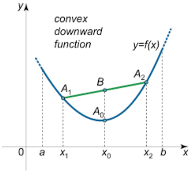
$$D(p(X,Y)||p(X)P(Y)) Thusmostimportantly$$

D(p(X,Y)——p(X)p(Y)) = I(X;Y)

2 Convexity and Jensen's inequality

- Convexity: For $x = px_1 + (1-p)x_2, p \in \mathbb{R}^+$, then

$$f(x) \leq pf(x_1) + (1-p)f(x_2)$$



- X: discrete random variable
- f: convex function

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

- f: concave function

$$E[f(X)] \leq f(E[X])$$

2 Maximal Entropy

- X is a discrete RV $x \in \mathcal{X}$
  - with $|\mathcal{X}|$ elements, and distribution p.
- Then

$$H(X) \leq \log |\mathcal{X}|$$

  - with equality if

$$p(x) = \frac{1}{|\mathcal{X}|}, \ \forall x \in \mathcal{X} : \quad \text{(uniform distribution)}$$

  - uniform is scenario with highest uncertainty.
- Proof Sketch: (any p, q uniform). Use[2] that $D(p||q) \geq 0$.

$$D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)} = \sum p(x) \log(p(x)|\mathcal{X}|)$$
$$= \sum p(x) \log p(x) + \sum p(x) \log |\mathcal{X}| = -H(X) + \log |\mathcal{X}| \geq 0$$

[2]From Log-Sum Ineq: $\sum_{i=1}^{|\mathcal{X}|} p_i \log \frac{p_i}{q_i} \geq (\sum_{i=1}^{|\mathcal{X}|} p_i) \log\big((\sum_{i=1}^{|\mathcal{X}|} p_i)/(\sum_{i=1}^{|\mathcal{X}|} q_i)\big)$.

- X is a R.V.

- Then

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

- You cannot increase the information you get for Z from X, by "massaging" X

2 Markov chains (MC): (X, Y, Z)

- (X, Y, Z) form a MC if and only if
  - $p(Z|X, Y) \overset{(mc1)}{=} p(Z|Y)$
  - $p(X, Y, Z) \overset{(mc2)}{=} p(X)p(Y|X)p(Z|Y)$
  - (mc2): Apply chain rule and then (mc1)
- Meaning: dependency of z from x happens through y only
- (mc1) and (mc2) are equivalent
  - Sketch of Alternate Proof that (mc1) $\rightarrow$ (mc2)

$$P(X, Y, Z) \overset{(br)}{=} P(Z|X, Y)P(X, Y)$$

$$\overset{(mc1)}{=} P(Z|Y)P(X, Y) \overset{(br)}{=} P(Z|Y)P(Y|X)P(X) \rightarrow \ (mc2)$$

2 Proof of DPI using Markov chains

- Want to prove that $I(X,Z) \leq I(X,Y)$
- Consider Information Chain Rule (icr)

$$
\begin{aligned}
I(X;Y,Z) &\overset{(icr)}{=} I(X;Y) + I(X;Z|Y) \\
&\overset{(icr)}{=} I(X;Z) + I(X;Y|Z) \\
&\Rightarrow I(X;Y) + I(X;Z|Y) = I(X;Z) + I(X;Y|Z) \quad (**)
\end{aligned}
$$

- Now note that:

$$P(X,Z|Y) \overset{(br)}{=} \frac{P(X,Y,Z)}{P(Y)} \overset{(br)}{=} \frac{P(Z|Y,X)P(X,Y)}{P(Y)} \overset{(mc+br)}{=} P(Z|Y)P(X|Y)$$

- Thus, given Y, then X and Z are independent
- Thus $I(X;Z|Y) = 0$
- Thus from $(**)$, we have

$$I(X;Y) = I(X;Z) + I(X;Y|Z)$$

- Thus $I(X,Z) \leq I(X,Y)$. (DPI proved)

2 Fano's inequality

- X: unknown R.V.
- We observe Y, correlated with X via $P(Y|X)$
- We want to build an estimate of X using Y
  - build $\hat{X} = g(Y)$
- Prob of error: P_e=Prob$(\hat{X} \neq X)$
- Fano bound: $P_e \geq$?

EURECOM
Sophia Antipolis

- Let E be a binary R.V. $E(error\ event) = \begin{cases} 0 & \hat{X} = X \quad 1 - P_e \\ 1 & \hat{X} \neq X \quad P_e \end{cases}$
- $H(E, X|Y) \overset{(cr)}{=} H(X|Y) + H(E|X, Y) = H(X|Y) \quad (*)$
  - Since $\rightarrow H(E|X, Y) = 0$, since E is fully determined by X and Y
- Rewrite same

$$H(E, X|Y) \overset{(cr)}{=} H(E|Y) + H(X|E, Y) \quad (**)$$

$$H(X|E, Y) = H(X|Y) - H(E|Y) \overset{(cre)}{\geq} H(X|Y) - H(E) = H(X|Y) - H(P_e) \ (*\ast$$

$$H(X|E, Y) = (1-P_e)H(X|Y, E = 0) + P_e H(X|Y, E = 1) \overset{(maxH)}{\leq} P_e \log(|\mathcal{X}|-1)$$

3 Asymptotic equipartition

- Definition: **Random process** $\{X_i\} : X_1, X_2, ..., X_n, \quad n \in \mathbb{Z}$
- Definition: i.i.d. random process
  - X_i, X_j are independent $\forall i \neq j$, and $p(X_i) = p(X_j)$
  - i.i.d. is worst case for compression.

3 (Weak) Law of Large Numbers

- Consider an i.i.d. R.P. $\{Z_i\}$
  - $E[Z_i] = \mu$, $E[(Z_i - \mu)^2] = \sigma^2$.
- Define sample mean: $\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$
- LLN: $\bar{Z}_n$ converges "in probability" to μ.

$$Prob(|\bar{Z}_n - \mu| \geq \epsilon) \xrightarrow{n\to\infty} 0, \quad \forall \epsilon > 0$$

- We say that $\bar{Z}_n \xrightarrow{probability} \mu$
- But LLN says even more ...

$$Prob(|\bar{Z}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon} \xrightarrow{n\to\infty} 0$$

- This convergence is slower for correlated data
  - more data needed to explore the entire space...!

- Example: $n = 365$, $X_i \in \mathcal{X} = \{r, s\}$
  - (weather data simplified to i.i.d.)
- $X^n = (X_1, X_2, ..., X_n)$, $Prob(r) = 10\%$, $Prob(s) = 90\%$

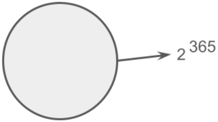
- $X^n = (r, r, r, ..., r), X^n = (s, s, s, ..., s)$ are atypical sequences:
  - possible mathematically, but really low probability.

- $X^n = (s, s, r, s, s, ..., r, s)$ more typical sequence: what we expect! about 10% of “r” and 90% “s”.

- Take $\{X_i\}$ on i.i.d. R.P. $p(X_i)$ is the distribution of $X_i \sim X$ $(X_i \in \mathbb{R})$
- Let $Z_i \triangleq logp(X_i)$. $\{Z_i\}$ also i.i.d. process. $(Z_i \in \mathbb{R}^-)$
  - i.e., Draw: $X^n = [X_1\ X_2 \cdots X_n]$
  - create instance $Z^n = [\log p(X_1)\ \log p(X_2)\ \dots\ \log p(X_n)] = [Z_1 \dots Z_n]$
- $\frac{1}{n}\sum_{i=1}^{n} Z_i \to E(Z_i) = E[\log p(X_i)] = \sum_{x\in\mathcal{X}} p(x)\log p(x) = -H(X)$

$$\frac{1}{n \log p(X_1,\dots,X_n) = \frac{1}{n}\log p(X^n) \xrightarrow{n\to\infty} -H(X)}$$

$$\Rightarrow P(X^n) \approx 2^{-nH(X)},\ n \to \infty \qquad \text{(simply intuition)}$$



- $P(X^n) \approx \frac{1}{2^{nH(X)}}$ and not $\frac{1}{2^{365}}$
  - Many sequences are negligible.
  - Key for Compression: place focus on "typical" sequences

EURECOM Sophia Antipolis

3 Asymptotic equipartition

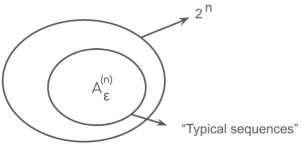
- LLN:

$$Prob(|\frac{1}{n} log P(X^n) + H(X)| > \epsilon) \to 0$$

- Define typical set: $\mathcal{A}_\epsilon^{(n)}$

$$\mathcal{A}_\epsilon^{(n)} = \{X^n \in \mathcal{X}^n \;\; s.t. \;\; 2^{-n(H(X)+\epsilon)} \leq P(X^n) \leq 2^{-n(H(X)-\epsilon)}\}$$

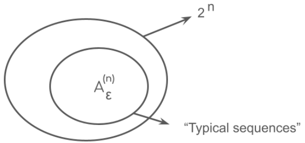
- Typical sequences: high likelihood to be observed



3 Basic properties of the typical set

For large n

- 1). $Prob(X^n \in \mathcal{A}_\epsilon^{(n)}) \overset{(LLN)}{\geq} 1-\epsilon$ (easy - restricted)
- 2). $|\mathcal{A}_\epsilon^{(n)}| \approx 2^{n(H(X))} << 2^n$ (binary)
  - from $\sum_{X^n \in \mathcal{A}_\epsilon^{(n)}} P(X^n) \approx 1$



3 Data Compression

- Compress: $X^n = (X_1, X_2, ..., X_n) \stackrel{(\xi)}{\rightarrow} \xi(X^n) = (0, 1, 1, ..., 0)$
  - to vector of length $\ell = \mathcal{L}(\xi(X^n)) << n$
- Recall $\mathcal{A}_\epsilon^{(n)}$ contains no more than $2^{n(H(X)+\epsilon)}$ elements/vectors

- Take a sequence X^n, examine if $X^n \in \mathcal{A}_\epsilon^{(n)}$ or not
  - a) if $X^n \in \mathcal{A}_\epsilon^{(n)}$, map X^n to index of $\lceil nH(X) + \epsilon \rceil$ bits
  - b) if $X^n \notin \mathcal{A}_\epsilon^{(n)}$, map X^n to index of $\lceil \log |X^n| \rceil$ bits.
  - c) add 1 bit to indicate if case a) or case b).
- This a compression based on typicality

- Define $\mathcal{L}(\xi(X^n))$: Length of the bit string X^n was mapped into

$$
\begin{aligned}
E[\mathcal{L}(\xi(X^n))] &= \sum_{X^n \in \mathcal{X}^n} \mathcal{L}(\xi(X^n))P(X^n) \\
&= \sum_{X^n \in \mathcal{A}_\epsilon^{(n)}} \mathcal{L}(\xi(X^n))P(X^n) + \sum_{X^n \notin \mathcal{A}_\epsilon^{(n)}} \mathcal{L}(\xi(X^n))P(X^n) \\
&\leq \sum_{X^n \in \mathcal{A}_\epsilon^{(n)}} (n(H(X)+\epsilon)+1+1)P(X^n) \\
&+ \sum_{X^n \notin \mathcal{A}_\epsilon^{(n)}} (\log|X^n|+1+1)P(X^n) \\
&\leq n(H(X)+\epsilon)+2+(\log|X^n|+2)\epsilon
\end{aligned}
$$

- Normalize... $\frac{E[\mathcal{L}(\xi(X^n))]}{n} \leq H(X)+\epsilon+\frac{2}{n}+\epsilon(log|X|+\frac{2}{n})$
- $\frac{E[\mathcal{L}(\xi(X^n))]}{n}$ is arbitrarily close to entropy (but scheme not scalable).

3 Compression of correlated sequences

- What if $\{X_i\}$ no longer iid ... identically distributed though

$$e.g. \quad X_i \in \mathcal{X} = \begin{cases} \text{r} & 0.1, \\ \text{s} & 0.9 \end{cases}$$

- But X_i and X_{i+1} correlated.

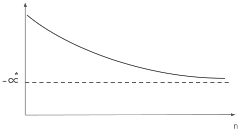
$$H(X_i) = H(X_j) = H(X), \quad \forall i \neq j$$

- Definition: **Entropy rate**
  - D_1: $H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, ..., X_n) = \lim_{n\to\infty} \frac{1}{n} H(X^n)$
  - D_2: $H(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_1, X_2, ..., X_{n-1})$
- $H(\mathcal{X})$ sets the limit of compression of correlated sources
- In general $H(\mathcal{X}) \leq H(X)$ (equality when $\{X_i\}$ iid)
  - Compressing correlated sources takes less space than iid data

- Note: D_1 same as D_2 (for stationary random process)

Proof:

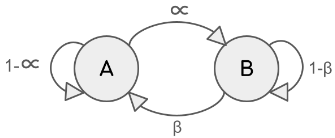
- $H(X_{n+1}|X_1, X_2, ..., X_n) =: \alpha_{n+1}$
- $H(\mathcal{X}) \overset{(D_2)}{=} \lim_{n\to\infty} \alpha_n$
- $\alpha_{n+1} = H(X_{n+1}|X_1, X_2, ..., X_n) \overset{(cre)}{\leq} H(X_{n+1}|X_2, X_3, ..., X_n) = H(X_n|X_1, X_2, ..., X_{n-1}) = \alpha_n$ (stationary)
- $\alpha_{n+1} \leq \alpha_n, \quad \Rightarrow \alpha_n \to \alpha^*$ when $n \to \infty$



- $H(\mathcal{X}) \overset{(D_1)}{=} \frac{1}{n} H(X_1, X_2, ..., X_n) \overset{(cr)}{=} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_1, X_2, ..., X_{i-1}) = \frac{1}{n} \sum_{i=1}^{n} \alpha_i \xrightarrow{n\to\infty} \alpha^*$ (because sequence converges)
- $\Rightarrow D_1 = D_2$

3 Application of entropy rate: Markov chains

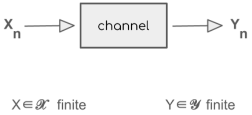
- Evaluate entropy rate for Binary Markov chains (BMC):



$$X_i = 0\,(A) \text{ or } 1\,(B), \quad P(X_i) \stackrel{(stnr)}{=} P(X_{i-1}), \quad P_0 = \frac{\beta}{\alpha+\beta},\; P_1 = \frac{\alpha}{\alpha+\beta} \quad (**)$$

- Entropy rate:
  $H(\mathcal{X}) \stackrel{(D_2)}{=} \lim_{n\to\infty} H(X_n|X_1, X_2, ..., X_{n-1}) \stackrel{(mc)}{=} H(X_n|X_{n-1})$
- $H(X_2|X_1) = ... = H(\alpha)P_0 + H(\beta)P_1$ (binary entropy)
- $H(\mathcal{X}) \stackrel{(**)}{=} \frac{H(\beta)\alpha+H(\alpha)\beta}{\alpha+\beta} \stackrel{(cre)}{<} H(X_2) = H(\frac{\alpha}{\alpha+\beta}) \stackrel{(sym)}{=} H(\frac{\beta}{\alpha+\beta})$ □
- Equality for $\alpha = \beta = 0.5$, no correlation anymore.
  - no gain, no difference between entropy rate and entropy.

EURECOM
Sophia Antipolis



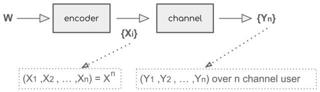
- Channel given by probability transition matrix:

$$Prob(Y_n = y | X_n = x), \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

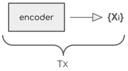
- Memoryless channel:
  - Y_n depends on X_n only, not on X_{n-1}, X_{n-2}, ...
  - Assume stationarity: the index n does not matter

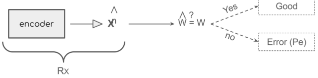
$$Prob(Y = y | X = x), \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

- Fundamental question: what is max reliable communication rate?



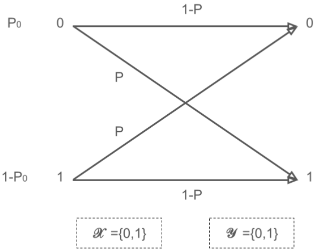
- $w \in \mathcal{W} = \{w^1, ..., w^M\}$: message set
- $\{X_i\}$: stream of bits in which the message is encoded
- Rate="number of information bits sent per channel use"
- $log_2 M$: information bits per message
  - as $M \uparrow$, messages too close, harder for RX to distinguish $\Rightarrow \hat{X}_i \neq X_i$
- Capacity C is maximal rate R such that $P_e \to 0$ as $n \to \infty$.





4 Channel Capacity for the BSC

- Let's explore the Binary Symmetric Channel (BSC)



- We will use (and prove later) that

$$C = \max_{P(X)} I(X;Y), \quad P(X) \text{ input distribution}$$

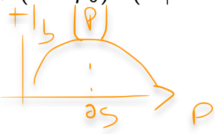
4 Channel Capacity for the BSC

- Start with $C = \max_{P(X)} I(X;Y)$
- $I(X;Y) = H(Y) - H(Y|X) \quad (*)$

$$H(Y|X) = \sum_{x\in\mathcal{X}} H(Y|X=x)P(x) = H(Y|X=0)p_0 + (1-p_0)H(Y|X=1)$$

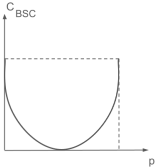
binary entropy

$$= H_b(p)p_0 + (1-p_0)H_b(p) = H_b(p) \quad (**)$$



- $I(X:Y) \stackrel{(*),(**)}{=} H(Y) - H_b(p) \leq 1 - H_b(p)$
  - $H(Y) \leq 1$ Maxed if $p_0 = 0.5$
- $C_{BSC} = \max_{P(X)} I(X;Y) = 1 - H_b(p)$
  - achieved when $P(X=0) = P(X=1) = \frac{1}{2}$ which yields maximizing $P(Y=0) = P(Y=1) = \frac{1}{2}$

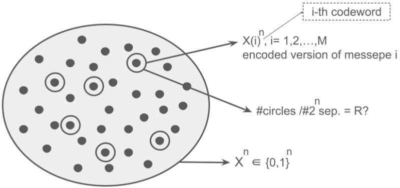
less randomness better



4 Encoding and Decoding

- Each message $i = \{1, ..., M\}$, maps onto long vector $X^n(i) = \{0, 1\}^n$ from Binary Code $\mathcal{C}$

iid



  > Recall: Entropy rate (affects goodness of code)

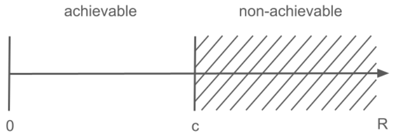
$$H = \lim_{n\to\infty} H(X_n|X_1, X_2, ..., X_{n-1}) \overset{(if\ \ mc)}{=} H(X_n|X_{n-1})$$

- Transmit vector $X^n(i)$, corrupted by channel, to yield Y^n
- Want to decode: $g(Y^n) = i \in \{1, 2, ..., M\}$

- $\mathcal{C} = codebook = \{X^n(1), X^n(2), ..., X^n(M)\}$
- Define probability error:
  - Def. $\lambda_i = Prob(g(Y^n) \neq i \mid X^n = X^n(i))$
  - Def. $\lambda^n = max_{i=1,2,\ldots,M}\ \lambda_i$ (worst case)
  - Def. $P_e = \frac{1}{M}\sum_{i=1}^{M} \lambda_i$ (average case)
- We want

$$\lambda^n \xrightarrow{n\to\infty} 0$$

- A rate R is achievable iff there exists a series of (M, n) codes such that $\lambda^n \xrightarrow{n\to\infty} 0$



- **Theorem:** the maximum of all achievable rates R is given by

$$C = max_{P(X)} I(X;Y)$$

Proof:

- Let us build a random code
- We generate M code words $X^n(1), X^n(2), ..., X^n(M)$.
  - According to $P(X^n) = \prod_{i=1}^{n} P(X_i)$, $X^n = (X_1, X_2, ..., X_n)$
- Assume that message i is transmitted ($X^n(i)$ is transmitted)
- Decoder receives Y^n
- Decoder finds message i s.t. $(X^n(i), Y^n)$ is jointly typical
  - Let's see what joint typicality (J.T.) means

- Def. Sequences X^n and Y^n are jointly typical if:
- 1. X^n is typical[3] with accuracy ϵ, i.e., if $|-\frac{1}{n}\log P(X^n) - H(X)| < \epsilon$
  - Recall $P(X^n) \approx 2^{-nH(X)}$
  - Recall $|\mathcal{A}_{X,\epsilon}| \approx 2^{nH(X)}$
- 2. Y^n is typical with accuracy ϵ, i.e., if $|-\frac{1}{n}\log P(Y^n) - H(Y)| < \epsilon$
  - Recall $P(Y^n) \approx 2^{-nH(Y)}$
  - Recall $|\mathcal{A}_{Y,\epsilon}| \approx 2^{nH(Y)}$
- 3. (X^n, Y^n) has to be such that $|-\frac{1}{n}\log P(X^n, Y^n) - H(X,Y)| < \epsilon$
  - $P(X^n, Y^n) \approx 2^{-nH(X,Y)}$
  - $|\mathcal{A}_\epsilon| \approx 2^{nH(X,Y)}$

1 2 3 not weird

[3] Make histogram, consider as pdf, calculate entropy.

Key property of joint-typicality:

- Recall: channel represented by $P(Y|X)$ i.e., by $P(X, Y)$
- Let (X^n, Y^n), drawn from $\prod_i^n P(X_i, Y_i)$ (i.e., channel related)
  - Y^n has been produced after sending X^n through channel
  - $\prod_i^n P(X_i, Y_i)$ since X_i iid and channel memoryless.

prob of   marginals

jointly   typical

Theorem

- 1. $\text{Prob}((X^n, Y^n) \in \mathcal{A}_\epsilon) \xrightarrow{n\to\infty} 1$
  - From LLN (as before - convergence of sequence of logs of joint distr.)
- 2. Let $\tilde{X}^n, \tilde{Y}^n \sim P(X^n)P(Y^n)$ generated
  - i.e. $\tilde{X}^n, \tilde{Y}^n$ are independent
  - Observing $\tilde{Y}^n$. Going through all X^n. If $\tilde{X}^n$ not the tx, then independent to $\tilde{Y}^n$. $(\tilde{X}^n, \tilde{Y}^n \sim P(X^n)P(Y^n))$
  - similarly: $\tilde{Y}^n$ is not "channel associated" to transmitted $\tilde{X}^n$
  - Then:

$$Prob((\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon) \approx 2^{-n(I(X,Y))} \tag{1}$$

vanishingly small probs

- Sketch Proof of (1):
  - How many jointly typical pairs come from distribution $P(X^n)P(Y^n)$

$$P((\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon) = \sum_{(x^n, y^n) \in \mathcal{A}_\epsilon^n} P(x^n)P(y^n)$$

  - Number of summands is $|\mathcal{A}_\epsilon^n| \approx 2^{nH(X,Y)}$
  - Pairs are jointly typical so $P(X^n) = 2^{-nH(X)}$, $P(Y^n) = 2^{-nH(Y)}$

$$P((\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon) \leq\approx \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{-nI(X;Y)}$$

- Example:

$$|\mathcal{A}_{\epsilon,X}| = |\mathcal{A}_{\epsilon,Y}| = 1000, \quad |\mathcal{A}_\epsilon| = 300$$

$$\Rightarrow P((\tilde{X}^n, \tilde{Y}^n) \in \mathcal{A}_\epsilon) \leq \frac{300}{10^6} = 3 \cdot 10^{-4}$$

- ...Continue toward proving $C = max_{P(X)} I(X;Y)$
- Consider random code $\mathcal{C}$
- Let us assume **message** $i = 1$ has been sent via $X^{(1)} \in \mathcal{C}$
- Let us assume that Y^n **was received**
- Error events : $E_i = \begin{cases} 1 & \text{If } Y^n \text{ is J.T. with } X^{(i)}, \\ 0 & \text{If not} \end{cases}$
- An error occurs if $E_1 = 0$ or $E_2 = 1$ or ... $E_M = 1$
  missed duplicated
- Thus

$$P_e \overset{(unionB)}{\leq} Prob(E_1 = 0) + \sum_{i=2}^{M} Prob(E_i = 1)$$

$$\leq \epsilon + (2^{nR} - 1)2^{-n(I(X,Y)-3\epsilon)}$$

> Since $Prob(E_1 = 0) \approx \epsilon \to 0$ from Theorem (part 1)
> Since $M = 2^{nR}$
> Since $Prob(E_i = 1) \approx 2^{-nI(X,Y)}$ from Theorem (part 2)

EURECOM
Sophia Antipolis

5 Deriving Channel Capacity

- Have seen that

$$P_e \leq \epsilon + (2^{nR} - 1)2^{-n(I(X,Y)-3\epsilon)}$$

- Thus

$$P_e \leq 3\epsilon \to 0, \quad n >> 1, \quad R < I(X, Y).$$

rate

- Thus $R < I(X, Y)$
- Eventually ...

$$C = max_{P(X)} I(X; Y)$$

$R = \frac{\log|M|}{n}$

EURECOM Sophia Antipolis

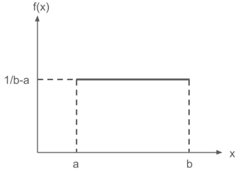
6 Information Theory of Continuous Data

- X is continuous R.V with distribution $f(x)$ (PDF)
- $F(x) = Prob(X \leq x)$ CDF $\rightarrow f(x) = \frac{dF(X)}{dx}$
- $\int_{-\infty}^{+\infty} f(x)\, dx = 1$

Definition: Differential entropy $X \sim f(X)$

$$h(x) \triangleq -E[\log f(x)] = \int_{-\infty}^{+\infty} -f(x) \log f(x)\, dx = \int_{Supp(X)} -f(x) \log f(x)\, dx$$

- $X \sim U[a, b]$
  $a, b \in \mathbb{R}, \quad a \leq b \quad \Rightarrow h(x) = -E[\log f(x)] = -E[\log \frac{1}{b-a}] = \log(b-a)$



Example: Gaussian pdf (maximizes Entropy)

- $X \sim N(0, \sigma^2)$, $f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{X^2}{2\sigma^2}}$

$$
\begin{aligned}
h(X) &= -\int_{-\infty}^{+\infty} f(X) \log f(X) = -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{X^2}{2\sigma^2}} \log_2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{X^2}{2\sigma^2}} dx \\
&= -\log \frac{1}{\sqrt{2\pi\sigma^2}} - \int_{-\infty}^{+\infty} -\frac{x^2 log_2 e}{2\sigma^2} f(x)\, dx \\
&= -\log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{\log_2 e}{2\sigma} \int_{-\infty}^{+\infty} X^2 f(X)\, dx
\end{aligned}
$$

- Since $f(X)$ and X^2 are symmetric, then second term goes to zero

$$\Rightarrow h(X) = \log \sqrt{2\pi e \sigma^2}$$

> The higher the σ^2, the more disorder we have, the higher $h(X)$

EURECOM Sophia Antipolis

6 Review of Gaussian Variables

- 1) Central limit theorem: $X_1, ..., X_n$ i.i.d arbitrary
  - > $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n$ empirical average

$$\Rightarrow \sqrt{n}\bar{X}_n \rightarrow \bar{X} \sim N(0, \sigma^2)$$

- 2) Uncorrelated vs. independent

> $E[XY] = 0 \Leftarrow$ f(X,Y)=f(X)f(Y)

$$\nRightarrow$$

- 3) Sum of independent Gaussians
  - > X, Y jointly Gaussian $\Rightarrow Z = X \pm Y \sim N(\mu_X \pm \mu_Y, \sigma_x^2 + \sigma_Y^2)$
- 4) Shift and scale

$$Z \sim N(0,1) \quad \Rightarrow \quad X = aZ + b \sim N(b, a^2) \; a, b \;\; \text{constants}$$

- 5) Multivariate Gaussian $\underline{X} = \{X_1, X_2, ..., X_n\} \sim N(\underline{\mu}, \Sigma)$
  - > with $\underline{\mu} = E[\underline{X}], \Sigma = E[(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})^T]$

$$f(\underline{X}) = \frac{1}{\sqrt[n]{2\pi|\Sigma|}} e^{-\underline{X}^T\Sigma^{-1}\underline{X}}$$

EURECOM
Sophia Antipolis

Def: Joint and Conditional Entropy

- Def: $h(X,Y) \triangleq -E[\log f(X,Y)] = -\int_{-\infty}^{+\infty} f(X,Y) \log f(X,Y)\, dxdy$
- Def: $h(X|Y) = -\int_{-\infty}^{+\infty} f(X,Y) \log f(X|Y)\, dxdy$

Def: KL-divergence and Mutual Information

- Def: $D(f(x)||g(x)) = \int_{-\infty}^{+\infty} f(X) \log \frac{f(X)}{g(X)}\, dx$
- Def: $I(X,Y) = D(f(X,Y)||f(X)f(Y))$

Also similar to before

$$I(X,Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$$

$$h(Y|X) \leq h(Y), \quad h(X|Y) \leq h(X)$$

- **Chain Rule**

$$h(X_1, ..., X_n) = \sum_{i=1}^{n} h(X_i|X_1, ..., X_{i-1})$$
$$= \sum_{i=1}^{n} h(X_i|X_{i+1}, ..., X_n) \leq \sum_{i=1}^{n} h(X_i)$$

  - Equality for X_i independent
  - Remember $h(X, Y) = h(X) + h(Y|X)$.

- **Scale and Shift** $X \sim f(X)$
  - $Y = X + C$ (C fixed) $\Rightarrow h(Y) = h(X)$
  - $Y = aX$ (a fixed) $\Rightarrow h(Y) = h(X) + log_2 a$ (just plug in def.)

6 Maximization of Differential Entropy

- Recall that $h(X + C) = h(X) \rightarrow$ we focus w.l.o.g. on $E[X] = 0$
- Focus on $var(x) = \sigma^2$ (fixed) otherwise ill-posed problem
- max h(x)= ?
- Proof as before but now let $g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$ (Gaussian)
- As before use that $D(f(X)||g(X)) > 0$ to get

$$h(X) \leq \frac{1}{2}\log(2\pi e\sigma^2)$$

  - with equality for $f(x) = g(x)$ (gaussian)

  - Whereas recall: For discrete case $H(X) \leq \log_2 |\mathcal{X}|$ (uniform)

6 ...Typical Sequences (similar to discrete, BUT....)

- $X^n = \{X_1, ..., X_n\}$ iid
- $\frac{1}{n}\sum_{i=1}^{n} log_2 f(X_i) \xrightarrow{n \to +\infty} E[\log_2 f(X_i)] = -h(X)$
- $\Delta_\epsilon^{(n)} = \{X^n \in \mathbb{R}^n | \ |-\frac{1}{n}\sum_{i=1}^{n} log(f(X_i)) - h(X)| < \epsilon\}$

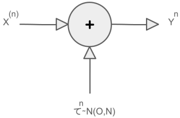
- Prop. 1

$$P(X^n \in \Delta_\epsilon^{(n)}) \overset{(LLN)}{\geq} 1 - \epsilon, \text{ as } n \to \infty$$

- Prop. 2

$$Vol(\Delta_\epsilon^{(n)}) \approx 2^{n(h(X))}$$

  - Where volume of set $S \subset \mathbb{R}^n$ is $Vol(S) = \int_S dx^n$



- Why Gaussian?
  - Mathematical Tractability and CLT!
  - Useful for the worst case!
  - AWGN Z_i a good assumption
- $w \in \mathcal{W} = \{1, 2, 3, ..., M\}$ message, which is encoded to $X^n(w) = (X_1, X_2, ..., X_n)$
- AWGN Channel

$$Y_i = X_i + Z_i \quad \forall i = \{1, ..., n\}, \text{Rate} \quad R = \frac{\log_2 M}{n}$$

Q: what is R_{max} s.t. $p(\hat{w} = w) \xrightarrow{n \to \infty} 1$

7 Capacity of Gaussian channel

Looking for

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

- Under power constraint $\begin{cases} E[X_i^2] \leq P & \forall i \quad \text{(instantaneous)} \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq P & \text{average} \end{cases}$
- Let us compute:

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(X+Z|X) = h(Y) - h(Z) \\ &= h(Y) - \frac{1}{2}\log(2\pi e N) \leq \frac{1}{2}\log(2\pi e \sigma_y^2) - \frac{1}{2}\log(2\pi e N) \end{aligned}$$

- $h(y)$ is always upper bounded by Gaussian entropy
- If $X \sim N(0,P)$ then Y Gaussian with $\sigma_y^2 = \sigma_x^2 + \sigma_z^2 \leq P + N$

$$I(X;Y) \leq \frac{1}{2}\log(2\pi e(P+N)) - \frac{1}{2}\log(2\pi e N)$$

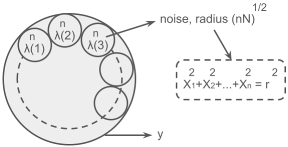
- Thus $\quad C = \frac{1}{2}\log(1+\frac{P}{N})$ bits/sec/Hz or bits/channel use

- $\mathcal{W} = \{1, 2, ..., M\} \xrightarrow{encoder} \mathcal{C} = \{X^n(1), X^n(2), ..., X^n(M)\}$ codebook
- $Y^n = X^n(w) + Z^n$ and

$$\sum_{i=1}^{n} X_i^n(w)^2 = nP \quad \text{(power constraint } \forall w)$$

- When n large

$$\frac{1}{n}\sum_{i=1}^{n} Z_i^2 \approx N \qquad \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \approx N + P$$



- Small Hypersphere of radius $\sum_{i=1}^{n} Z_i^2 \approx nN$
- Big Hypersphere of radius $\sum_{i=1}^{n} Y_i^2 \approx n(N + P)$
- Decoder works if small bubbles do not touch (and fails otherwise)

EURECOM Sophia Antipolis

7 Capacity of gaussian channel – Sphere packing argument

- volume of n-dimensional hyper-sphere is $vol(n) = \alpha_n r^n$
  - volume of hyper-sphere of radius $\sqrt{n(N+P)}$ is $\alpha_n(n(N+P))^{\frac{n}{2}}$
  - volume of "noisy bubble" is $\alpha_n(nN)^{\frac{n}{2}}$
- M (number of bubbles/messages) can not be bigger than

$$M \leq \frac{\alpha_n(n(N+P))^{\frac{n}{2}})}{\alpha_n(nN)^{n/2}} = (1 + \frac{P}{N})^{\frac{n}{2}}$$

- This corresponds to $log((1+\frac{P}{N})^{\frac{n}{2}})$ bits, over n dimensions
- Thus

$$R \leq \frac{1}{2}\log(1+\frac{P}{N}) \quad \text{per real dimension}$$