Information Theory & Coding

Nov 9th 2020

Last Time:

"differential Entropy":

- $h(X) = E\left[\log \frac{1}{f_X(X)}\right]$ $\quad f_X$: probability density function of X

- $h(X^n) = E\left[\log \frac{1}{f_{X^n}(X^n)}\right]$ $\quad f_{X^n}$: joint density of the random vector X^n

- $h(X^n) = \sum_{i=1}^{n} h(X_i | X^{i-1})$.

- $h(X+a) = h(X)$
- $h(aX) = \log|a| + h(X)$

- $$\begin{aligned} I(X;Y) &= h(X) + h(Y) - h(XY) \\ &= h(X) - h(X|Y) \\ &= h(Y) - h(Y|X) \\ &= \lim_{\substack{\varepsilon \downarrow 0 \\ \delta \downarrow 0}} \underline{I(X_\varepsilon; Y_\delta)} \end{aligned}$$

with $X_\varepsilon = \varepsilon \left\lfloor \frac{X}{\varepsilon} \right\rfloor$

$Y_\delta = \delta \left\lfloor Y/\delta \right\rfloor$

mean

$\underline{Ex}$: $X \sim N(\mu, \sigma^2)$ — variance $= E(X^2) - E(X)^2$

normal $\equiv$ Gaussian

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

– Maximal entropy property of Gaussians.

$\underline{Ex}$: A Gaussian vector $X^n = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$ is a

random vector whose p.d.f is of the form

$$f_{X^n}(\vec{x}^n) = \frac{1}{\det(2\pi K)^{1/2}} \exp -\frac{1}{2}(\vec{x}-\vec{\mu})^T K^{-1}(\vec{x}-\vec{\mu}).$$

where K is an $n \times n$ matrix & $\vec{\mu}$ is n-vector

$\underline{\vec{\mu} = E(X^n)}$, K is the covariance matrix,

$\underline{K_{ij}} = E[(X_i - \mu_i)(X_j - \mu_j)]$.

Such an X^n is said to be $N(\underline{\vec{\mu}}, \underline{\underline{K}})$

Suppose $\vec{\mu} = \vec{0}$: Consider

$$\underbrace{v^T K v} = \sum_{i,j} v_i K_{ij} v_j$$

$$= \sum_{i,j} v_i E(X_i X_j) v_j$$

$$= E\left[\sum_{i,j} v_i X_i v_j X_j \right)$$

$$= E\left[\left(\sum_i v_i X_i \right)^2 \right) \geq 0.$$

$\Rightarrow$ the matrix K is non-negative definite matrix.
(positive semi ")

Suppose $Z^n = \underline{N(0, I_n)}$, i.e.,
$\underbrace{Z_1 \ldots Z_n \text{ are } N(0,1), \text{ independent.}}$

Now consider, $X^n = A Z^n$

$$X_1 = \sum_j A_{1j} Z_j$$

$$X_2 = \sum A_{2j} Z_j$$

$$\vdots$$

$$X_n = \sum A_{nj} Z_j$$

$$E(X_i X_j) = E\Big(\sum_k A_{ik} Z_k \sum_l A_{jl} Z_l\Big)$$

$$= \sum_{k,l} A_{ik} A_{jl} E(Z_k Z_l)$$

$$= \sum_k A_{ik} A_{jk} = (AA^T)_{ij}$$

$$\Rightarrow \underbrace{X^n}_{} = N(\vec{0}, \underbrace{AA^T}_{}).$$

$\Rightarrow$ for any pos. semi def K $\exists$ a Gaussian,

$$X^n \sim N(0, K).$$

Q: what is $h(X^n)$ if $X^n \sim N(0,K)$.

$$\log \frac{1}{f_{X^n}(\vec{x})} = \frac{1}{2}\log\det(2\pi K) + \frac{1}{2}\vec{x}^T K^{-1}\vec{x}\,(\log e)$$

$$h(X^n) = \frac{1}{2}\log\det(2\pi K) + \frac{1}{2}\underbrace{E\big[X^T K^{-1} X\big]}_{}\log e$$

$$X^T K^{-1} X = \sum_{i,j} X_i (K^{-1})_{ij} X_j$$

$$E(\quad) = \sum_{i,j} (K^{-1})_{ij} K_{ij}$$

$$= \sum_{i,j} (K^{-1})_{ij} K_{ji}$$

$$= \sum_i (\underbrace{K^{-1} K}_{I})_{ii}$$

$$= n$$

$$\Rightarrow h(X^n) = \frac{1}{2} \log \det(2\pi K) + \frac{1}{2} n (\log e)$$

$$= \frac{1}{2} \log \left(e^n \det(2\pi K) \right)$$

$$= \frac{1}{2} \log \det(2\pi e K).$$

So the entropy of a $N(\bar{\mu}, K)$ is

$$\frac{1}{2} \log \det(2\pi e K).$$

<u>Thm</u>: if X^n is any Random vector $(\mathbb{R}^n)$ with covariance matrix K then

$$h(X^n) \le \frac{1}{2} \log \det (2\pi e K).$$

Pf: Assume without loss of generality that

$$E(X^n) = \vec{0}. \text{ so } K \equiv E(XX^T).$$

let f be the p.d.f of X^n

let g be " " " a gaussian,

$$g(\bar{x}) = \frac{1}{\det(2\pi K)^{1/2}} \exp\left(-\frac{1}{2} \bar{x}^T K^{-1} \bar{x}\right).$$

note that

$$\int g(x) \log \frac{1}{g(x)}\, dx = \int f(x) \log \frac{1}{g(x)}\, dx$$

because $\log \frac{1}{g(x)} = \text{const} + \sum_{i,j}' x_i x_j c_{ij}$.

$$h(X^n) = \int f(x) \log \frac{1}{f(x)}\, dx$$

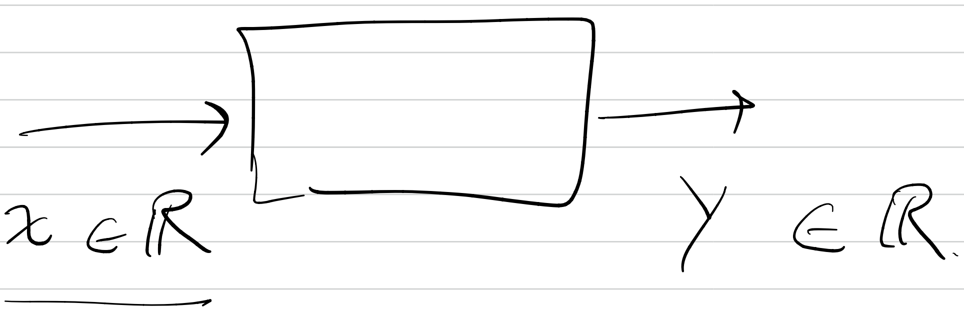
$$\frac{1}{2} \log \det (2\pi e K) = \int g(x) \log \frac{1}{g(x)}\, dx = \int f(x) \log \frac{1}{g(x)}\, dx$$

$$h(X^n) - \frac{1}{2}\log_2 \det(2\pi e K)$$

$$= \int f(x) \log \frac{g(x)}{f(x)}\, dx$$

$$\leq \log \int \cancel{f(x)} \frac{g(x)}{\cancel{f(x)}}\, dx = \log 1 = 0. \quad //$$

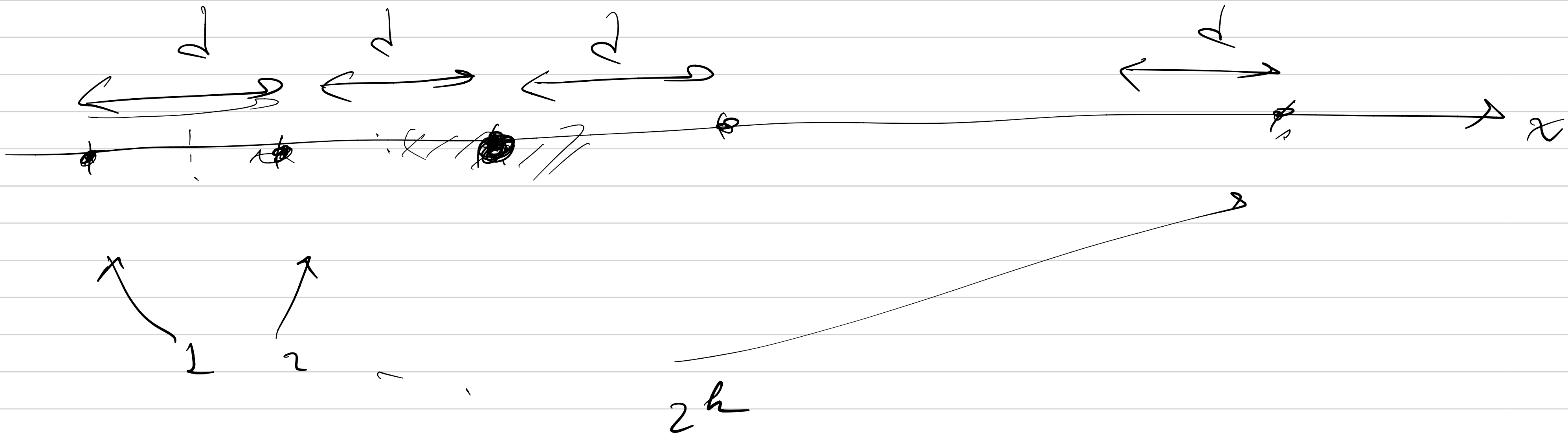
Additive Gaussian Noise Channel:



$$Y = N(x, \sigma^2) \equiv Y = x + Z$$

where Z is $N(0, \sigma^2)$, the statistics of Z does not depend on x.

With no restrictions on x, we see that we can send arbitrary number k of bits / channel use with arbitrary reliability ε.



$\Pr(|Z| > \frac{d}{2}) < \varepsilon.$

$\Rightarrow$ $\Pr($ Y is closer to the sent x then any other possible sent points $) \geq 1 - \varepsilon$

$\Rightarrow$ we have "code" ($n = 1$, $M = 2^k$ codewords)

& $P(\text{error}) < \varepsilon.$

$\Rightarrow$ Unconstrained capacity of this channel is ∞.

- The more natural problem is to impose some constraints on the input.

One possibility : $|x| < A$. i.e. $x \in [-A, A]$.

& compute $\max_{f_X : \text{supported on } [-A, A]} I(X;Y) = C(A)$

Another possibility : $E(X^2) \leq P$, we now want to

compute $\max_{f_X : E(X^2) \leq P} I(X;Y) =: C(P)$.

Thm: for the AGN channel, $Y = X + Z$, $Z \sim N(0, \sigma^2)$

with the constraint $E(X^2) \leq P$,

$C(P) =$

Pf : $\max_{X : E(X^2) \leq P} I(X;Y) = \max h(Y) - h(Y|X)$.

$h(Y|X) = h(X+Z|X) = h(Z|X) = h(Z)$

$= \frac{1}{2} \log(2\pi e \sigma^2)$

So $C(P) = \underbrace{\max h(Y)}_{\dots} - \frac{1}{2} \log(2\pi e \sigma^2)$.

$Y = X + Z$. $Var(Y) = Var(X) + Var(Z)$

$\leq P + \sigma^2$ (= iff $E(X) = 0$).

$\Rightarrow h(Y) \leq \frac{1}{2} \log(2\pi e (P + \sigma^2))$ (= iff Y is Gaussian)

$$I(X;Y) \le \frac{1}{2}\log 2\pi e(P+\sigma^2) - \frac{1}{2}\log 2\pi e\,\sigma^2$$

$$= \frac{1}{2}\log\frac{P+\sigma^2}{\sigma^2} = \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right).$$

equality will hold if $E(X)=0$ $E(X^2)=P$.

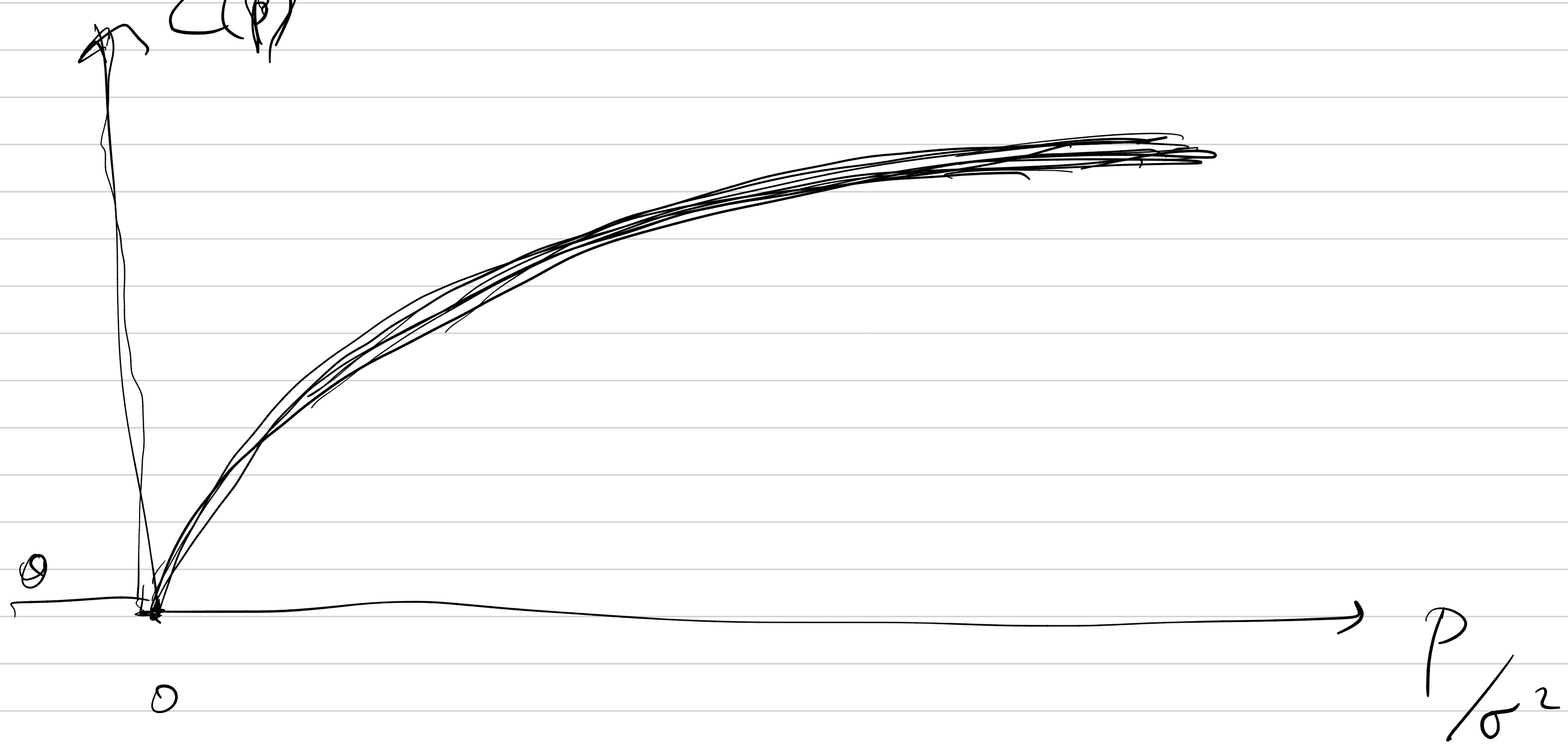
& Y is Gaussian.

$\equiv$ if $X \sim N(0,P)$.

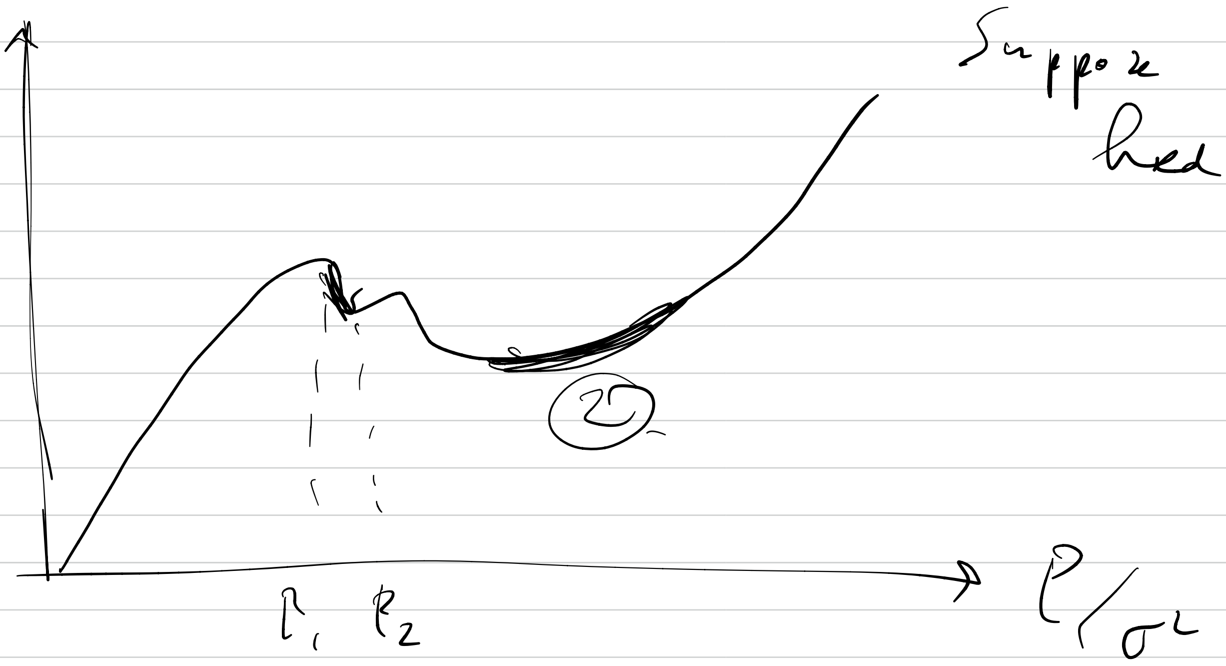
$$\Rightarrow \max I(X;Y) = \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) =: C(P).$$

$\frac{P}{\sigma^2}$ = signal to noise ratio. $P = E(X^2)$

= SNR $\sigma^2 = E(Z^2)$



$P \mapsto C(P)$ is $\nearrow$ and concave



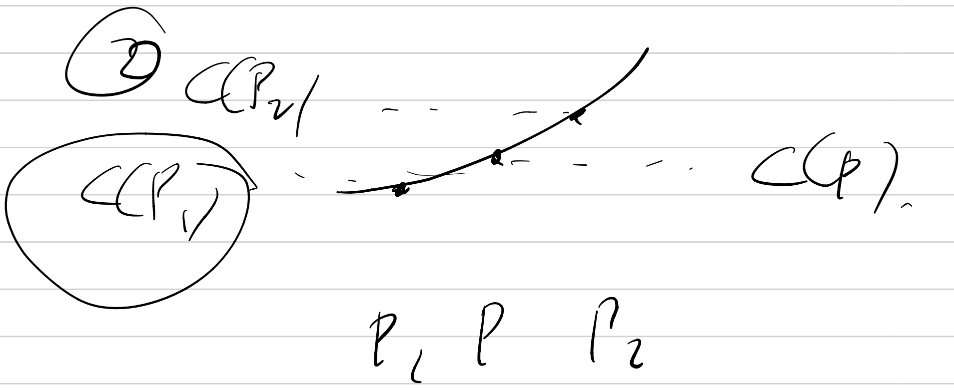
Suppose our computation had yielded

This could not have been correct:

① ↘ : can't happen.

$$\max_{E(X^2)\le P_1} I(X;Y) \le \max_{E(X^2)\le P_2} I(X;Y)$$

②



$P = \lambda P_1 + (1-\lambda)P_2$, but

$C(P) < \lambda C(P_1) + (1-\lambda) C(P_2)$

but $\exists$ a comm. scheme with $E(X^2) \le P_1$ & transmits $C(P_1)$ bits/ch. use ①

$\exists$ " " " " $\le P_2$ & " $C(P_2)$ " " ②

we can build a new system, which simply uses

system ① on λ - fraction of the time

② " $(1-\lambda)$ " " "

this new system will have $\underline{E(X^2)} \leq \underline{\lambda P_1 + (1-\lambda) P_2}$

& will transmit $\underline{\lambda C(P_1) + (1-\lambda) C(P_2)}$ bits/ch use

$\Rightarrow \underline{C(\lambda P_1 + (1-\lambda) P_2)} \geq \underline{\lambda C(P_1) + (1-\lambda) C(P_2)}$.

$\Rightarrow$ C can't be non-concave. //

Soft vs Hard constraints: A hard constraint of the type $X \in [a,b)$ can be expressed as:

$$\underline{E[\mathbb{1}_{(a,b)^c}(X)]} = 0.$$

by choosing $b(X) = \begin{cases} 0 & \text{if } a \leq x \leq b \\ 1 & \text{else} \end{cases}$ we turn

the "hard" constraint on X to a soft constraint.

---

constraint

$X \in [-A, A)$, $\underline{C(A)}$.

The argument we used above proves the following Thm

Thm: given a channel $p(y|x)$, and a function $b : \mathcal{X} \to \mathbb{R}$ the function

$$\beta \mapsto C(\beta) = \max_{X : E[b(X)] \leq \beta} I(X;Y)$$

is non-decreasing and concave.

Remember that we had proved that

$$C(\beta) = \max_{E(b(X)) \leq \beta} I(X;Y)$$ is the capacity of a

channel with constraint β for discrete, stationary, memoryless channels.

Q: is the theorem still true when $\mathcal{X}, \mathcal{Y}$ are non-discrete.

A: Yes, because $I(X;Y) \approx I(X_\varepsilon ; Y_\delta)$. ///