

Lec 5 : Functional (In)equalities

Yanjun Han



Recall: Shannon-type inequalities, i.e. all entropy inequalities that can be derived using:

$$\textcircled{1} \text{ monotonicity: } H(X) \leq H(X, Y)$$

$$\textcircled{2} \text{ submodularity: } I(X; Y|Z) \geq 0$$

This lecture will cover some non-Shannon-type inequalities.

Def (differential entropy). For a RV X with a density f on \mathbb{R}^d , its differential entropy is defined as

$$h(X) := h(f) := \int_{\mathbb{R}^d} -f(x) \log f(x) dx.$$

Note: ① $h(X) \in \mathbb{R} \cup \{\pm\infty\}$. In particular, it can be negative.

$$\textcircled{2} \quad h(aX) = h(X) + \log a, \text{ for } a \in \mathbb{R}$$

③ $h(X) \leq h(X, Y)$ no longer holds. However, it's still true that

$$I(X; Y) = h(X) + h(Y) - h(X, Y) \geq 0.$$

Example. If $X \sim N(\mu, \Sigma)$, then $f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$, so

$$h(X) = \mathbb{E}_{X \sim f} \left[\frac{1}{2} \log((2\pi)^d \det(\Sigma)) + \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right]$$

$$= \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log \det \Sigma.$$

Thm (entropy power inequality, EPI) For independent RVs X, Y on \mathbb{R}^d ,

$$e^{\frac{2}{d}h(X+Y)} \geq e^{\frac{2}{d}h(X)} + e^{\frac{2}{d}h(Y)}.$$

Note: ① Equality holds iff X, Y are Gaussian, and $\Sigma_X = c\Sigma_Y$.

② EPI shows that, for given values of $h(X)$ and $h(Y)$, $h(X+Y)$ is minimized when X, Y are Gaussian.

We will present the proof in Stam (1959).

Detour: Fisher information. For a RV X with density f , the Fisher information is

$$J(X) := \int_{\mathbb{R}} \frac{(f'(x))^2}{f(x)} dx$$

Recall: Fisher information $I(\theta)$ in Lec 3: for $Y \sim P_\theta$.

$$I(\theta) := I^Y(\theta) := \int \frac{\left(\frac{\partial}{\partial \theta} p_\theta\right)^2}{p_\theta} dx.$$

They are connected via $I^Y(\theta) \equiv J(X)$ when $Y = \theta + X$.

Properties: ① $J(aX) = \frac{1}{a^2} J(X)$

② DPI: $I^Y(\theta) \leq I^X(\theta)$ if $\theta - X - Y$ is a Markov chain.

(Pf: $I^Y(\theta) = \lim_{\alpha \rightarrow 0} \frac{1}{\alpha^2} X^2(P_{Y|\theta+\alpha} \| P_{Y|\theta}) \leq \lim_{\alpha \rightarrow 0} \frac{1}{\alpha^2} X^2(P_{X|\theta+\alpha} \| P_{X|\theta}) = I^X(\theta)$)

Thm (Stam) For independent X_1, X_2 :

$$\frac{1}{J(X_1+X_2)} \geq \frac{1}{J(X_1)} + \frac{1}{J(X_2)}.$$

or equivalently, $(a+b)^2 J(X_1+X_2) \leq a^2 J(X_1) + b^2 J(X_2)$. $\forall a, b > 0$.

Pf. Write $Y_1 = a\theta + X_1$, $Y_2 = b\theta + X_2$, then

$$I^{Y_1}(\theta) = I^{Y_1/a}(\theta) = J\left(\frac{X_1}{a}\right) = a^2 J(X_1).$$

$$\text{Therefore, } (a+b)^2 J(X_1+X_2) = I^{Y_1+Y_2}(\theta) \leq I^{Y_1, Y_2}(\theta) = a^2 J(X_1) + b^2 J(X_2). \quad \square$$

Thm (de Bruijn). For $Z \sim N(0, 1)$ independent of X , then for $a > 0$,

$$\frac{d}{da} h(X + \sqrt{a}Z) = \frac{1}{2} J(X + \sqrt{a}Z).$$

Pf. Let $p_a = p * N(0, a)$ be the density of $X + \sqrt{a}Z$, then

$$\frac{\partial p_a}{\partial a} = \frac{1}{2} p_a'' \quad (*)$$

To see (*), just note that for any test function f ,

$$\begin{aligned}
 \frac{\partial}{\partial a} \mathbb{E}_{P_a}[f] &= \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \mathbb{E}[f(X + \sqrt{a}\Delta Z) - f(X + \sqrt{a}Z)] \\
 &= \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \mathbb{E}[f(X + \sqrt{a}Z + \sqrt{\Delta}Z') - f(X + \sqrt{a}Z)], \quad Z' \text{ independent copy of } Z \\
 &= \lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \mathbb{E}[f'(X + \sqrt{a}Z) \cdot \sqrt{\Delta}Z' + \frac{1}{2} f''(X + \sqrt{a}Z) \cdot \Delta Z'^2 + o(\Delta)] \\
 &= \frac{1}{2} \mathbb{E}_{P_a}[f''] = \frac{1}{2} \int f'' P_a'' . \quad (\text{integration by parts})
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \frac{d}{da} h(X + \sqrt{a}Z) &= - \int (1 + \log P_a) \frac{\partial P_a}{\partial a} = -\frac{1}{2} \int (1 + \log P_a) P_a'' \\
 &= \frac{1}{2} \int \left(\frac{P_a'}{P_a} \right)^2 P_a'' \quad (\text{integration by parts}) \\
 &= \frac{1}{2} J(X + \sqrt{a}Z). \quad \blacksquare
 \end{aligned}$$

Proof of EPI. ① $d=1$. Let $X_\lambda = X * N(0, f(\lambda))$, $Y_\lambda = Y * N(0, g(\lambda))$,
for some functions f, g TBD. Since

$$\frac{d}{d\lambda} [e^{2H(X_\lambda)}] = 2e^{2H(X_\lambda)} J(X_\lambda) f'(\lambda),$$

we have

$$\frac{d}{d\lambda} \left[\frac{e^{2H(X_\lambda)} + e^{2H(Y_\lambda)}}{e^{2H(X_\lambda+Y_\lambda)}} \right] = \frac{2}{e^{2H(X_\lambda+Y_\lambda)}} \left(e^{2H(X_\lambda)} J(X_\lambda) f'(\lambda) + e^{2H(Y_\lambda)} J(Y_\lambda) g'(\lambda) \right) - \left(e^{2H(X_\lambda)} + e^{2H(Y_\lambda)} \right) J(X_\lambda+Y_\lambda) (f'(\lambda) + g'(\lambda))$$

Choosing $f'(\lambda) = e^{2H(X_\lambda)}$, $g'(\lambda) = e^{2H(Y_\lambda)}$, then

$$\frac{d}{d\lambda} \left[\frac{e^{2H(X_\lambda)} + e^{2H(Y_\lambda)}}{e^{2H(X_\lambda+Y_\lambda)}} \right] \geq 0, \quad \forall \lambda > 0.$$

As $\lambda \rightarrow \infty$, both X_λ and Y_λ are "more and more Gaussian", the ratio $\rightarrow 1$.

Therefore, this ratio at $\lambda=0$ must be ≤ 1 , which is the EPI.

② General $d \geq 2$ by induction:

$$\begin{aligned}
 h(X^d + Y^d) &= h(X^{d-1} + Y^{d-1}) + h(X_d + Y_d | X^{d-1} + Y^{d-1}) \\
 &\geq h(X^{d-1} + Y^{d-1}) + h(X_d + Y_d | X^{d-1}, Y^{d-1}) \quad (\text{conditioning reduces entropy}) \\
 &\geq \frac{d-1}{2} \log(e^{\frac{2}{d-1} h(X^{d-1})} + e^{\frac{2}{d-1} h(Y^{d-1})}) \quad (\text{induction hypothesis}) \\
 &\quad + \frac{1}{2} \mathbb{E}_{X^{d-1}, Y^{d-1}} \log(e^{2h(X_d | X^{d-1} = x^{d-1})} + e^{2h(Y_d | Y^{d-1} = y^{d-1})}) \quad (X \perp\!\!\!\perp Y) \\
 &\geq \frac{1}{2} \log(e^{2h(X_d | X^{d-1})} + e^{2h(Y_d | Y^{d-1})}) \quad \text{by convexity of } (x, y) \mapsto \log(e^x + e^y) \\
 &\geq \frac{d}{2} \log(e^{\frac{2}{d} h(X^{d-1})} + e^{\frac{2}{d} h(Y^{d-1})}) \quad \text{by convexity of } (x, y) \mapsto \log(e^x + e^y) \text{ again} \\
 &= \frac{d}{2} \log(e^{\frac{2}{d} h(X^d)} + e^{\frac{2}{d} h(Y^d)}). \quad \square
 \end{aligned}$$

Example. Let X_1, X_2, \dots be i.i.d., $\mathbb{E}[X_i] = 0$, $\text{Var}(X_i) = 1$, and $h(X_i) > -\infty$.

Let $T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ be the standardized sum. Then by EPI,

$$\begin{aligned}
 h(T_{n+m}) &= h(\sqrt{\frac{m}{n+m}} \cdot \frac{1}{\sqrt{m}} \sum_{i=1}^m X_i + \sqrt{\frac{n}{n+m}} \cdot \frac{1}{\sqrt{n}} \sum_{i=m+1}^{n+m} X_i) \\
 &\geq \frac{1}{2} \log(e^{2h(\sqrt{\frac{m}{n+m}} T_m)} + e^{2h(\sqrt{\frac{n}{n+m}} T_n)}) \\
 &= \frac{1}{2} \log\left(\frac{m}{n+m} e^{2h(T_m)} + \frac{n}{n+m} e^{2h(T_n)}\right).
 \end{aligned}$$

In other words, the sequence $a_n := n e^{2h(T_n)}$ is super-additive:

$$a_{n+m} \geq a_n + a_m, \quad \forall n, m.$$

Moreover, since $\text{Var}(T_n) = 1$, the maximum entropy principle implies (exercise)

$$h(T_n) \leq \frac{1}{2} \log(2\pi e), \text{ so that } \frac{a_n}{n} \leq 2\pi e.$$

Therefore, $\frac{a_n}{n}$ must have a limit, i.e. $h(T_n) \rightarrow h^*$, and

$$D_{KL}(P_{T_n} || N(0, 1)) = -h(T_n) + \frac{1}{2} \log(2\pi e) \rightarrow D^*.$$

Barron (1986) shows that $D^* = 0$, a result known as the entropic CLT.

Information and estimation in Gaussian model

Let X be a general RV

$$Y_r = \sqrt{r}X + Z, \quad Z \sim N(0,1) \text{ independent of } X$$

$r > 0$: SNR parameter

Thm (I-MMSE)

$$\frac{d}{dr} I(X; Y_r) = \frac{1}{2} \mathbb{E}[(X - \mathbb{E}[X|Y_r])^2] =: \frac{1}{2} \text{mmse}(X|Y_r)$$

Note: 1. Perhaps the most surprising part is that this is an equality.

$$2. \text{mmse}(X|Y) = \mathbb{E}[|X - \mathbb{E}[X|Y]|^2] = \min_{f(\cdot)} \mathbb{E}[|X - f(Y)|^2]$$

is called the minimum mean squared error for estimating X based on Y .

There are several proofs for the I-MMSE formula, but the most generalizable one is via SDEs:

A more general result: if $dY_t = X_t dt + dB_t$, $t \in [0, T]$, then

$$I(X^T; Y^T) = \frac{1}{2} \int_0^T \mathbb{E}[(X_t - \mathbb{E}[X_t|Y^t])^2] dt.$$

To see how it implies the I-MMSE formula, take $X_t \equiv X$. Then Y_T is a sufficient statistic of Y^T for estimating X , i.e.

$$I(X^T; Y^T) = I(X; Y_T), \quad \mathbb{E}[X_t | Y^t] = \mathbb{E}[X | Y_t].$$

Moreover, $\frac{Y_T}{\sqrt{T}} = \sqrt{T}X + N(0, 1)$, so the SNR parameter is T .

The proof of the general result uses the filtering theory for BMs.

Lemma 1. For $dY_t = f(t)dt + dB_t$ with $f(t)$ adapted to the filtration \mathcal{F}^Y , then

$$\log \frac{dP_{Y^T}}{dP_{B^T}}(g^T) = \int_0^T f(s)dB_s - \frac{1}{2} \int_0^T f(s)^2 ds.$$

Intuition. For $t > 0$ and small $\Delta > 0$, the conditional distribution of $\hat{Y}_{t+\Delta} - \hat{Y}_t | \mathcal{F}^Y$ is

$$\begin{cases} N\left(\int_t^{t+\Delta} f(s)ds, \Delta\right) & \text{under } P_{Y^T} \\ N(0, \Delta) & \text{under } P_{B^T} \end{cases}$$

$$\begin{aligned} \text{so the log-likelihood ratio is } & \frac{1}{2} \int_t^{t+\Delta} f(s)ds \cdot (\hat{Y}_{t+\Delta} - \hat{Y}_t) - \frac{1}{2\Delta} \left(\int_t^{t+\Delta} f(s)ds \right)^2 \\ & \approx f(t)(\hat{Y}_{t+\Delta} - \hat{Y}_t) - \frac{\Delta}{2} f(t)^2. \end{aligned}$$

$$\text{Summing up gives } \sum_i f(t_i)(\hat{Y}_{t_i+\Delta} - \hat{Y}_{t_i}) - \frac{\Delta}{2} \sum_i f(t_i)^2 \xrightarrow{\Delta \rightarrow 0} \int_0^T f(t)d\hat{Y}_t - \frac{1}{2} \int_0^T f(t)^2 dt.$$

(Think: where did we use that f is adapted to \mathcal{F}^Y ?)

Lemma 2. For $dY_t = X_t dt + dB_t$, then

$$\tilde{B}_t = Y_t - \int_0^t \mathbb{E}[X_s | \mathcal{F}^Y] ds$$

is a BM adapted to \mathcal{F}^Y .

(A major difference is that X_t could be an unknown signal not adapted to \mathcal{F}^Y ; however, $\mathbb{E}[X_t | \mathcal{F}^Y]$ is always adapted to \mathcal{F}^Y)

Pf. Clearly \tilde{B}_t is adapted to \mathcal{F}^Y . In addition,

$$\tilde{B}_t = \int_0^t (X_s - \mathbb{E}[X_s | \mathcal{F}^Y]) ds + B_t$$

is an \mathcal{F}^Y -adapted martingale, satisfies $\tilde{B}_0 = 0$, and has quadratic variation t .

By Lévy's criterion, \tilde{B}_t is a BM. □

(Think: B_t is a BM; but is it adapted to \mathcal{F}^Y ?)

Returning to the proof.

$$I(X^T; Y^T) = \mathbb{E}_{P_{X^T, Y^T}} \left[\log \frac{P_{Y^T|X^T}}{P_{Y^T}} \right] = \mathbb{E}_{P_{X^T, Y^T}} \left[\log \frac{P_{Y^T|X^T}}{P_{B^T}} \right] - \mathbb{E}_{P_{X^T, Y^T}} \left[\log \frac{P_{Y^T}}{P_{B^T}} \right].$$

For the first term, since X^T is given (conditioned), Lemma 1 gives

$$\mathbb{E}_{P_{X^T, Y^T}} \left[\log \frac{P_{Y^T|X^T}}{P_{B^T}} \right] = \mathbb{E} \left[\int_0^T X_t dY_t - \frac{1}{2} \int_0^T X_t^2 dt \right].$$

For the second term, Lemma 2 tells that $\tilde{B}_t = Y_t - \int_0^t \mathbb{E}[X_s | Y^s] ds$ is a \mathcal{F}^Y -BM,

so

$$\log \frac{P_{Y^T}}{P_{B^T}} (g^T) = \log \frac{P_{Y^T}}{P_{\tilde{B}^T}} (g^T) \stackrel{\text{Lemma 1 again}}{\downarrow} = \int_0^T \mathbb{E}[X_t | Y^t] dY_t - \frac{1}{2} \int_0^T \mathbb{E}[X_t | Y^t]^2 dt$$

$$\Rightarrow \mathbb{E} \left[\log \frac{P_{Y^T}}{P_{B^T}} \right] = \mathbb{E} \left[\int_0^T \mathbb{E}[X_t | Y^t] dY_t - \frac{1}{2} \int_0^T \mathbb{E}[X_t | Y^t]^2 dt \right].$$

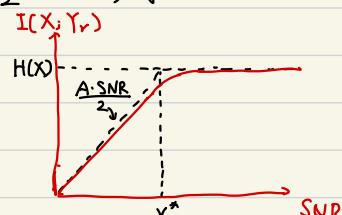
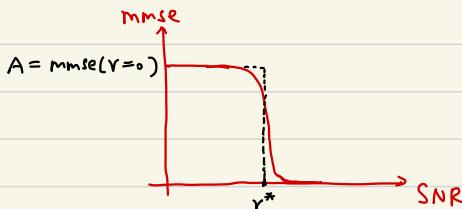
Therefore,

$$\begin{aligned} I(X^T; Y^T) &= \mathbb{E} \left[\int_0^T (X_t - \mathbb{E}[X_t | Y^t]) dY_t + \frac{1}{2} \int_0^T (\mathbb{E}[X_t | Y^t]^2 - X_t^2) dt \right] \\ &= \mathbb{E} \left[\int_0^T ((X_t - \mathbb{E}[X_t | Y^t]) X_t + \frac{1}{2} (\mathbb{E}[X_t | Y^t]^2 - X_t^2)) dt \right] \\ &= \int_0^T \mathbb{E}[(X_t - \mathbb{E}[X_t | Y^t])^2] dt. \end{aligned}$$



How is the I-MMSE formula useful in statistics?

Suppose we expect a problem to have a sharp phase transition at $\text{SNR} = r^*$, we can try to show that $I(X; Y_r) \geq \frac{A_r}{2} (1 - o(1))$ for all $r \leq (1 - \varepsilon) r^*$ (see picture)



In this case,

$$\frac{(1-\varepsilon)r^*}{2} \text{mmse}(0)(1-o(1)) \leq I(X; Y_{(1-\varepsilon)r^*}) = \frac{1}{2} \int_0^{(1-\varepsilon)r^*} \text{mmse}(r) dr$$

$r \mapsto \text{mmse}(r)$
is non-increasing

$$\leq \frac{(1-2\varepsilon)r^*}{2} \cdot \text{mmse}(0) + \frac{\varepsilon r^*}{2} \text{mmse}((1-2\varepsilon)r)$$

$$\Rightarrow \text{mmse}((1-2\varepsilon)r^*) \geq (1-o(1)) \text{mmse}(0), \text{ i.e. the MMSE does not really drop before } r = r^*.$$

Comparison with Fano: Recall that at a high level, Fano's inequality shows that the estimation error is large when the information $I(X; Y)$ is small. Surprisingly, the I-MMSE formula shows that this is also the case if $I(X; Y)$ is too large, and is particularly good at showing sharp transitions and identifying the exact threshold.

An example.

Consider the "sparse" mean estimation problem: $Y \sim N(\theta, 1)$, with $\theta \sim (1-p)\delta_0 + p\delta_\mu$. $p = o(1)$.

Thm. If $\mu \leq \sqrt{2(1-\varepsilon)\log \frac{1}{p}}$, then

$$\text{mmse}(\theta | Y) \geq (1-o(1)) \mathbb{E}[\theta^2] = (1-o(1)) p\mu^2.$$

(In other words, the mmse is essentially attained by the best estimator $\hat{\theta} = p\mu$ without seeing Y .)

Pf sketch. Let $X \sim (1-p)\delta_0 + p\delta_\mu$, $\mu = \sqrt{r}$. then $Y \stackrel{d}{=} Y_r = \sqrt{r}X + N(0, 1)$.

The mutual information can be computed as

$$I(X; Y_r) = \mathbb{E} \left[\log \frac{P_{Y_r|X}}{P_{Y_r}} \right] = \mathbb{E} \left[\log \frac{P_{Y_r|X}}{Q_{Y_r}} \right] - D_{KL}(P_{Y_r} || Q_{Y_r}) \text{ for any } Q.$$

Choose $Q_{Y_r} = N(p\sqrt{r}, 1)$, then

$$\mathbb{E} \left[\log \frac{P_{Y_r|X}}{Q_{Y_r}} \right] = \mathbb{E} \left[D_{KL}(P_{Y_r|X} \| Q_{Y_r}) \right] = \mathbb{E} \left[\frac{(\sqrt{r}X - p\sqrt{r})^2}{2} \right] = \frac{p(1-p)}{2} r,$$

$D_{KL}(P_{Y_r} \| Q_{Y_r}) = o(pr)$ after some algebra if $r < 2(1-\epsilon) \log \frac{1}{p}$.

$$\Rightarrow I(X; Y_r) \geq \frac{p(1-p)}{2} r (1-o(1)) \text{ if } r < 2(1-\epsilon) \log \frac{1}{p}.$$

Now using the previous I-MMSE program proves that

$$\text{mmse}(X|Y_r) \geq (1-o(1)) \text{Var}(X) = (1-o(1)) p \quad \text{if } r < 2(1-\epsilon) \log \frac{1}{p}$$

$$\Rightarrow \text{mmse}(\Theta|Y) = r \cdot \text{mmse}(X|Y_r) \geq (1-o(1)) pr^2 \quad \text{if } r < \sqrt{2(1-\epsilon) \log \frac{1}{p}}$$

◻

Tensorization of I-MMSE

Thm. If $Y_r = \sqrt{r}X + N(0, I_n)$, then

$$\frac{d}{dr} I(X; Y_r) = \frac{1}{2} \mathbb{E} [\|X - \mathbb{E}[X|Y_r]\|_2^2] =: \frac{1}{2} \text{mmse}(X|Y_r).$$

Pf. Consider the model where $Y_i = \sqrt{r_i}X_i + N(0, 1)$ for possibly different (r_1, \dots, r_n) , then

$$\frac{\partial}{\partial r_i} I(X; Y^n) = \frac{\partial}{\partial r_i} I(X_i; Y^n) + \underbrace{\frac{\partial}{\partial r_i} I(X_{\sim i}; Y^n | X_i)}_{=0 \text{ as } \sqrt{r_i}X_i \text{ can be subtracted from } Y_i \text{ when } X_i \text{ is known}}$$

$$= \underbrace{\frac{\partial}{\partial r_i} I(X_i; Y_{\sim i})}_{=0} + \frac{\partial}{\partial r_i} I(X_i; Y_i | Y_{\sim i})$$

by 1-D I-MMSE

$$= \frac{1}{2} \text{mmse}(X_i | Y^n)$$

$$\Rightarrow \frac{d}{dr} I(X; Y_r) = \sum_{i=1}^n \frac{\partial}{\partial r_i} I(X; Y_r) \Big|_{r_i=r} = \frac{1}{2} \text{mmse}(X | Y_r).$$

◻

Area theorem: a related result based on a similar tensorization idea

Consider the communication problem over a BEC ($Y = \{X\}?$ w.p. $1-\varepsilon$, w.p. ε). with input $X^n \sim \text{Unif}(C) = \text{Unif}(\{\overset{\uparrow}{X_1}, \dots, X_n\})$, with $M = e^{nR}$.

How to find a codebook s.t. $\frac{1}{n} \sum_{i=1}^n H(X_i | Y^n) \xrightarrow{\text{average bit error rate}} 0$ when $R < C = 1 - \varepsilon$?

Defn (EXIT function) $h_i(\varepsilon) = H(X_i | Y_{\sim i}), i \in [n]$
 $h(\varepsilon) = \frac{1}{n} \sum_{i=1}^n h_i(\varepsilon)$.

Lemma. $H(X_i | Y^n) = \varepsilon h_i(\varepsilon)$.

$$\begin{aligned} \text{Pf. } H(X_i | Y^n) &= (1-\varepsilon)H(X_i | Y_{\sim i}, Y_i \neq ?) + \varepsilon H(X_i | Y_{\sim i}, Y_i = ?) \\ &= \varepsilon H(X_i | Y_{\sim i}) = \varepsilon h_i(\varepsilon) \quad \blacksquare \end{aligned}$$

($h_i(\varepsilon)$ is the error probability of decoding X_i in the "non-trivial" scenario $Y_i = ?$)

Lemma. $\frac{d}{d\varepsilon} H(X^n | Y(\varepsilon)) = nh(\varepsilon)$.

Pf. Again, think of n independent channels with different erasure probabilities $(\varepsilon_1, \dots, \varepsilon_n)$. Then

$$\begin{aligned} \frac{\partial}{\partial \varepsilon_i} H(X^n | Y^n) &= \frac{\partial}{\partial \varepsilon_i} H(X_i | Y^n) + \underbrace{\frac{\partial}{\partial \varepsilon_i} H(X_{\sim i} | X_i, Y^n)}_{= H(X_{\sim i} | X_i, Y_{\sim i}) \text{, so derivative}} = 0 \\ &\stackrel{\text{by previous lemma}}{=} \frac{\partial}{\partial \varepsilon_i} (\varepsilon_i h_i(\varepsilon_i)) = h_i(\varepsilon_i). \end{aligned}$$

$$\Rightarrow \frac{d}{d\varepsilon} H(X^n | Y(\varepsilon)) = \left. \sum_{i=1}^n H(X_i | Y_{\sim i}) \right|_{\varepsilon_1 = \dots = \varepsilon_n = \varepsilon} = nh(\varepsilon). \quad \blacksquare$$

Area Thm (BEC): $\int_0^1 h(\varepsilon) d\varepsilon = R$.

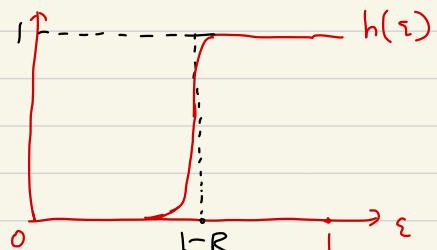
$$\text{Pf. } \int_0^1 h(\varepsilon) d\varepsilon = \frac{1}{n} \int_0^1 \frac{d}{d\varepsilon} H(X^n | Y(\varepsilon)^n) d\varepsilon = \frac{H(X^n | Y(1)^n) - H(X^n | Y(0)^n)}{n} \\ = \frac{H(X^n)}{n} = R. \quad \blacksquare$$

What does the area thm tell us? For a capacity-achieving code of rate $R = C$, it must hold that $h(\varepsilon) = 0(1)$ when $\varepsilon < 1-R$.

However, since $h(\varepsilon) \leq 1$ and $\int_0^1 h(\varepsilon) d\varepsilon = R$,

it must be the case that $h(\varepsilon) \approx 1$ for every $\varepsilon > 1-R$, i.e. the code is really bad in the

high-noise regime. Therefore, any capacity-achieving code must have a sharp transition for the decoding error.



Special topic: any "symmetric" linear code achieves the capacity of BEC

Linear code: $C = \{X_1^n, \dots, X_m^n\}$ is a linear subspace of \mathbb{F}_2^n .

(The encoding step of linear codes is easy: just a matrix-vector product)

"Symmetry": for all $i \neq k, j \neq l, \exists \pi \in S_n$ s.t. $\pi(i) = j, \pi(k) = l$, and
 $\pi C = C$ (πC applies the permutation π to all vectors in C)

Thm. For every symmetric linear code with $\frac{\log M}{n} \rightarrow R$, it attains the BEC capacity under the bit-MAP decoding.

$$(\text{i.e. } \hat{x}_i = \underset{x_i \in \{0,1\}}{\operatorname{argmax}} P(x_i | y^n))$$

(In the coding literature, this shows that the Reed-Muller code, which is symmetric and admits efficient encoding and decoding algorithms, is capacity-achieving.)

Proof ingredient I: Boolean function

Let $\Omega \subseteq \{0,1\}^n$. We call Ω :

① monotone: if $x \in \Omega$ and $x \leq x'$, then $x' \in \Omega$

② symmetric: if for all $i, j \in [n]$, $\exists \pi \in S_n$ s.t. $\pi(i) = j$ and $\pi(\Omega) = \Omega$.

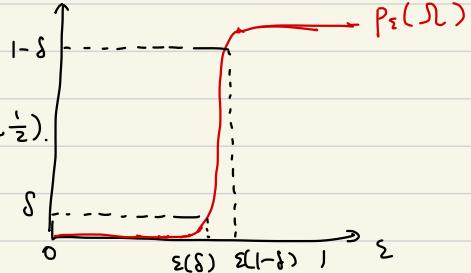
Also, for $\varepsilon \in [0,1]$, define a probabilistic object

$$p_\varepsilon(\Omega) = P(\text{Bern}(\varepsilon)^{\otimes n} \in \Omega).$$

(By monotonicity, $\varepsilon \mapsto p_\varepsilon(\Omega)$ is non-decreasing; for symmetry, we shall only need that all influence functions of Ω are the same. i.e. $I_1(\Omega) = \dots = I_n(\Omega)$, with

$$I_i(\Omega) = P_\varepsilon(x \in \{0,1\}^n : (x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n) \notin \Omega \text{ and } (x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_n) \in \Omega)$$

Let $\varepsilon(\delta) = \max \{ \varepsilon : p_\varepsilon(\Omega) \leq \delta \}$.



Thm. $\varepsilon(1-\delta) - \varepsilon(\delta) = o(1)$, $\forall \delta \in (0, \frac{1}{2})$.

(This shows that the function $\varepsilon \mapsto p_\varepsilon(\Omega)$ has a sharp threshold.)

Pf sketch. A classical result shows that

$$\frac{d}{d\varepsilon} p_\varepsilon(\Omega) = \sum_{i=1}^n I_i(\Omega) = n I_1(\Omega) \text{ by symmetry.}$$

It remains to show that $n I_1(\Omega) = w(1)$ whenever $p_\varepsilon(\Omega) \in [\delta, 1-\delta]$.

Classical Efron-Stein bound: $p_\varepsilon(\Omega)(1-p_\varepsilon(\Omega)) \lesssim \sum_{i=1}^n I_i(\Omega)$ only shows $n I_1(\Omega) = \Omega(1)$.

Key improvement (KKL theorem): $\frac{\log n}{n} \cdot p_\varepsilon(\Omega)(1-p_\varepsilon(\Omega)) \lesssim \max \{ I_1(\Omega), \dots, I_n(\Omega) \}$

↑
essentially the log-Sobolev inequality on the hypercube $\Rightarrow n I_1(\Omega) = \Omega(\log n) = w(1)$. ◻

Proof ingredient II: area theorem.

For a given linear code C , define

$$\Omega_i = \left\{ \text{all erasure patterns } w \in \{0,1\}^{n-1} \text{ such that } w \odot x_{n-i} \text{ fails to decode } x_i, \text{ for some } x \in C \right\}$$

(1 represents erasure, 0 represents non-erasure)

Since C is linear, WLOG can assume that $x=0$, i.e.

$$\Omega_i = \left\{ w \in \{0,1\}^{n-1} : \exists x_{n-i} \leq w \text{ st. } (x_{n-i}, 1) \in C \right\}.$$

Then: ① Ω_i is monotone (obvious)

② Ω_i is symmetric (follows from symmetry of C)

③ $P_\varepsilon(\Omega_i) = P(Y_{n-i} \text{ fails to decode } X_i) = h_i(\varepsilon)$

④ $h_i(\varepsilon) = h(\varepsilon)$ (symmetry of C again)

By the previous part, $\varepsilon \mapsto h(\varepsilon) = P_\varepsilon(\Omega_i)$ has a sharp threshold.
In addition, $\int_0^1 h(\varepsilon) d\varepsilon = R$ by area theorem.
This threshold can only be $\varepsilon^* = 1-R$, i.e. capacity-achieving!