Lec 5: Functional (In)equalities

Yanjun Han

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

① monotonicity: $H(X) \le H(X,Y)$

② submodularity: $I(X;Y \mid Z) \geq 0$

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

<u>Def (differential entropy)</u>. 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.

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

③ $h(X) \le h(X,Y)$ <u>no longer</u> holds. However, it's still true that

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

<u>Example</u>. 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

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

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

Easy fact (maximum entropy principle): If $\mathrm{Cov}(X) = \Sigma$, then $h(X) \le h(N(0,\Sigma))$.

<u>Pf</u>. $0 \le D_{KL}(P_X \,\|\, N(\mathbb{E}X, \Sigma)) = -h(X) + h(N(0,\Sigma))$ (check!). ☐

<u>Thm (entropy power inequality, EPI)</u> 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).

<u>Detour: Fisher information</u>. For a RV X with density f, the Fisher information is

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

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

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

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

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

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

$\left(\text{Pf: } I^Y(\theta) = \lim_{\Delta\to 0} \frac{1}{\Delta^2} \chi^2(P_{Y|\theta+\Delta} \| P_{Y|\theta}) \le \lim_{\Delta\to 0} \frac{1}{\Delta^2} \chi^2(P_{X|\theta+\Delta} \| P_{X|\theta}) = I^X(\theta) \right)$

<u>Thm (Stam)</u> For independent X_1, X_2:

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

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

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

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

Therefore, $(a+b)^2 J(X_1+X_2) = I^{Y_1+Y_2}(\theta) \le I^{Y_1,Y_2}(\theta) = a^2 J(X_1) + b^2 J(X_2)$. ☒

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

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

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

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

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

$$\frac{\partial}{\partial a}\mathbb{E}_{p_a}[f] = \lim_{\Delta\to 0}\frac{1}{\Delta}\mathbb{E}[f(X+\sqrt{a+\Delta}\,Z)-f(X+\sqrt{a}\,Z)]$$

$$= \lim_{\Delta\to 0}\frac{1}{\Delta}\mathbb{E}[f(X+\sqrt{a}\,Z+\sqrt{\Delta}\,Z')-f(X+\sqrt{a}\,Z)], \quad Z' \text{ independent copy of } Z$$

$$= \lim_{\Delta\to 0}\frac{1}{\Delta}\mathbb{E}[f'(X+\sqrt{a}\,Z)\cdot\sqrt{\Delta}\,Z' + \tfrac{1}{2}f''(X+\sqrt{a}\,Z)\cdot\Delta Z'^2 + o(\Delta)]$$

$$= \frac{1}{2}\mathbb{E}_{p_a}[f''] = \frac{1}{2}\int f p_a''. \quad \text{(integration by parts)}$$

Therefore,

$$\frac{d}{da}h(X+\sqrt{a}\,Z) = -\int(1+\log p_a)\frac{\partial p_a}{\partial a} = -\frac{1}{2}\int(1+\log p_a)\,p_a''$$

$$= \frac{1}{2}\int\frac{(p_a')^2}{p_a} \quad \text{(integration by parts)}$$

$$= \frac{1}{2}J(X+\sqrt{a}\,Z). \qquad \blacksquare$$

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

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

$$-(e^{2h(X_\lambda)}+e^{2h(Y_\lambda)})J(X_\lambda+Y_\lambda)(f'(\lambda)+g'(\lambda))\Big).$$

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 $\lambda\to\infty$, both X_λ and Y_λ are "more and more Gaussian", the ratio $\to 1$. Therefore, this ratio at $\lambda=0$ must be ≤ 1, which is the EPI.

② General $d \geq 2$ by induction:

$$h(X^d + Y^d) = h(X^{d-1} + Y^{d-1}) + h(X_d + Y_d \mid X^{d-1} + Y^{d-1})$$

$$\geq h(X^{d-1} + Y^{d-1}) + h(X_d + Y_d \mid X^{d-1}, Y^{d-1}) \quad \text{(conditioning reduces entropy)}$$

$$\geq \frac{d-1}{2} \log\left(e^{\frac{2}{d-1} h(X^{d-1})} + e^{\frac{2}{d-1} h(Y^{d-1})}\right) \quad \text{(induction hypothesis)}$$

$$+ \underbrace{\frac{1}{2} \mathbb{E}_{x^{d-1}, y^{d-1}} \log\left(e^{2h(X_d \mid X^{d-1} = x^{d-1})} + e^{2h(Y_d \mid Y^{d-1} = y^{d-1})}\right)}_{\geq \frac{1}{2}\log\left(e^{2h(X_d \mid X^{d-1})} + e^{2h(Y_d \mid Y^{d-1})}\right) \text{ by convexity of } (x,y) \mapsto \log(e^x + e^y)} \quad (X \perp\!\!\!\perp Y)$$

$$\geq \frac{d}{2} \log\left(e^{\frac{2}{d} h(X^{d-1}) + \frac{2}{d} h(X_d \mid X^{d-1})} + e^{\frac{2}{d} h(Y^{d-1}) + \frac{2}{d} h(Y_d \mid Y^{d-1})}\right)$$

by convexity of $(x, y) \mapsto \log(e^x + e^y)$ again

$$= \frac{d}{2} \log\left(e^{\frac{2}{d} h(X^d)} + e^{\frac{2}{d} h(Y^d)}\right). \quad \blacksquare$$

<u>Example</u>. Let $X_1, X_2, \cdots$ be i.i.d., $\mathbb{E}[X_1] = 0$, $\mathrm{Var}(X_1) = 1$, and $h(X_1) > -\infty$. Let $T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ be the standardized sum. Then by EPI,

$$h(T_{n+m}) = h\left(\sqrt{\frac{m}{n+m}} \cdot \frac{1}{\sqrt{m}} \sum_{i=1}^m X_i + \sqrt{\frac{n}{n+m}} \cdot \frac{1}{\sqrt{n}} \sum_{i=m+1}^{m+n} X_i\right)$$

$$\geq \frac{1}{2} \log\left(e^{2h\left(\sqrt{\frac{m}{n+m}} T_m\right)} + e^{2h\left(\sqrt{\frac{n}{n+m}} T_n\right)}\right)$$

$$= \frac{1}{2} \log\left(\frac{m}{n+m} e^{2h(T_m)} + \frac{n}{n+m} e^{2h(T_n)}\right).$$

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 $\mathrm{Var}(T_n) = 1$, the maximum entropy principle implies

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

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

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

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

Information and estimation in Gaussian model

Let X be a general RV

$Y_\gamma = \sqrt{\gamma}\, X + Z$, $\quad Z \sim N(0,1)$ independent of X

$\gamma > 0$: SNR parameter

Thm (I-MMSE)

$$\frac{d}{d\gamma} I(X; Y_\gamma) = \frac{1}{2}\mathbb{E}\left[(X - \mathbb{E}[X|Y_\gamma])^2\right] =: \frac{1}{2}\mathrm{mmse}(X|Y_\gamma)$$

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

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

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

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

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

$$\log \frac{dP_{Y^T}}{dP_{B^T}}(\xi^T) = \int_0^T f(t)\,d\xi_t - \frac{1}{2}\int_0^T f(t)^2\,dt.$$

Intuition. For $t \geq 0$ and small $\Delta > 0$, the conditional distribution of $\xi_{t+\Delta} - \xi_t \mid \xi^t$ is

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

so the log-likelihood ratio is
$$\frac{1}{\Delta}\int_t^{t+\Delta} f(s)\,ds \cdot (\xi_{t+\Delta} - \xi_t) - \frac{1}{2\Delta}\left(\int_t^{t+\Delta} f(s)\,ds\right)^2$$
$$\approx f(t)(\xi_{t+\Delta} - \xi_t) - \frac{\Delta}{2} f(t)^2.$$

Summing up gives $\sum_i f(t_i)(\xi_{t_i+\Delta} - \xi_{t_i}) - \frac{\Delta}{2} f(t_i)^2 \xrightarrow{\Delta \to 0} \int_0^T f(t)\,d\xi_t - \frac{1}{2}\int_0^T f(t)^2\,dt.$

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

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

$$\tilde{B}_t = Y_t - \int_0^t \mathbb{E}[X_s \mid Y^s]\,ds$$

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, $\mathbb{E}[X_t \mid Y^t]$ is always adapted to $\mathcal{F}^Y$)

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

$$\tilde{B}_t = \int_0^t \left(X_s - \mathbb{E}[X_s \mid Y^s]\right) ds + B_t$$

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. $\boxed{\text{Q}}$

(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 $\tilde{B}_t = Y_t - \int_0^t \mathbb{E}[X_s \mid Y^s]\, ds$ is a $\mathcal{F}^Y$-BM,

so

$$\log \frac{P_{Y^T}}{P_{B^T}}(\xi^T) = \log \frac{P_{Y^T}}{P_{\tilde{B}^T}}(\xi^T) \overset{\text{Lemma 1 again}}{=} \int_0^T \mathbb{E}[X_t \mid Y^t]\, d\xi_t - \frac{1}{2}\int_0^T \mathbb{E}[X_t \mid Y^t]^2\, dt$$

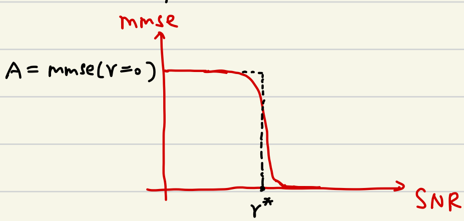
$$\Rightarrow \mathbb{E}\left[\log \frac{P_{Y^T}}{P_{B^T}}\right] = \mathbb{E}\left[\int_0^T \mathbb{E}[X_t \mid Y^t]\, dY_t - \frac{1}{2}\int_0^t \mathbb{E}[X_t \mid Y^t]^2\, dt\right].$$

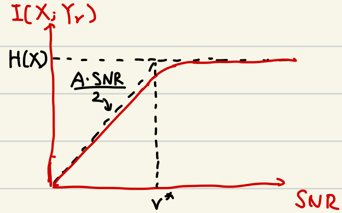
Therefore,

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

How is the I-MMSE formula useful in statistics?

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





In this case,

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

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

$$\le \frac{(1-2\varepsilon)r^*}{2}\cdot \mathrm{mmse}(0) + \frac{\varepsilon r^*}{2}\,\mathrm{mmse}((1-2\varepsilon)r)$$

$\Longrightarrow$ $\mathrm{mmse}((1-2\varepsilon)r^*) \ge (1-o(1))\,\mathrm{mmse}(0)$, 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

$$\theta \sim (1-p)\delta_0 + p\,\delta_\mu, \quad p = o(1).$$

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

$$\mathrm{mmse}(\theta \mid Y) \ge (1-o(1))\,\mathbb{E}[\theta^2] = (1-o(1))\,p\mu^2.$$

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

Pf sketch. Let $X \sim (1-p)\delta_0 + p\delta_1$, $\mu = \sqrt{r}$. then $Y \overset{d}{=} Y_r = \sqrt{r}X + N(0,1)$.

The mutual information can be computed as

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

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

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

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

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

Now using the previous I-MMSE program proves that

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

$$\Rightarrow \text{mmse}(\theta | Y) = r \cdot \text{mmse}(X | Y_{\mu^2}) \geq (1-o(1))\rho\mu^2 \quad \text{if} \quad \mu < \sqrt{2(1-\varepsilon)\log\frac{1}{\rho}}$$

$\blacksquare$

Tensorization of I-MMSE.

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

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

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

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

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

by 1-D I-MMSE

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

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

$\blacksquare$

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

Consider the communication problem over a BEC ($Y = \begin{cases} X & \text{w.p. } 1-\varepsilon \\ ? & \text{w.p. } \varepsilon \end{cases}$), with input $X^n \sim \mathrm{Unif}(\underset{\text{codebook}}{C}) = \mathrm{Unif}(\{x^n_{(1)}, \dots, x^n_{(M)}\})$, with $M = e^{nR}$.

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

Defn (EXIT function) $\quad h_i(\varepsilon) = H(X_i \mid Y_{\sim i})$, $i \in [n]$

$$h(\varepsilon) = \frac{1}{n}\sum_{i=1}^n h_i(\varepsilon).$$

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

Pf. $H(X_i \mid Y^n) = (1-\varepsilon) H(X_i \mid Y_{\sim i}, Y_i \neq ?) + \varepsilon H(X_i \mid Y_{\sim i}, Y_i = ?)$

$$= \varepsilon H(X_i \mid Y_{\sim i}) = \varepsilon h_i(\varepsilon)$$ ∎

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

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

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

$$\frac{\partial}{\partial \varepsilon_i} H(X^n \mid Y^n) = \frac{\partial}{\partial \varepsilon_i} H(X_i \mid Y^n) + \frac{\partial}{\partial \varepsilon_i} \overbrace{H(X_{\sim i} \mid X_i, Y^n)}^{= H(X_{\sim i} \mid X_i, Y_{\sim i}) \text{, so derivative } = 0}$$

by previous lemma

$$= \frac{\partial}{\partial \varepsilon_i}\left(\varepsilon_i H(X_i \mid Y_{\sim i})\right) = H(X_i \mid Y_{\sim i}).$$

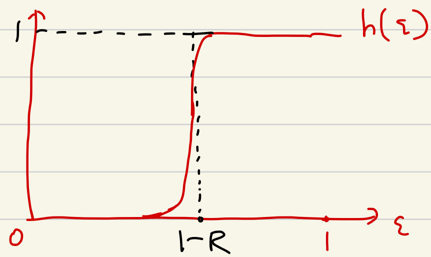
$$\Rightarrow \frac{d}{d\varepsilon} H(X^n \mid Y(\varepsilon)^n) = \sum_{i=1}^n H(X_i \mid Y_{\sim i})\Big|_{\varepsilon_1 = \dots = \varepsilon_n = \varepsilon} = n h(\varepsilon).$$ ∎

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

Pf. $\int_0^1 h(\varepsilon)\,d\varepsilon = \frac{1}{n}\int_0^1 \frac{d}{d\varepsilon} H(X^n \mid Y(\varepsilon)^n)\,d\varepsilon = \frac{H(X^n \mid Y(1)^n) - H(X^n \mid Y(0)^n)}{n}$

$= \frac{H(X^n)}{n} = R.$ $\square$

What does the area thm tell us? For a capacity-achieving code of rate $R = C$, it must hold that $h(\varepsilon) = o(1)$ when $\varepsilon < 1-R$. However, since $h(\varepsilon) \le 1$ and $\int_0^1 h(\varepsilon)\,d\varepsilon = R$, it must be the case that $h(\varepsilon) \simeq 1$ 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: $C = \{x^n_{(1)}, \cdots, x^n_{(M)}\}$ 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 $i \neq k$, $j \neq \ell$, $\exists\, \pi \in S_n$ s.t. $\pi(i) = j$, $\pi(k) = \ell$, and $\pi C = C$ (πC applies the permutation π to all vectors in C)

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

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

(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 $x \le x'$, then $x' \in \Omega$

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

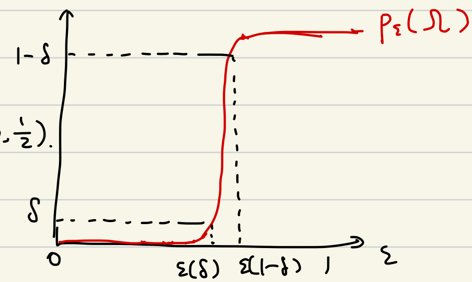
Also, for $\varepsilon \in [0,1]$, define a probabilistic object

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

(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. $I_1(\Omega) = \cdots = I_n(\Omega)$, with

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

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



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

(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 $n I_1(\Omega) = \omega(1)$ 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): $\frac{\log n}{n} \cdot p_\varepsilon(\Omega)(1 - p_\varepsilon(\Omega)) \lesssim \max\{I_1(\Omega), \cdots, I_n(\Omega)\}$

essentially the log-Sobolev inequality on the hypercube

$$\Rightarrow n I_1(\Omega) = \Omega(\log n) = \omega(1).$$ $\square$

Proof ingredient II: area theorem.

For a given linear code C, define

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

(1 represents erasure, 0 represents non-erasure)

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

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

Then:
1. Ω_i is monotone (obvious)
2. Ω_i is symmetric (follows from symmetry of C)
3. $p_\varepsilon(\Omega_i) = \mathbb{P}(Y_{\sim i} \text{ fails to decode } X_i) = h_i(\varepsilon)$
4. $h_i(\varepsilon) \equiv h(\varepsilon)$ (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.

$\rightarrow$ This threshold can only be $\varepsilon^* = 1 - R$, i.e. capacity-achieving!