5.1 Extremization of mutual information for memoryless sources and channels

Theorem 5.1. *(Joint M.I. vs. marginal M.I.)*

(1) If $P_{Y^n|X^n} = \prod P_{Y_i|X_i}$ *then*

$$I(X^n;Y^n) \le \sum I(X_i;Y_i) \tag{5.1}$$

with equality iff $P_{Y^n} = \prod P_{Y_i}$*. Consequently,*

$$\max_{P_{X^n}} I(X^n;Y^n) = \sum_{i=1}^n \max_{P_{X_i}} I(X_i;Y_i).$$

(2) If $X_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n$ *then*

$$I(X^n;Y^n) \ge \sum I(X_i;Y_i) \tag{5.2}$$

with equality iff $P_{X^n|Y^n} = \prod P_{X_i|Y_i}$ P_{Y^n}*-almost surely*[1]*. Consequently,*

$$\min_{P_{Y^n|X^n}} I(X^n;Y^n) = \sum_{i=1}^n \min_{P_{Y_i|X_i}} I(X_i;Y_i).$$

Proof. (1) Use $I(X^n;Y^n) - \sum I(X_j;Y_j) = D(P_{Y^n|X^n} \| \prod P_{Y_i|X_i} | P_{X^n}) - D(P_{Y^n} \| \prod P_{Y_i})$

(2) Reverse the role of X and Y: $I(X^n;Y^n) - \sum I(X_j;Y_j) = D(P_{X^n|Y^n} \| \prod P_{X_i|Y_i} | P_{Y^n}) - D(P_{X^n} \| \prod P_{X_i})$ □

Note: The moral of this result is that

1. For product channel, the MI-maximizing input is a product distribution
2. For product source, the MI-minimizing channel is a product channel

This type of result is often known as **single-letterization** in information theory, which tremendously simplifies the optimization problem over a large-dimensional (multi-letter) problem to a scalar (single-letter) problem. For example, in the simplest case where X^n, Y^n are binary vectors, optimizing $I(X^n;Y^n)$ over P_{X^n} and $P_{Y^n|X^n}$ entails optimizing over 2^n-dimensional vectors and $2^n \times 2^n$ matrices, whereas optimizing each $I(X_i;Y_i)$ individually is easy.
Example:

[1] That is, if $P_{X^n,Y^n} = P_{Y^n} \prod P_{X_i|Y_i}$ as measures.

1. *(5.1) fails for non-product channels.* $X_1 \perp\!\!\!\perp X_2 \sim \text{Bern}(1/2)$ on $\{0,1\} = \mathbb{F}_2$:

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_1 \\ I(X_1;Y_1) &= I(X_2;Y_2) = 0 \quad \text{but} \quad I(X^2;Y^2) = 2 \text{ bits} \end{aligned}$$

2. *Strict inequality in (5.1).*

$$\forall k\ Y_k = X_k = U \sim \text{Bern}(1/2) \quad \Rightarrow \quad \begin{aligned} &I(X_k;Y_k) = 1 \\ &I(X^n;Y^n) = 1 < \sum I(X_k;Y_k) \end{aligned}$$

3. *Strict inequality in (5.2).* $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n$

$$Y_1 = X_2, Y_2 = X_3, \ldots, Y_n = X_1 \quad \Rightarrow \quad \begin{aligned} &I(X_k;Y_k) = 0 \\ &I(X^n;Y^n) = \sum H(X_i) > 0 = \sum I(X_k;Y_k) \end{aligned}$$

5.2* Gaussian capacity via orthogonal symmetry

Multi-dimensional case (WLOG assume $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n$ iid), for a memoryless channel:

$$\max_{\mathbb{E}[\sum X_k^2] \le nP} I(X^n; X^n + Z^n) \le \max_{\mathbb{E}[\sum X_k^2] \le nP} \sum_{k=1}^n I(X_k; X_k + Z_k)$$

Given a distribution $P_{X_1} \cdots P_{X_n}$ satisfying the constraint, form the "average of marginals" distribution $\bar{P}_X = \frac{1}{n}\sum_{k=1}^n P_{X_k}$, which also satisfies the single letter constraint $\mathbb{E}[X^2] = \frac{1}{n}\sum_{k=1}^n \mathbb{E}[X_k^2] \le P$. Then from concavity in P_X of $I(P_X, P_{Y|X})$

$$I(\bar{P}_X; P_{Y|X}) \ge \frac{1}{n}\sum_{k=1}^n I(P_{X_k}, P_{Y|X})$$

So $\bar{P}_X$ gives the same or better MI, which shows that the extremization above ought to have the form $nC(P)$ where $C(P)$ is the single letter capacity. Now suppose $Y^n = X^n + Z_G^n$ where $Z_G^n \sim \mathcal{N}(0, \mathbf{I}_n)$. Since an isotropic Gaussian is rotationally symmetric, for any orthogonal transformation $U \in O(n)$, the additive noise has the same distribution $Z_G^n \sim UZ_G^n$, so that $P_{UY^n|UX^n} = P_{Y^n|X^n}$, and

$$I(P_{X^n}, P_{Y^n|X^n}) = I(P_{UX^n}, P_{UY^n|UX^n}) = I(P_{UX^n}, P_{Y^n|X^n})$$

From the "average of marginal" argument above, averaging over many rotations of X^n can only make the mutual information larger. Therefore, the optimal input distribution P_{X^n} can be chosen to be invariant under orthogonal transformations. Consequently, the (unique!) capacity achieving output distribution $P^*_{Y^n}$ must be rotationally invariant. Furthermore, from the conditions for equality in (5.1) we conclude that $P^*_{Y^n}$ must have independent components. Since the only product distribution satisfying the power constraints and having rotational symmetry is an isotropic Gaussian, we conclude that $P_{Y^n} = (P^*_Y)^n$ and $P^*_Y = \mathcal{N}(0, P\mathbf{I}_n)$.

For the other direction in the Gaussian saddle point problem:

$$\min_{P_N : \mathbb{E}[N^2]=1} I(X_G; X_G + N)$$

This uses the same trick, except here the input distribution is automatically invariant under orthogonal transformations.

5.3 Information measures and probability of error

Let W be a random variable and $\hat{W}$ be our prediction. There are three types of problems:

1. Random guessing: $W \quad \hat{W}$.
2. Guessing with data: $W \to X \to \hat{W}$.
3. Guessing with noisy data: $W \to X \to Y \to \hat{W}$.

We want to draw converse statements, e.g., if the uncertainty of W is high or if the information provided by the data is too little, then it is difficult to guess the value of W.

Theorem 5.2. *Let $|\mathcal{X}| = M < \infty$ and $P_{\max} \triangleq \max_{x \in \mathcal{X}} P_X(x)$. Then*

$$H(X) \le (1 - P_{\max}) \log(M-1) + h(P_{\max}) \triangleq F_M(P_{\max}), \tag{5.3}$$

with equality iff $P_X = (P_{\max}, \underbrace{\tfrac{1-P_{\max}}{M-1}, \ldots, \tfrac{1-P_{\max}}{M-1}}_{M-1})$.

Proof. First proof: Write RHS-LHS as a divergence. Let $P = (P_{\max}, P_2, \ldots, P_M)$ and introduce $Q = (P_{\max}, \frac{1-P_{\max}}{M-1}, \ldots, \frac{1-P_{\max}}{M-1})$. Then RHS-LHS $= D(P\|Q) \ge 0$, with inequality iff $P = Q$.

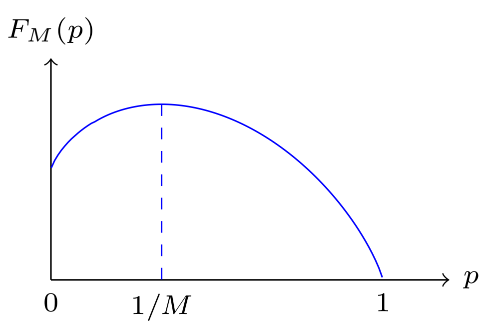
Second proof: Given any $P = (P_{\max}, P_2, \ldots, P_M)$, apply a random permutation π to the last $M-1$ atoms to obtain the distribution P_π. Then averaging P_π over all permutation π gives Q. Then use concavity of entropy or "conditioning reduces entropy": $H(Q) \ge H(P_\pi|\pi) = H(P)$.

Third proof: Directly solve the convex optimization $\max\{H(P) : p_i \le P_{\max}, i = 1, \ldots, M\}$.

Fourth proof: Data processing inequality. Later. □

Note: Similar to Shannon entropy H, $P_{\max}$ is also a reasonable measure for randomness of P. In fact, $\log \frac{1}{P_{\max}}$ is known as the *Rényi entropy of order* ∞, denoted by $H_\infty(P)$. Note that $H_\infty(P) = \log M$ iff P is uniform; $H_\infty(P) = 0$ iff P is a point mass.

Note: The function F_M on the RHS of (5.3) looks like



which is concave with maximum $\log M$ at maximizer $1/M$, but not monotone. However, $P_{\max} \ge \frac{1}{M}$ and F_M is decreasing on $[\frac{1}{M}, 1]$. Therefore (5.3) gives a lower bound on $P_{\max}$ in terms of entropy.

Interpretation: Suppose one is trying to guess the value of X without any information. Then the best bet is obviously the most likely outcome, i.e., the maximal probability of success among all estimators is

$$\max_{\hat{X} \perp\!\!\!\perp X} \mathbb{P}[X = \hat{X}] = P_{\max} \tag{5.4}$$

Thus (5.3) means: It is hard to predict something of large entropy.

Conceptual question: Is it true (for every predictor $\hat{X} \perp\!\!\!\perp X$) that

$$H(X) \le F_M(\mathbb{P}[X = \hat{X}])\,? \tag{5.5}$$

This is not obvious from (5.3) and (5.4) since $p \mapsto F_M(p)$ is not monotone. To show (5.5) consider the data processor $(X, \hat{X}) \mapsto \mathbf{1}_{\{X = \hat{X}\}}$:

$$\begin{array}{rcl} P_{X\hat{X}} & = & P_X P_{\hat{X}} \\ Q_{X\hat{X}} & = & U_X P_{\hat{X}} \end{array} \Rightarrow \begin{array}{rcl} \mathbb{P}[X = \hat{X}] & \triangleq & P_S \\ \mathbb{Q}[X = \hat{X}] & = & \frac{1}{M} \end{array} \Rightarrow \begin{array}{rcl} d\Big(P_S \| \frac{1}{M}\Big) & \le & D(P_{X\hat{X}} \| Q_{X\hat{X}}) \\ & = & \log M - H(X) \end{array}$$

where inequality follows by the data-processing for divergence. ☐

The benefit of this proof is that it trivially generalizes to (possibly randomized) estimators $\hat{X}(Y)$, which depend on some observation Y correlated with X:

Theorem 5.3 (Fano's inequality). *Let $|\mathcal{X}| = M < \infty$ and $X \to Y \to \hat{X}$. Then*

$$H(X|Y) \le F_M(\mathbb{P}[X = \hat{X}(Y)]) = \mathbb{P}[X \ne \hat{X}] \log(M-1) + h(\mathbb{P}[X \ne \hat{X}]). \tag{5.6}$$

Thus, if in addition X is uniform, then

$$I(X;Y) = \log M - H(X|Y) \ge \mathbb{P}[X = \hat{X}] \log M - h(\mathbb{P}[X \ne \hat{X}]). \tag{5.7}$$

Proof. Apply data processing to P_{XY} vs. $U_X P_Y$ and the data processor (kernel) $(X, Y) \mapsto \mathbf{1}_{\{X \ne \hat{X}\}}$ (note that $P_{\hat{X}|Y}$ is fixed). ☐

Remark: We can also derive Fano's Inequality as follows: Let $\epsilon = \mathbb{P}[X \ne \hat{X}]$. Apply data processing for M.I.

$$I(X;Y) \ge I(X;\hat{X}) \ge \min_{P_{Z|X}} \{ I(P_X, P_{Z|X}) : \mathbb{P}[X = Z] \ge 1 - \epsilon \}.$$

This minimum will not be zero since if we force X and Z to agree with some probability, then $I(X;Z)$ cannot be too small. It remains to compute the minimum, which is a nice convex optimization problem. (Hint: look for invariants that the matrix $P_{Z|X}$ must satisfy under permutations $(X, Z) \mapsto (\pi(X), \pi(Z))$ then apply the convexity of $I(P_X, \cdot)$).

Theorem 5.4 (Fano inequality: general). *Let $X, Y \in \mathcal{X}$, $|\mathcal{X}| = M$ and let $Q_{XY} = P_X P_Y$, then*

$$\begin{aligned} I(X;Y) &\ge d(\mathbb{P}[X = Y] \| \mathbb{Q}[X = Y]) \\ &\ge \mathbb{P}[X = Y] \log \frac{1}{\mathbb{Q}[X = Y]} - h(\mathbb{P}[X = Y]) \\ (&= \mathbb{P}[X = Y] \log M - h(\mathbb{P}[X = Y]) \quad \textit{if } P_X \textit{ or } P_Y = \textit{uniform}) \end{aligned}$$

Proof. Apply data processing to P_{XY} and Q_{XY}. Note that if P_X or P_Y = uniform, then $\mathbb{Q}[X = Y] = \frac{1}{M}$ always. ☐

The following result is useful in providing converses for data transmission.

Corollary 5.1 (Lower bound on average probability of error). *Let $W \to X \to Y \to \hat{W}$ and W is uniform on $[M] \triangleq \{1, \ldots, M\}$. Then*

$$P_e \triangleq \mathbb{P}[W \ne \hat{W}] \ge 1 - \frac{I(X;Y) + h(P_e)}{\log M} \tag{5.8}$$

$$\ge 1 - \frac{I(X;Y) + \log 2}{\log M}. \tag{5.9}$$

Proof. Apply Theorem 5.3 and the data processing for M.I.: $I(W;\hat{W}) \le I(X;Y)$. ☐

5.4 Fano, LeCam and minimax risks

In order to show an application to statistical decision theory, consider the following setting:

- Parameter space $\theta \in [0,1]$
- Observation model X_i – i.i.d. Bern(θ)
- Quadratic loss function:
$$\ell(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$
- Fundamental limit:
$$R^*(n) \triangleq \sup_{\theta_0 \in [0,1]} \inf_{\hat{\theta}} \mathbb{E}[(\hat{\theta}(X^n) - \theta)^2 | \theta = \theta_0]$$

A natural estimator to consider is the empirical mean:

$$\hat{\theta}_{emp}(X^n) = \frac{1}{n} \sum_i X_i$$

It achieves the loss

$$\sup_{\theta_0} \mathbb{E}[(\hat{\theta}_{emp} - \theta)^2 | \theta = \theta_0] = \sup_{\theta_0} \frac{\theta_0(1-\theta_0)}{n} = \frac{1}{4n} . \tag{5.10}$$

The question is how close this is to the optimal.

First, recall the *Cramer-Rao lower bound*: Consider an arbitrary statistical estimation problem $\theta \to X \to \hat{\theta}$ with $\theta \in \mathbb{R}$ and $P_{X|\theta}(dx|\theta_0) = f(x|\theta)\mu(dx)$ with $f(x|\theta)$ is differentiable in θ. Then for any $\hat{\theta}(x)$ with $\mathbb{E}[\hat{\theta}(X)|\theta] = \theta + b(\theta)$ and smooth $b(\theta)$ we have

$$\mathbb{E}[(\hat{\theta} - \theta)^2 | \theta = \theta_0] \ge b(\theta_0)^2 + \frac{(1 + b'(\theta_0))^2}{J_F(\theta_0)} , \tag{5.11}$$

where $J_F(\theta_0) = \mathrm{Var}[\frac{\partial \ln f(X|\theta)}{\partial \theta} | \theta = \theta_0]$ is the Fisher information (4.6). In our case, for any *unbiased* estimator (i.e. $b(\theta) = 0$) we have

$$\mathbb{E}[(\hat{\theta} - \theta)^2 | \theta = \theta_0] \ge \frac{\theta_0(1-\theta_0)}{n} ,$$

and we can see from (5.10) that $\hat{\theta}_{emp}$ is optimal in the class of unbiased estimators.

How do we show that biased estimators can not do significantly better? One method is the following. Suppose some estimator $\hat{\theta}$ achieves

$$\mathbb{E}[(\hat{\theta} - \theta)^2 | \theta = \theta_0] \le \Delta_n^2 \tag{5.12}$$

for all θ_0. Then, setup the following probability space:

$$W \to \theta \to X^n \to \hat{\theta} \to \hat{W}$$

- $W \sim$ Bern$(1/2)$
- $\theta = 1/2 + \kappa(-1)^W \Delta_n$ where $\kappa > 0$ is to be specified later
- X^n is i.i.d. Bern(θ)

- $\hat{\theta}$ is the given estimator
- $\hat{W} = 0$ if $\hat{\theta} > 1/2$ and $\hat{W} = 1$ otherwise

The idea here is that we use our high-quality estimator to distinguish between two hypotheses $\theta = 1/2 \pm \kappa\Delta_n$. Notice that for probability of error we have:

$$\mathbb{P}[W \neq \hat{W}] = \mathbb{P}[\hat{\theta} > 1/2 | \theta = 1/2 - \kappa\Delta_n] \leq \frac{\mathbb{E}[(\hat{\theta} - \theta)^2]}{\kappa^2 \Delta_n^2} \leq \frac{1}{\kappa^2}$$

where the last steps are by Chebyshev and (5.12), respectively. Thus, from Theorem 5.3 we have

$$I(W; \hat{W}) \geq \left(1 - \frac{1}{\kappa^2}\right) \log 2 - h(\kappa^{-2}) .$$

On the other hand, from data-processing and golden formula we have

$$I(W; \hat{W}) \leq I(\theta; X^n) \leq D(P_{X^n|\theta} \| \mathrm{Bern}(1/2)^n | P_\theta)$$

Computing the last divergence we get

$$D(P_{X^n|\theta} \| \mathrm{Bern}(1/2)^n | P_\theta) = nd(1/2 - \kappa\Delta_n \| 1/2) = n(\log 2 - h(1/2 - \kappa\Delta_n))$$

As $\Delta_n \to 0$ we have

$$h(1/2 - \kappa\Delta_n) = \log 2 - 2 \log e \cdot (\kappa\Delta_n)^2 + o(\Delta_n^2) .$$

So altogether, we get that for every fixed κ we have

$$\left(1 - \frac{1}{\kappa^2}\right) \log 2 - h(\kappa^{-2}) \leq 2n \log e \cdot (\kappa\Delta_n)^2 + o(n\Delta_n^2) .$$

In particular, by optimizing over κ we get that for some constant $c \approx 0.015 > 0$ we have

$$\Delta_n^2 \geq \frac{c}{n} + o(1/n) .$$

Together with (5.10), we have

$$\frac{0.015}{n} + o(1/n) \leq R^*(n) \leq \frac{1}{4n} ,$$

and thus the empirical-mean estimator is *rate-optimal*.

We mention that for this particular problem (estimating mean of Bernoulli samples) the minimax risk is known exactly:

$$R^*(n) = \frac{1}{4(1 + \sqrt{n})^2}$$

but obtaining this requires rather sophisticated methods. In fact, even showing $R^*(n) = \frac{1}{4n} + o(1/n)$ requires careful priors on θ (unlike the simple two-point prior we used above).[2]

We demonstrated here the essense of the *Fano method* of proving lower (impossibility) bounds in statistical decision theory. Namely, given an estimation task we select a prior on θ which on one hand yields a rather small information $I(\theta; X)$ and on the other hand has sufficiently separated points which thus should be distinguishable by a good estimator. For more see [Yu97].

[2] In fact, getting this result is not hard if one accepts the following *Bayesian Cramer-Rao lower bound*: For any estimator $\hat{\theta}$ and for any prior $\pi(\theta)d\theta$ with smooth density π we have

$$\mathbb{E}_{\theta\sim\pi}[(\hat{\theta}(X) - \theta)^2] \geq \frac{1}{\mathbb{E}[J_F(\theta)] + J_F(\pi)} ,$$

5.5 Entropy rate

Definition 5.1. The entropy rate of a process $\mathbb{X} = (X_1, X_2, \dots)$ is

$$H(\mathbb{X}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(X^n) \tag{5.13}$$

provided the limit exists.

Stationarity is a sufficient condition for entropy rate to exist. Essentially, stationarity means invariance w.r.t. time shift. Formally, $\mathbb{X}$ is stationary if $(X_{t_1}, \dots, X_{t_n}) \overset{D}{=} (X_{t_1+k}, \dots, X_{t_n+k})$ for any $t_1, \dots, t_n, k \in \mathbb{N}$.

Theorem 5.5. *For any stationary process* $\mathbb{X} = (X_1, X_2, \dots)$

1. $H(X_n|X^{n-1}) \le H(X_{n-1}|X^{n-2})$
2. $\frac{1}{n} H(X^n) \ge H(X_n|X^{n-1})$
3. $\frac{1}{n} H(X^n) \le \frac{1}{n-1} H(X^{n-1})$
4. $H(\mathbb{X})$ *exists and* $H(\mathbb{X}) = \lim_{n\to\infty} \frac{1}{n} H(X^n) = \lim_{n\to\infty} H(X_n|X^{n-1})$.
5. *For double-sided process* $\mathbb{X} = (\dots, X_{-1}, X_0, X_1, X_2, \dots)$, $H(\mathbb{X}) = H(X_1|X_{-\infty}^0)$ *provided that* $H(X_1) < \infty$.

Proof.

1. Further conditioning + stationarity: $H(X_n|X^{n-1}) \le H(X_n|X_2^{n-1}) = H(X_{n-1}|X^{n-2})$
2. Using chain rule: $\frac{1}{n} H(X^n) = \frac{1}{n} \sum H(X_i|X^{i-1}) \ge H(X_n|X^{n-1})$
3. $H(X^n) = H(X^{n-1}) + H(X_n|X^{n-1}) \le H(X^{n-1}) + \frac{1}{n} H(X^n)$
4. $n \mapsto \frac{1}{n} H(X^n)$ is a decreasing sequence and lower bounded by zero, hence has a limit $H(\mathbb{X})$. Moreover by chain rule, $\frac{1}{n} H(X^n) = \frac{1}{n} \sum_{i=1}^n H(X_i|X^{i-1})$. Then $H(X_n|X^{n-1}) \to H(\mathbb{X})$. Indeed, from part 1 $\lim_n H(X_n|X^{n-1}) = H'$ exists. Next, recall from calculus: if $a_n \to a$, then the Cesàro's mean $\frac{1}{n} \sum_{i=1}^n a_i \to a$ as well. Thus, $H' = H(\mathbb{X})$.
5. Assuming $H(X_1) < \infty$ we have from (3.10):

$$\lim_{n\to\infty} H(X_1) - H(X_1|X_{-n}^0) = \lim_{n\to\infty} I(X_1; X_{-n}^0) = I(X_1; X_{-\infty}^0) = H(X_1) - H(X_1|X_{-\infty}^0)$$

□

---

where $J_F(\theta)$ is as in (5.11), $J_F(\pi) \triangleq \int \frac{(\pi'(\theta))^2}{\pi(\theta)} d\theta$. Then taking π supported on a $\frac{1}{n^{\frac{1}{4}}}$-neighborhood surrounding a given point θ_0 we get that $\mathbb{E}[J_F(\theta)] = \frac{n}{\theta_0(1-\theta_0)} + o(n)$ and $J_F(\pi) = o(n)$, yielding

$$R^*(n) \ge \frac{\theta_0(1-\theta_0)}{n} + o(1/n)\,.$$

This is a rather general phenomenon: Under regularity assumptions in any iid estimation problem $\theta \to X^n \to \hat{\theta}$ *with quadratic loss* we have

$$R^*(n) = \frac{1}{\inf_\theta J_F(\theta)} + o(1/n)\,.$$

Example: (Stationary processes)

1. $\mathbb{X}$ – iid source $\Rightarrow H(\mathbb{X}) = H(X_1)$

2. $\mathbb{X}$ – mixed sources: Flip a coin with bias p at time $t = 0$, if head, let $\mathbb{X} = \mathbb{Y}$, if tail, let $\mathbb{X} = \mathbb{Z}$. Then $H(\mathbb{X}) = pH(\mathbb{Y}) + \bar{p}H(\mathbb{Z})$.

3. $\mathbb{X}$ – stationary Markov chain : $\quad X_1 \to X_2 \to X_3 \to \cdots$

$$H(X_n|X^{n-1}) = H(X_n|X_{n-1}) \Rightarrow H(\mathbb{X}) = H(X_2|X_1) = \sum_{a,b} \mu(a) P_{b|a} \log \frac{1}{P_{b|a}}$$

   where μ is an invariant measure (possibly non-unique; unique if the chain is ergodic).

4. $\mathbb{X}$– hidden Markov chain : Let $X_1 \to X_2 \to X_3 \to \cdots$ be a Markov chain. Fix $P_{Y|X}$. Let $X_i \xrightarrow{P_{Y|X}} Y_i$. Then $\mathbb{Y} = (Y_1, \ldots)$ is a stationary process. Therefore $H(\mathbb{Y})$ exists but it is very difficult to compute (no closed-form solution to date), even if $\mathbb{X}$ is a binary Markov chain and $P_{Y|X}$ is a BSC.

5.6 Entropy and symbol (bit) error rate

In this section we show that the entropy rates of two processes $\mathbb{X}$ and $\mathbb{Y}$ are close whenever they can be "coupled". Coupling of two processes means defining them on a common probability space so that average distance between their realizations is small. In our case, we will require that the symbol error rate be small, i.e.

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{P}[X_j \neq Y_j] \le \epsilon \,. \tag{5.14}$$

Notice that if we define the Hamming distance as

$$d_H(x^n, y^n) \triangleq \sum_{j=1}^{n} 1\{x_j \neq y_j\}$$

then indeed (5.14) corresponds to requiring

$$\mathbb{E}[d_H(X^n, Y^n)] \le n\epsilon \,.$$

Before showing our main result, we show that Fano's inequality Theorem 5.3 can be tensorized:

Proposition 5.1. *Let X_k take values on a finite alphabet $\mathcal{X}$. Then*

$$H(X^n|Y^n) \le nF_{|\mathcal{X}|}(1 - \delta) \,, \tag{5.15}$$

where

$$\delta = \frac{1}{n} \mathbb{E}[d_H(X^n, Y^n)] = \frac{1}{n} \sum_{j=1}^{n} \mathbb{P}[X_j \neq Y_j] \,.$$

Proof. For each $j \in [n]$ consider $\hat{X}_j(Y^n) = Y_j$. Then from (5.6) we get

$$H(X_j|Y^n) \le F_M(\mathbb{P}[X_j = Y_j]) \,, \tag{5.16}$$

where we denoted $M = |\mathcal{X}|$. Then, upper-bounding joint entropy by the sum of marginals, cf. (1.1), and combining with (5.16) we get

$$H(X^n|Y^n) \le \sum_{j=1}^{n} H(X_j|Y^n) \tag{5.17}$$

$$\le \sum_{j=1}^{n} F_M(\mathbb{P}[X_j = Y_j]) \tag{5.18}$$

$$\le nF_M(\frac{1}{n}\sum_{j=1}^{n}\mathbb{P}[X_j = Y_j])\,. \tag{5.19}$$

where in the last step we used concavity of F_M and Jensen's inequality. Noticing that

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{P}[X_j = Y_j] = 1 - \delta$$

concludes the proof. ☐

Corollary 5.2. *Consider two processes $\mathbb{X}$ and $\mathbb{Y}$ with entropy rates $H(\mathbb{X})$ and $H(\mathbb{Y})$. If*

$$\mathbb{P}[X_j \ne Y_j] \le \epsilon$$

for every j and if $\mathbb{X}$ takes values on a finite alphabet of size M, then

$$H(\mathbb{X}) - H(\mathbb{Y}) \le F_M(1-\epsilon)\,.$$

If both processes have alphabets of size M then

$$|H(\mathbb{X}) - H(\mathbb{Y})| \le \epsilon \log M + h(\epsilon) \to 0 \qquad \text{as } \epsilon \to 0$$

Proof. There is almost nothing to prove:

$$H(X^n) \le H(X^n, Y^n) = H(Y^n) + H(X^n|Y^n)$$

and apply (5.15). For the last statement just recall the expression for F_M. ☐

5.7 Mutual information rate

Definition 5.2 (Mutual information rate)**.**

$$I(\mathbb{X};\mathbb{Y}) = \lim_{n\to\infty}\frac{1}{n}I(X^n;Y^n)$$

provided the limit exists.

Example: *Gaussian processes.* Consider $\mathbb{X}, \mathbb{N}$ two stationary Gaussian processes, independent of each other. Assume that their auto-covariance functions are absolutely summable and thus there exist continuous power spectral density functions f_X and f_N. Without loss of generality, assume all means are zero. Let $c_X(k) = \mathbb{E}[X_1 X_{k+1}]$. Then f_X is the Fourier transform of the auto-covariance function c_X, i.e., $f_X(\omega) = \sum_{k=-\infty}^{\infty} c_X(k)e^{i\omega k}$. Finally, assume $f_N \ge \delta > 0$. Then recall from Lecture 2:

$$\begin{aligned} I(X^n; X^n + N^n) &= \frac{1}{2}\log\frac{\det(\Sigma_{X^n} + \Sigma_{N^n})}{\det \Sigma_{N^n}} \\ &= \frac{1}{2}\sum_{i=1}^{n}\log\sigma_i - \frac{1}{2}\sum_{i=1}^{n}\log\lambda_i\,, \end{aligned}$$

where σ_j, λ_j are the eigenvalues of the covariance matrices $\Sigma_{Y^n} = \Sigma_{X^n} + \Sigma_{N^n}$ and Σ_{N^n}, which are all Toeplitz matrices, e.g., $(\Sigma_{X^n})_{ij} = \mathbb{E}[X_i X_j] = c_X(i-j)$. By Szegö's theorem (see Section 5.8*):

$$\frac{1}{n}\sum_{i=1}^{n} \log \sigma_i \to \frac{1}{2\pi}\int_0^{2\pi} \log f_Y(\omega) d\omega$$

Note that $c_Y(k) = \mathbb{E}[(X_1+N_1)(X_{k+1}+N_{k+1})] = c_X(k) + c_N(k)$ and hence $f_Y = f_X + f_N$. Thus, we have

$$\frac{1}{n} I(X^n; X^n + N^n) \to I(\mathbb{X}; \mathbb{X}+\mathbb{N}) = \frac{1}{4\pi}\int_0^{2\pi} \log \frac{f_X(w) + f_N(\omega)}{f_N(\omega)} d\omega$$

(Note: maximizing this over $f_X(\omega)$ leads to the famous water filling solution $f_X^*(\omega) = |T - f_N(\omega)|^+$.)

5.8* Toeplitz matrices and Szegö's theorem

Theorem 5.6 (Szegö). *Let $f : [0, 2\pi) \to \mathbb{R}$ be the Fourier transform of a summable sequence $\{a_k\}$, that is*

$$f(\omega) = \sum_{k=-\infty}^{\infty} e^{ik\omega} a_k\,, \qquad \sum |a_k| < \infty$$

Then for any $\phi : \mathbb{R} \to \mathbb{R}$ continuous on the closure of the range of f, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \phi(\sigma_{n,j}) = \frac{1}{2\pi}\int_0^{2\pi} \phi(f(\omega)) d\omega\,,$$

where $\{\sigma_{n,j}, j = 1, \ldots, n\}$ are the eigenvalues of the Toeplitz matrix $T_n = \{a_{\ell-m}\}_{\ell,m=1}^{n}$.

Proof sketch. The idea is to approximate ϕ by polynomials, while for polynomials the statement can be checked directly. An alternative interpretation of the strategy is the following: Roughly speaking we want to show that the empirical distribution of the eigenvalues $\frac{1}{n}\sum_{j=1}^{n} \delta_{\sigma_{n,j}}$ converges weakly to the distribution of $f(W)$, where W is uniformly distributed on $[0, 2\pi]$. To this end, let us check that all moments converge. Usually this does not imply weak convergence, but in this case an argument can be made.

For example, for $\phi(x) = x^2$ we have

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} \sigma_{n,j}^2 &= \frac{1}{n} \operatorname{tr} T_n^2 \\
&= \frac{1}{n}\sum_{\ell,m=1}^{n} (T_n)_{\ell,m} (T_n)_{m,\ell} \\
&= \frac{1}{n}\sum_{\ell,m} a_{\ell-m} a_{m-\ell} \\
&= \frac{1}{n}\sum_{\ell=-n-1}^{n-1} (n - |\ell|) a_\ell a_{-\ell} \\
&= \sum_{x \in (-1,1) \cap \frac{1}{n}\mathbb{Z}} (1 - |x|) a_{nx} a_{-nx}\,,
\end{aligned}$$

Substituting $a_\ell = \frac{1}{2\pi}\int_0^{2\pi} f(\omega) e^{i\omega\ell}$ we get

$$\frac{1}{n}\sum_{j=1}^{n} \sigma_{n,j}^2 = \frac{1}{(2\pi)^2} \iint f(\omega) f(\omega') \theta_n(\omega - \omega')\,, \tag{5.20}$$

where

$$\theta_n(u) = \sum_{x \in (-1,1) \cap \frac{1}{n}\mathbb{Z}} (1 - |x|) e^{-inux}$$

is a Fejer kernel and converges to a δ-function: $\theta_n(u) \to 2\pi\delta(u)$ (in the sense of convergence of Schwartz distributions). Thus from (5.20) we get

$$\frac{1}{n} \sum_{j=1}^{n} \sigma_{n,j}^2 \to \frac{1}{(2\pi)^2} \iint f(\omega) f(\omega') 2\pi\delta(\omega - \omega') d\omega d\omega' = \frac{1}{2\pi} \int_0^{2\pi} f^2(\omega) d\omega$$

as claimed. $\square$