LECTURE NOTES ON INFORMATION THEORY

Preface

> “There is a whole book of readymade, long and convincing, lavishly composed telegrams for all occasions. Sending such a telegram costs only twenty-five cents. You see, what gets transmitted over the telegraph is *not the text* of the telegram, but simply *the number* under which it is listed in the book, and the signature of the sender. This is quite a funny thing, reminiscent of Drugstore Breakfast #2. Everything is served up in a ready form, and the customer is totally freed from the unpleasant necessity to think, and to spend money on top of it.”
>
> *Little Golden America.* Travelogue by I. Ilf and E. Petrov, 1937.
>
> [Pre-Shannon encoding, courtesy of M. Raginsky]

These notes are a graduate-level introduction to the mathematics of Information Theory. They were created by Yury Polyanskiy and Yihong Wu, who used them to teach at MIT (2012, 2013 and 2016) and UIUC (2013, 2014). The core structure and flow of material is largely due to Prof. Sergio Verdú, whose wonderful class at Princeton University [Ver07] shaped up our own perception of the subject. Specifically, we follow Prof. Verdú’s style in relying on single-shot results, Feinstein’s lemma and information spectrum methods. We have added a number of technical refinements and new topics, which correspond to our own interests (e.g., modern aspects of finite blocklength results and applications of information theoretic methods to statistical decision theory).

Compared to the more popular “typicality” and “method of types” approaches (as in Cover-Thomas [CT06] and Csiszár-Körner [CK81]), these notes prepare the reader to consider delay-constraints (“non-asymptotics”) and to simultaneously treat continuous and discrete sources/channels.

We are especially thankful to Dr. O. Ordentlich, who contributed a lecture on lattice codes. Initial version was typed by Qingqing Huang and Austin Collins,who also created many graphics. Rachel Cohen have also edited various parts. Aolin Xu,Pengkun Yang and Ganesh Ajjanagadde have contributed suggestions and corrections to the content.
We are indebted to all of them.

Y. Polyanskiy
Y. Wu

27 Feb 2015

Contents

Part I

Information measures

§ 1. Information measures: entropy and divergence

Review: Random variables

- Two methods to describe a random variable (R.V.) X:
  1. a function $X : \Omega \to \mathcal{X}$ from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a target space $\mathcal{X}$.
  2. a distribution P_X on some measurable space $(\mathcal{X}, \mathcal{F})$.
- Convention: capital letter – RV (e.g. X); small letter – realization (e.g. x_0).
- X — discrete if there exists a countable set $\mathcal{X} = \{x_j, j = 1, \ldots\}$ such that $\sum_{j=1}^{\infty} P_X(x_j) = 1$. $\mathcal{X}$ is called alphabet of X, $x \in \mathcal{X}$ – atoms and $P_X(x_j)$ – probability mass function (pmf).
- For discrete RV support $\text{supp} P_X = \{x : P_X(x) > 0\}$.
- Vector RVs: $X_1^n \triangleq (X_1, \ldots, X_n)$. Also denoted just X^n.
- For a vector RV X^n and $S \subset \{1, \ldots, n\}$ we denote $X_S = \{X_i, i \in S\}$.

1.1 Entropy

Definition 1.1 (Entropy)**.** For a discrete R.V. X with distribution P_X:

$$\begin{aligned} H(X) &= \mathbb{E}\Big[\log \frac{1}{P_X(X)}\Big] \\ &= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} \;. \end{aligned}$$

Definition 1.2 (Joint entropy)**.** $X^n = (X_1, X_2, \ldots, X_n)$ – a random vector with n components.

$$H(X^n) = H(X_1, \ldots, X_n) = \mathbb{E}\Big[\log \frac{1}{P_{X_1, \ldots, X_n}(X_1, \ldots, X_n)} .\Big]$$

Definition 1.3 (Conditional entropy)**.**

$$H(X|Y) = \mathbb{E}_{y \sim P_Y}[H(P_{X|Y=y})] = \mathbb{E}\Big[\log \frac{1}{P_{X|Y}(X|Y)}\Big],$$

i.e., the entropy of $H(P_{X|Y=y})$ averaged over P_Y.

Note:

- Q: Why such definition, why log, why entropy?
  Name comes from thermodynamics. Definition is justified by theorems in this course (e.g. operationally by compression), but also by a number of experiments. For example, we can measure time it takes for ants-scouts to describe location of the food to ants-workers. It was found that when nest is placed at a root of a full binary tree of depth d and food at one of the leaves, the time was proportional to $\log 2^d = d$ – entropy of the random variable describing food location. It was estimated that ants communicate with about $0.7 - 1$ bit/min. Furthermore, communication time reduces if there are some regularities in path-description (e.g., paths like "left,right,left,right,left,right" were described faster). See [RZ86] for more.

- We agree that $0 \log \frac{1}{0} = 0$ (by continuity of $x \mapsto x \log \frac{1}{x}$)

- Also write $H(P_X)$ instead of $H(X)$ (abuse of notation, as customary in information theory).

- Basis of log — units

$\log_2$	$\leftrightarrow$	bits
$\log_e$	$\leftrightarrow$	nats
$\log_{256}$	$\leftrightarrow$	bytes
$\log$	$\leftrightarrow$	arbitrary units, base always matches exp

Example (Bernoulli): $X \in \{0,1\}, \quad \mathbb{P}[X=1] = P_X(1) \triangleq p$

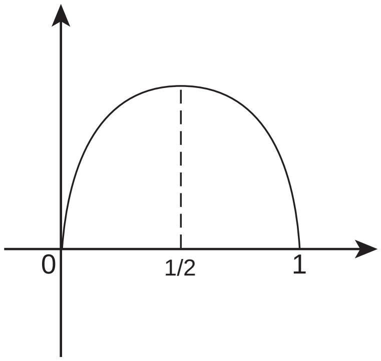
$$H(X) = h(p) \triangleq p \log \frac{1}{p} + \bar{p} \log \frac{1}{\bar{p}}$$

where $h(\cdot)$ is called the **binary entropy function**.

Proposition 1.1. *$h(\cdot)$ is continuous, concave on $[0,1]$ and*

$$h'(p) = \log \frac{\bar{p}}{p}$$

with infinite slope at 0 and 1.



Example (Geometric): $X \in \{0,1,2,\dots\} \quad \mathbb{P}[X=i] = P_x(i) = p \cdot (\bar{p})^i$

$$\begin{aligned} H(X) &= \sum_{i=0}^{\infty} p \cdot \bar{p}^i \log \frac{1}{p \cdot \bar{p}^i} = \sum_{i=0}^{\infty} p\bar{p}^i \Big(i \log \frac{1}{\bar{p}} + \log \frac{1}{p} \Big) \\ &= \log \frac{1}{p} + p \cdot \log \frac{1}{\bar{p}} \cdot \frac{1-p}{p^2} = \frac{h(p)}{p} \end{aligned}$$

Example (Infinite entropy): Can $H(X) = +\infty$? Yes, $\mathbb{P}[X=k] = \frac{c}{k \ln^2 k}, k = 2, 3, \cdots$

Review: Convexity

- *Convex set*: A subset S of some vector space is convex if $x, y \in S \Rightarrow \alpha x + \bar{\alpha} y \in S$ for all $\alpha \in [0,1]$. (Notation: $\bar{\alpha} \triangleq 1 - \alpha$.)

  e.g., unit interval $[0,1]$; $S = \{\text{probability distributions on } \mathcal{X}\}$, $S = \{P_X : \mathbb{E}[X] = 0\}$.

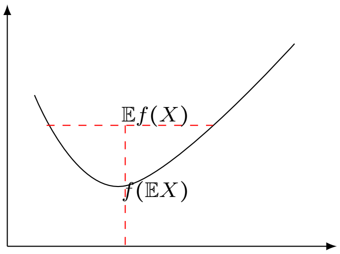
- *Convex function*: $f : S \to \mathbb{R}$ is

  - convex if $f(\alpha x + \bar{\alpha} y) \le \alpha f(x) + \bar{\alpha} f(y)$ for all $x, y \in S, \alpha \in [0,1]$.
  - strictly convex if $f(\alpha x + \bar{\alpha} y) < \alpha f(x) + \bar{\alpha} f(y)$ for all $x \neq y \in S, \alpha \in (0,1)$.
  - (strictly) concave if $-f$ is (strictly) convex.

  e.g., $x \mapsto x \log x$ is strictly convex; the mean $P \mapsto \int x \mathrm{d}P$ is convex but not strictly convex, variance is concave (Q: is it strictly concave? Think of zero-mean distributions.).

- *Jensen's inequality*: For any S-valued random variable X

  - f is convex $\Rightarrow f(\mathbb{E}X) \le \mathbb{E}f(X)$
  - f is strictly convex $\Rightarrow f(\mathbb{E}X) < \mathbb{E}f(X)$ unless X is a constant ($X = \mathbb{E}X$ a.s.)



Famous puzzle: A man says, "I am the average height and average weight of the population. Thus, I am an average man." However, he is still considered to be a little overweight. Why?
Answer: The weight is roughly proportional to the volume, which is roughly proportional to the third power of the height. Let P_X denote the distribution of the height among the population. So by Jensen's inequality, since $x \mapsto x^3$ is strictle convex on $x > 0$, we have $(\mathbb{E}X)^3 < \mathbb{E}X^3$, regardless of the distribution of X.
Source: [Yos03, Puzzle 94] or online [Har].

Theorem 1.1. *Properties of H:*

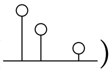
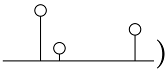
1. *(Positivity) $H(X) \ge 0$ with equality iff $X = x_0$ a.s. for some $x_0 \in \mathcal{X}$.*
2. *(Uniform maximizes entropy) $H(X) \le \log |\mathcal{X}|$, with equality iff X is uniform on $\mathcal{X}$.*
3. *(Invariance under relabeling) $H(X) = H(f(X))$ for any bijective f.*
4. *(Conditioning reduces entropy)*
$$H(X|Y) \le H(X), \quad \text{with equality iff } X \text{ and } Y \text{ are independent.}$$

5. *(Small chain rule)*
$$H(X,Y) = H(X) + H(Y|X) \le H(X) + H(Y)$$

6. *(Entropy under functions) $H(X) = H(X, f(X)) \ge H(f(X))$ with equality iff f is one-to-one on the support of P_X,*

7. *(Full chain rule)*
$$H(X_1,\dots,X_n) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \le \sum_{i=1}^{n} H(X_i), \tag{1.1}$$
$$\uparrow \textit{ equality iff } X_1,\dots,X_n \textit{ mutually independent} \tag{1.2}$$

Proof. 1. Expectation of non-negative function
2. Jensen's inequality
3. H only depends on the values of P_X, not locations:

$H($  $) = H($  $)$

4. Later (Lecture 2)
5. $\mathbb{E}\log\frac{1}{P_{XY}(X,Y)} = \mathbb{E}\Big[\log\frac{1}{P_X(X)\cdot P_{Y|X}(Y|X)}\Big]$
6. *Intuition:* $(X, f(X))$ contains the same amount of information as X. Indeed, $x \mapsto (x, f(x))$ is 1-1. Thus by 3 and 5:
$$H(X) = H(X, f(X)) = H(f(X)) + H(X|f(X)) \ge H(f(X))$$
The bound is attained iff $H(X|f(X)) = 0$ which in turn happens iff X is a *constant* given $f(X)$.
7. Telescoping:
$$P_{X_1X_2\cdots X_n} = P_{X_1}P_{X_2|X_1}\cdots P_{X_n|X^{n-1}}$$
□

Note: To give a preview of the *operational meaning* of entropy, let us play the following game. We are allowed to make queries about some unknown discrete R.V. X by asking yes-no questions. The objective of the game is to guess the realized value of the R.V. X. For example, $X \in \{a,b,c,d\}$ with $\mathbb{P}[X=a] = 1/2$, $\mathbb{P}[X=b] = 1/4$, and $\mathbb{P}[X=c] = \mathbb{P}[X=c] = 1/8$. In this case, we can ask "$X = a$?". If not, proceed by asking "$X = b$?". If not, ask "$X = c$?", after which we will know for sure the realization of X. The resulting average number of questions is $1/2 + 1/4\times 2 + 1/8\times 3 + 1/8\times 3 = 1.75$, which equals $H(X)$ in bits. It turns out (chapter 2) that the minimal average number of yes-no questions to pin down the value of X is always between $H(X)$ `bits` and $H(X)+1$ `bits`. In this special case the above scheme is optimal because (intuitively) it always splits the probability in half.

1.1.1 Entropy: axiomatic characterization

One might wonder why entropy is defined as $H(P) = \sum p_i \log\frac{1}{p_i}$ and if there are other definitions. Indeed, the information-theoretic definition of entropy is related to entropy in statistical physics. Also, it arises as answers to specific operational problems, e.g., the minimum average number of bits to describe a random variable as discussed above. Therefore it is fair to say that it is not pulled out of thin air.

Shannon has also showed that entropy can be defined *axiomatically*, as a function satisfying several natural conditions. Denote a probability distribution on m letters by $P = (p_1,\dots,p_m)$ and consider a functional $H_m(p_1,\dots,p_m)$. If H_m obeys the following axioms:

a) Permutation invariance

b) Expansible: $H_m(p_1, \dots, p_{m-1}, 0) = H_{m-1}(p_1, \dots, p_{m-1})$.

c) Normalization: $H_2(\frac{1}{2}, \frac{1}{2}) = \log 2$.

d) Continuity: $H_2(p, 1-p) \to 0$ as $p \to 0$.

e) Subadditivity: $H(X,Y) \le H(X) + H(Y)$. Equivalently, $H_{mn}(r_{11}, \dots, r_{mn}) \le H_m(p_1, \dots, p_m) + H_n(q_1, \dots, q_n)$ whenever $\sum_{j=1}^n r_{ij} = p_i$ and $\sum_{i=1}^m r_{ij} = q_j$.

f) Additivity: $H(X,Y) = H(X)+H(Y)$ if $X \perp\!\!\!\perp Y$. Equivalently, $H_{mn}(p_1q_1, \dots, p_mq_n) \le H_m(p_1, \dots, p_m) + H_n(q_1, \dots, q_n)$.

then $H_m(p_1, \dots, p_m) = \sum_{i=1}^m p_i \log \frac{1}{p_i}$ is the only possibility. The interested reader is referred to [CT06, p. 53] and the reference therein.

1.1.2 History of entropy

In the early days of industrial age, engineers wondered if it is possible to construct a perpetual motion machine. After many failed attempts, a law of conservation of energy was postulated: a machine cannot produce more work than the amount of energy it consumed from the ambient world (this is also called the *first law* of thermodynamics). The next round of attempts was then to construct a machine that would draw energy in the form of heat from a warm body and convert it to equal (or approximately equal) amount of work. An example would be a steam engine. However, again it was observed that all such machines were highly inefficiencient, that is the amount of work produced by absorbing heat Q was $\ll Q$. The remainder of energy was dissipated to the ambient world in the form of heat. Again after many rounds of attempting various designs Clausius and Kelvin proposed another law:

> *Second law* of thermodynamics: There does not exist a machine that operates in a cycle (i.e. returns to its original state periodically), produces useful work and whose only other effect on the outside world is drawing heat from a warm body. (That is, every such machine, should expend some amount of heat to some cold body too!)[1]

Equivalent formulation is: There does not exist a cyclic process that transfers heat from a cold body to a warm body (that is, every such process needs to be helped by expending some amount of external work).

Notice that there is something annoying about the second law as compared to the first law. In the first law there is a quantity that is conserved, and this is somehow logically easy to accept. The second law seems a bit harder to believe in (and some engineers did not, and only their recurrent failures to circumvent it finally convinced them). So Clausius, building on an ingenious work of S. Carnot, figured out that there is an "explanation" to why any cyclic machine should expend heat. He proposed that there must be some hidden quantity associated to the machine, entropy of it (translated as transformative content), whose value must return to its original state. Furthermore, under any reversible (i.e. quasi-stationary, or "very slow") process operated on this machine the change of entropy is proportional to the ratio of absorbed heat and the temperature of the machine:

$$\Delta S = \frac{\Delta Q}{T} . \tag{1.3}$$

[1] Note that the reverse effect (that is converting work into heat) is rather easy: friction is an example.

So that if heat Q is absorbed at temperature T_{hot} then to return to the original state, one must return some Q' amount of heat. Q' can be significantly smaller than Q if Q' is returned at temperature $T_{cold} < T_{hot}$. Further logical arguments can convince one that for irreversible cyclic process the change of entropy at the end of the cycle can only be positive, and hence *entropy cannot reduce*.

There were a great many experimentally verified consequences that second law produced. However, what is surprising is that the mysterious entropy did not have any formula for it (unlike say energy), and thus had to be computed indirectly on the basis of relation (1.3). This was changed with the revolutionary work of Boltzmann and Gibbs, who showed that for a system of n particles the entropy of a given macro-state can be computed as

$$S = kn \sum_{j=1}^{\ell} p_j \log \frac{1}{p_j} \,,$$

where k is the Boltzmann constant, we assume that each particle can only be in one of ℓ molecular states (e.g. spin up/down, or if we quantize the phase volume into ℓ subcubes) and p_j is the fraction of particles in j-th molecular state.

1.1.3* Entropy: submodularity

Recall that $[n]$ denotes a set $\{1, \ldots, n\}$, $\binom{S}{k}$ denotes subsets of S of size k and 2^S denotes all subsets of S. A set function $f : 2^S \to \mathbb{R}$ is called submodular if for any $T_1, T_2 \subset S$

$$f(T_1 \cup T_2) + f(T_1 \cap T_2) \le f(T_1) + f(T_2)$$

Submodularity is similar to concavity, in the sense that "adding elements gives diminishing returns". Indeed consider $T' \subset T$ and $b \notin T$. Then

$$f(T \cup b) - f(T) \le f(T' \cup b) - f(T') \,.$$

Theorem 1.2. *Let X^n be discrete RV. Then $T \mapsto H(X_T)$ is submodular.*

Proof. Let $A = X_{T_1 \setminus T_2}, B = X_{T_1 \cap T_2}, C = X_{T_2 \setminus T_1}$. Then we need to show

$$H(A, B, C) + H(B) \le H(A, B) + H(B, C) \,.$$

This follows from a simple chain

$$\begin{aligned} H(A, B, C) + H(B) &= H(A, C|B) + 2H(B) & (1.4)\\ &\le H(A|B) + H(C|B) + 2H(B) & (1.5)\\ &= H(A, B) + H(B, C) & (1.6) \end{aligned}$$

□

Note that entropy is not only submodular, but also monotone:

$$T_1 \subset T_2 \implies H(X_{T_1}) \le H(X_{T_2}) \,.$$

So fixing n, let us denote by Γ_n the set of all non-negative, monotone, submodular set-functions on $[n]$. Note that via an obvious enumeration of all non-empty subsets of $[n]$, Γ_n is a closed convex cone in $\mathbb{R}_+^{2^n-1}$. Similarly, let us denote by Γ_n^* the set of all set-functions corresponding to distributions on X^n. Let us also denote $\bar{\Gamma}^*_n$ the closure of Γ^*_n. It is not hard to show, cf. [ZY97], that $\bar{\Gamma}^*_n$ is also a closed convex cone and that

$$\Gamma^*_n \subset \bar{\Gamma}^*_n \subset \Gamma_n \,.$$

The astonishing result of [ZY98] is that

$$\Gamma^*_2 = \bar{\Gamma}^*_2 = \Gamma_2 \tag{1.7}$$
$$\Gamma^*_3 \subsetneq \bar{\Gamma}^*_3 = \Gamma_3 \tag{1.8}$$
$$\Gamma^*_n \subsetneq \bar{\Gamma}^*_n \subsetneq \Gamma_n \qquad n \ge 4\,. \tag{1.9}$$

This follows from the fundamental new information inequality not implied by the submodularity of entropy (and thus called *non-Shannon inequality*). Namely, [ZY98] shows that for any 4 discrete random variables:

$$I(X_3;X_4) - I(X_3;X_4|X_1) - I(X_3;X_4|X_2) \le \frac{1}{2}I(X_1;X_2) + \frac{1}{4}I(X_1;X_3,X_4) + \frac{1}{4}I(X_2;X_3,X_4)\,.$$

(see Definition 2.3).

1.1.4 Entropy: Han's inequality

Theorem 1.3 (Han's inequality)**.** *Let X^n be discrete n-dimensional RV and denote $\bar{H}_k(X^n) = \frac{1}{\binom{n}{k}}\sum_{T\subset\binom{[n]}{k}} H(X_T)$ – the average entropy of a k-subset of coordinates. Then $\frac{\bar{H}_k}{k}$ is decreasing in k:*

$$\frac{1}{n}\bar{H}_n \le \cdots \le \frac{1}{k}\bar{H}_k \cdots \le \bar{H}_1\,. \tag{1.10}$$

Furthermore, the sequence $\bar{H}_k$ is increasing and concave in the sense of decreasing slope:

$$\bar{H}_{k+1} - \bar{H}_k \le \bar{H}_k - \bar{H}_{k-1}\,. \tag{1.11}$$

Proof. Denote for convenience $\bar{H}_0 = 0$. Note that $\frac{\bar{H}_m}{m}$ is an average of differences:

$$\frac{1}{m}\bar{H}_m = \frac{1}{m}\sum_{k=1}^{m}(\bar{H}_k - \bar{H}_{k-1})$$

Thus, it is clear that (1.11) implies (1.10) since increasing m by one adds a smaller element to the average. To prove (1.11) observe that from submodularity

$$H(X_1,\ldots,X_{k+1}) + H(X_1,\ldots,X_{k-1}) \le H(X_1,\ldots,X_k) + H(X_1,\ldots,X_{k-1},X_{k+1})\,.$$

Now average this inequality over all $n!$ permutations of indices $\{1,\ldots,n\}$ to get

$$\bar{H}_{k+1} + \bar{H}_{k-1} \le 2\bar{H}_k$$

as claimed by (1.11).

Alternative proof: Notice that by "conditioning decreases entropy" we have

$$H(X_{k+1}|X_1,\ldots,X_k) \le H(X_{k+1}|X_2,\ldots,X_k)\,.$$

Averaging this inequality over all permutations of indices yields (1.11). □

Note: Han's inequality holds for any submodular set-function.
Example: Another submodular set-function is

$$S \mapsto I(X_S; X_{S^c}) \,.$$

Han's inequality for this one reads

$$0 = \frac{1}{n} I_n \le \cdots \le \frac{1}{k} I_k \cdots \le I_1 \,,$$

where $I_k = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} I(X_S; X_{S^c})$ – gauges the amount of k-subset coupling in the random vector X^n.

1.2 Divergence

Review: Measurability

In this course we will assume that all alphabets are standard Borel spaces. Some of the nice properties of standard Borel spaces:

- all complete separable metric spaces, endowed with Borel σ-algebras are standard Borel. In particular, countable alphabets and $\mathbb{R}^n$ and $\mathbb{R}^\infty$ (space of sequences) are standard Borel.
- if $\mathcal{X}_i, i = 1, \ldots$ are s.B.s. then so is $\prod_{i=1}^\infty \mathcal{X}_i$
- singletons $\{x\}$ are measurable sets
- diagonal $\Delta = \{(x,x) : x \in \mathcal{X}\}$ is measurable in $\mathcal{X} \times \mathcal{X}$
- (Most importantly) for any probability distribution $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$ there exists a transition probability kernel (also called a regular branch of a conditional distribution) $P_{Y|X}$ s.t.
$$P_{X,Y}[E] = \int_{\mathcal{X}} P_X(dx) \int_{\mathcal{Y}} P_{Y|X=x}(dy) 1\{(x,y) \in E\} \,.$$

Intuition: $D(P\|Q)$ gauges the **dissimilarity** between P and Q.

Definition 1.4 (Divergence)**.** Let P, Q be distributions on

- $\mathcal{A}$ = discrete alphabet (finite or countably infinite)

$$D(P\|Q) \triangleq \sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)} \,,$$

where we agree:

(1) $0 \cdot \log \frac{0}{0} = 0$

(2) $\exists a : Q(a) = 0, P(a) > 0 \Rightarrow D(P\|Q) = \infty$

- $\mathcal{A} = \mathbb{R}^k$, P and Q have densities f_P and f_Q

$$D(P\|Q) = \begin{cases} \int_{\mathbb{R}^k} \log \frac{f_P(x^k)}{f_Q(x^k)} f_P(x^k) dx^k & , \text{ Leb}\{f_P > 0, f_Q = 0\} = 0 \\ +\infty & , \text{ otherwise} \end{cases}$$

- $\mathcal{A}$ — measurable space:

$$D(P\|Q) = \begin{cases} \mathbb{E}_Q \frac{dP}{dQ} \log \frac{dP}{dQ} = \mathbb{E}_P \log \frac{dP}{dQ} & , \quad P \ll Q \\ +\infty & , \quad \text{otherwise} \end{cases}$$

(Also known as information divergence, Kullback–Leibler divergence, relative entropy.)

Notes:

- (Radon-Nikodym theorem) Recall that for two measures P and Q, we say P is absolutely continuous w.r.t. Q (denoted by $P \ll Q$) if $Q(E) = 0$ implies $P(E) = 0$ for all measurable E. If $P \ll Q$, then there exists a function $f : \mathcal{X} \to \mathbb{R}_+$ such that for any measurable set E,

  $$P(E) = \int_E f \mathrm{d}Q. \qquad \text{[change of measure]}$$

  Such f is called a density (or a Radon-Nikodym derivative) of P w.r.t. Q, denoted by $\frac{\mathrm{d}P}{\mathrm{d}Q}$. For finite alphabets, we can just take $\frac{dP}{dQ}(x)$ to be the ratio of the pmfs. For P and Q on $\mathbb{R}^n$ possessing pdfs we can take $\frac{dP}{dQ}(x)$ to be the ratio of pdfs.

- (Infinite values) $D(P\|Q)$ can be ∞ also when $P \ll Q$, but the two cases of $D(P\|Q) = +\infty$ are consistent since $D(P\|Q) = \sup_\Pi D(P_\Pi\|Q_\Pi)$, where Π is a finite partition of the underlying space $\mathcal{A}$ (proof: later)

- (Asymmetry) $D(P\|Q) \neq D(Q\|P)$. Asymmetry can be very useful. Example: $P(H) = P(T) = 1/2$, $Q(H) = 1$. Upon observing HHHHHHH, one tends to believe it is Q but can never be absolutely sure; Upon observing HHT, know for sure it is P. Indeed, $D(P\|Q) = \infty$, $D(Q\|P) = 1\,\texttt{bit}$.

- (Pinsker's inequality) There are many other measures for dissimilarity, e.g., total variation (L_1-distance)

  $$\begin{aligned} \mathrm{TV}(P,Q) &\triangleq \sup_E P[E] - Q[E] && (1.12) \\ &= \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q| \qquad = \text{(discrete case)} \frac{1}{2}\sum_x |P(x) - Q(x)|\,. && (1.13) \end{aligned}$$

  This one is symmetric. There is a famous Pinsker's (or Pinsker-Csiszár) inequality relating D and TV:

  $$\mathrm{TV}(P,Q) \le \sqrt{\frac{1}{2\log e} D(P\|Q)}\,. \qquad (1.14)$$

- (Other divergences) A general class of divergence-like measures was proposed by Csiszár. Fixing a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(1) = 0$ we define *f-divergence* D_f as

  $$D_f(P\|Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]. \qquad (1.15)$$

This encompasses total variation, χ^2-distance, Hellinger, Tsallis etc. Inequalities between various f-divergences such as (1.14) was once an active field of research. It was made largely irrelevant by a work of Harremoës and Vajda [HV11] giving a simple method for obtaining best possible inequalities between any two f-divergences.

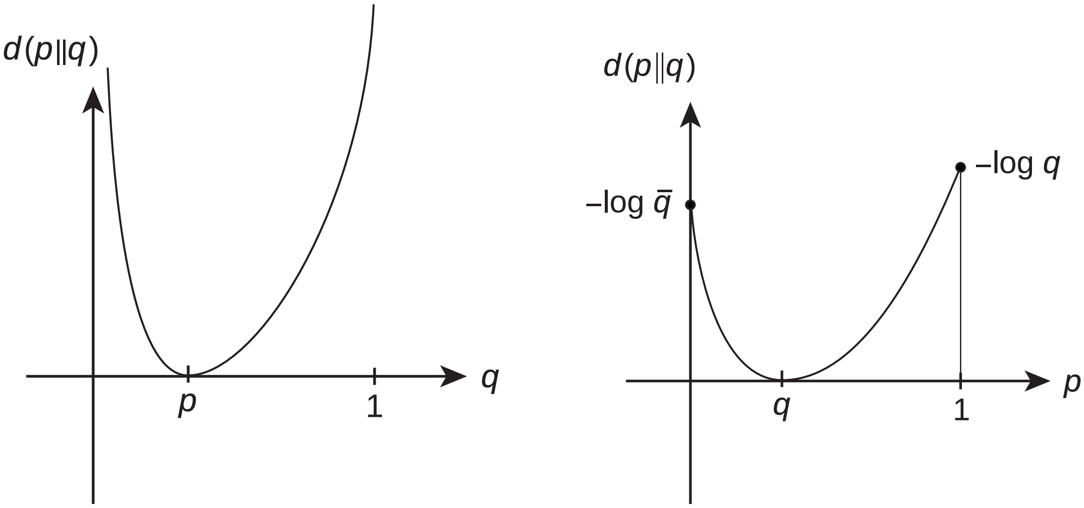
Theorem 1.4 (H v.s. D)**.** *If distribution P is supported on $\mathcal{A}$ with $|\mathcal{A}| < \infty$, then*

$$H(P) = \log|\mathcal{A}| - D(P\| \underbrace{U_{\mathcal{A}}}_{\textit{uniform distribution on } \mathcal{A}}).$$

Example (Binary divergence): $\mathcal{A} = \{0,1\}$; $P = [p, \bar{p}]$; $Q = [q, \bar{q}]$

$$D(P\|Q) = d(p\|q) \triangleq p\log\frac{p}{q} + \bar{p}\log\frac{\bar{p}}{\bar{q}}$$

Here is how $d(p\|q)$ depends on p and q:



Quadratic lower bound (homework):

$$d(p\|q) \geq 2(p-q)^2 \log e$$

Example (Real Gaussian): $\mathcal{A} = \mathbb{R}$

$$D(\mathcal{N}(m_1, \sigma_1^2)\|\mathcal{N}(m_0, \sigma_0^2)) = \frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2} + \frac{1}{2}\Big[\frac{(m_1 - m_0)^2}{\sigma_0^2} + \frac{\sigma_1^2}{\sigma_0^2} - 1\Big]\log e \tag{1.16}$$

Example (Complex Gaussian): $\mathcal{A} = \mathbb{C}$. The pdf of $\mathcal{N}_c(m, \sigma^2)$ is $\frac{1}{\pi\sigma^2}e^{-|x-m|^2/\sigma^2}$, or equivalently:

$$\mathcal{N}_c(m, \sigma^2) = \mathcal{N}\left(\begin{bmatrix}\mathrm{Re}(m) & \mathrm{Im}(m)\end{bmatrix}, \begin{bmatrix}\sigma^2/2 & 0 \\ 0 & \sigma^2/2\end{bmatrix}\right) \tag{1.17}$$

$$D(\mathcal{N}_c(m_1, \sigma_1^2)\|\mathcal{N}_c(m_0, \sigma_0^2)) = \log\frac{\sigma_0^2}{\sigma_1^2} + \Big[\frac{|m_1 - m_0|^2}{\sigma_0^2} + \frac{\sigma_1^2}{\sigma_0^2} - 1\Big]\log e \tag{1.18}$$

Example (Vector Gaussian): $\mathcal{A} = \mathbb{C}^k$

$$\begin{aligned} D(\mathcal{N}_c(m_1, \Sigma_1)\|\mathcal{N}_c(m_0, \Sigma_0)) \quad = \quad & \log\det\Sigma_0 - \log\det\Sigma_1 + (m_1 - m_0)^H \Sigma_0^{-1}(m_1 - m_0)\log e \\ & + \mathrm{tr}(\Sigma_0^{-1}\Sigma_1 - I)\log e \end{aligned}$$

(assume $\det \Sigma_0 \neq 0$).

Note: The definition of $D(P\|Q)$ extends verbatim to measures P and Q (not necessarily probability measures), in which case $D(P\|Q)$ can be negative. A sufficient condition for $D(P\|Q) \geq 0$ is that P is a probability measure and Q is a sub-probability measure, i.e., $\int \mathrm{d}Q \leq 1 = \int \mathrm{d}P$.

1.3 Differential entropy

The notion of *differential entropy* is simply the divergence with respect to the Lebesgue measure:

Definition 1.5. The differential entropy of a random vector X^k is

$$h(X^k) = h(P_{X^k}) \triangleq -D(P_{X^k}\|\mathrm{Leb}). \tag{1.19}$$

In particular, if X^k has probability density function (pdf) p, then $h(X^k) = \mathbb{E}\log\frac{1}{p(X^k)}$; otherwise $h(X^k) = -\infty$. Conditional differential entropy $h(X^k|Y) \triangleq \mathbb{E}\log\frac{1}{p_{X^k|Y}(X^k|Y)}$ where $p_{X^k|Y}$ is a conditional pdf.

Warning: Even for X with pdf $h(X)$ can be positive, negative, take values of $\pm\infty$ or even be undefined[2].

Nevertheless, differential entropy shares many properties with the usual entropy:

Theorem 1.5 (Properties of differential entropy)**.** *Assume that all differential entropies appearing below exists and are finite (in particular all RVs have pdfs and conditional pdfs). Then the following hold :*

1. *(Uniform maximizes diff. entropy) If $\mathbb{P}[X^n \in S] = 1$ then $h(X^n) \leq \mathrm{Leb}\{S\}$ with equality iff X^n is uniform on S.*
2. *(Conditioning reduces diff. entropy) $h(X|Y) \leq h(X)$ (here Y could be arbitrary, e.g. discrete)*
3. *(Chain rule)*
$$h(X^n) = \sum_{k=1}^{n} h(X_k|X^{k-1}).$$
4. *(Submodularity) The set-function $T \mapsto h(X_T)$ is submodular.*
5. *(Han's inequality) The function $k \mapsto \frac{1}{k\binom{n}{k}}\sum_{T\in\binom{[n]}{k}} h(X_T)$ is decreasing in k.*

1.3.1 Application of differential entropy: Loomis-Whitney and Bollobás-Thomason

The following famous result shows that n-dimensional rectangle simultaneously minimizes volumes of all projections:[3]

Theorem 1.6 (Bollobás-Thomason Box Theorem)**.** *Let $K \subset \mathbb{R}^n$ be a compact set. For $S \subset [n]$ denote by K_S – projection of K on the subset S of coordinate axes. Then there exists a rectangle A s.t. $\mathrm{Leb}\{A\} = \mathrm{Leb}\{K\}$ and for all $S \subset [n]$:*

$$\mathrm{Leb}\{A_S\} \leq \mathrm{Leb}\{K_S\}$$

[2] For an example, consider piecewise-constant pdf taking value $e^{(-1)^n n}$ on the n-th interval of width $\Delta_n = \frac{c}{n^2}e^{-(-1)^n n}$.

[3] Note that since K is compact, its projection and slices are all compact and hence measurable.

Proof. Let X^n be uniformly distributed on K. Then $h(X^n) = \log \mathrm{Leb}\{K\}$. Let A be rectangle $a_1 \times \cdots \times a_n$ where

$$\log a_i = h(X_i | X^{i-1}) \,.$$

Then, we have by 1. in Theorem 1.5

$$h(X_S) \le \log \mathrm{Leb}\{K_S\}$$

On the other hand, by the chain rule

$$\begin{aligned} h(X_S) &= \sum_{i=1}^{n} 1\{i \in S\} h(X_i | X_{[i-1]\cap S}) && (1.20)\\ &\ge \sum_{i \in S} h(X_i | X^{i-1}) && (1.21)\\ &= \log \prod_{i \in S} a_i && (1.22)\\ &= \log \mathrm{Leb}\{A_S\} && (1.23) \end{aligned}$$

□

Corollary 1.1 (Loomis-Whitney)**.** *Let K be a compact subset of $\mathbb{R}^n$ and let K_{j^c} denote projection of K on coordinate axes $[n] \setminus j$. Then*

$$\mathrm{Leb}\{K\} \le \prod_{j=1}^{n} \mathrm{Leb}\{K_{j^c}\}^{\frac{1}{n-1}} \,. \tag{1.24}$$

Proof. Apply previous theorem to construct rectangle A and note that

$$\mathrm{Leb}\{K\} = \mathrm{Leb}\{A\} = \prod_{j=1}^{n} \mathrm{Leb}\{A_{j^c}\}^{\frac{1}{n-1}}$$

By previous theorem $\mathrm{Leb}\{A_{j^c}\} \le \mathrm{Leb}\{K_{j^c}\}$. □

The meaning of Loomis-Whitney inequality is best understood by introducing the average width of K in direction j: $w_j \triangleq \frac{\mathrm{Leb}\{K\}}{\mathrm{Leb}\{K_{j^c}\}}$. Then (1.24) is equivalent to

$$\mathrm{Leb}\{K\} \ge \prod_{j=1}^{n} w_j \,,$$

i.e. that volume of K is greater than volume of the rectangle of average widths.

§ 2. Information measures: mutual information

2.1 Divergence: main inequality

Theorem 2.1 (Information Inequality)**.**

$$D(P\|Q) \ge 0\ ; \quad D(P\|Q) = 0 \quad \textit{iff}\ P = Q$$

Proof. Let $\varphi(x) \triangleq x \log x$, which is strictly convex, and use Jensen's Inequality:

$$D(P\|Q) = \sum_{\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} = \sum_{\mathcal{X}} Q(x) \varphi\left(\frac{P(x)}{Q(x)}\right) \ge \varphi\left(\sum_{\mathcal{X}} Q(x) \frac{P(x)}{Q(x)}\right) = \varphi(1) = 0$$

□

2.2 Conditional divergence

The main objects in our course are random variables. The main operation for creating new random variables, and also for defining relations between random variables, is that of a random transformation:

Definition 2.1. Conditional probability distribution (aka random transformation, transition probability kernel, Markov kernel, channel) $K(\cdot|\cdot)$ has two arguments: first argument is a measurable subset of $\mathcal{Y}$, second argument is an element of $\mathcal{X}$. It must satisfy:

1. For any $x \in \mathcal{X}$: $K(\,\cdot\,|x)$ is a probability measure on $\mathcal{Y}$
2. For any measurable A function $x \mapsto K(A|x)$ is measurable on $\mathcal{X}$.

In this case we will say that K acts from $\mathcal{X}$ to $\mathcal{Y}$. In fact, we will abuse notation and write $P_{Y|X}$ instead of K to suggest what spaces $\mathcal{X}$ and $\mathcal{Y}$ are[1]. Furthermore, if X and Y are connected by the random transformation $P_{Y|X}$ we will write $X \xrightarrow{P_{Y|X}} Y$.

Remark 2.1. (Very technical!) Unfortunately, condition 2 (standard for probability textbooks) will frequently not be sufficiently strong for this course. The main reason is that we want Radon-Nikodym derivatives such as $\frac{dP_{Y|X=x}}{dQ_Y}(y)$ to be jointly measurable in (x, y). See Section **??** for more.

Example:

1. deterministic system: $Y = f(X) \Leftrightarrow P_{Y|X=x} = \delta_{f(x)}$
2. decoupled system: $Y \perp\!\!\!\perp X \Leftrightarrow P_{Y|X=x} = P_Y$

[1] Another reason for writing $P_{Y|X}$ is that from any joint distribution $P_{X,Y}$ (on standard Borel spaces) one can extract a random transformation by conditioning on X.

3. additive noise (convolution): $Y = X + Z$ with $Z \perp\!\!\!\perp X \Leftrightarrow P_{Y|X=x} = P_{x+Z}$.

Multiplication:

$$\begin{tikzcd}
P_X \arrow[dr] & & \\
& X \arrow[r, "P_{Y|X}"] & Y
\end{tikzcd}$$

to get $P_{XY} = P_X P_{Y|X}$:

$$P_{XY}(x,y) = P_{Y|X}(y|x) P_X(x)\,.$$

Composition (Marginalization): $P_Y = P_{Y|X} \circ P_X$, that is $P_{Y|X}$ acts on P_X to produce P_Y:

$$P_Y(y) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x) P_X(x)\,.$$

Will also write $P_X \xrightarrow{P_{Y|X}} P_Y$.

Definition 2.2 (Conditional divergence)**.**

$$\begin{aligned} D(P_{Y|X}\|Q_{Y|X}|P_X) &= \mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}\|Q_{Y|X=x})] && (2.1)\\ &= \sum_{x\in\mathcal{X}} P_X(x) D(P_{Y|X=x}\|Q_{Y|X=x})\,. && (2.2)\end{aligned}$$

Note: $H(X|Y) = \log|\mathcal{A}| - D(P_{X|Y}\|U_X|P_Y)$, where U_X is is uniform distribution on $\mathcal{X}$.

Theorem 2.2 (Properties of Divergence)**.**

1. $D(P_{Y|X}\|Q_{Y|X}|P_X) = D(P_X P_{Y|X}\|P_X Q_{Y|X})$

2. *(Simple chain rule)* $D(P_{XY}\|Q_{XY}) = D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_X\|Q_X)$

3. *(Monotonicity)* $D(P_{XY}\|Q_{XY}) \geq D(P_Y\|Q_Y)$

4. *(Full chain rule)*

$$D(P_{X_1\cdots X_n}\|Q_{X_1\cdots X_n}) = \sum_{i=1}^n D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}})$$

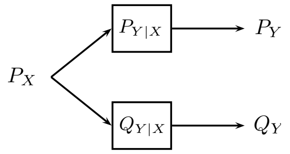
In the special case of $Q_{X^n} = \prod_i Q_{X_i}$ *we have*

$$D(P_{X_1\cdots X_n}\|Q_{X_1}\cdots Q_{X_n}) = D(P_{X_1\cdots X_n}\|P_{X_1}\cdots P_{X_n}) + \sum D(P_{X_i}\|Q_{X_i})$$

5. *(**Conditioning increases divergence**) Let* $P_{Y|X}$ *and* $Q_{Y|X}$ *be two kernels, let* $P_Y = P_{Y|X}\circ P_X$ *and* $Q_Y = Q_{Y|X}\circ P_X$. *Then*

$$\begin{aligned} D(P_Y\|Q_Y) &\leq D(P_{Y|X}\|Q_{Y|X}|P_X)\\ &\quad \textit{equality iff } D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0\end{aligned}$$

Pictorially:



6. *(**Data-processing for divergences**) Let $P_Y = P_{Y|X} \circ P_X$*

$$\left.\begin{array}{rcl} P_Y & = & \int P_{Y|X}(\cdot|x)dP_X \\ Q_Y & = & \int P_{Y|X}(\cdot|x)dQ_X \end{array}\right\} \Longrightarrow D(P_Y\|Q_Y) \le D(P_X\|Q_X) \tag{2.3}$$

Pictorially:

$$\begin{tikzcd} P_X \arrow[dr] & & & P_Y \\ & X \arrow[r, "P_{Y|X}"] & Y \arrow[ur] \arrow[dr] & \\ Q_X \arrow[ur] & & & Q_Y \end{tikzcd}$$

$$\Longrightarrow D(P_X\|Q_X) \ge D(P_Y\|Q_Y)$$

Proof. We only illustrate these results for the case of finite alphabets. General case follows by doing a careful analysis of Radon-Nikodym derivatives, introduction of regular branches of conditional probability etc. For certain cases (e.g. separable metric spaces), however, we can simply discretize alphabets and take granularity of discretization to 0. This method will become clearer in Lecture 4, once we understand continuity of D.

1. $\mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}\|Q_{Y|X=x})] = \mathbb{E}_{(X,Y)\sim P_X P_{Y|X}}\left[\log \frac{P_{Y|X}}{Q_{Y|X}}\frac{P_X}{P_X}\right]$

2. Disintegration: $\mathbb{E}_{(X,Y)}\left[\log \frac{P_{XY}}{Q_{XY}}\right] = \mathbb{E}_{(X,Y)}\left[\log \frac{P_{Y|X}}{Q_{Y|X}} + \log\frac{P_X}{Q_X}\right]$

3. Apply 2. with X and Y interchanged and use $D(\cdot\|\cdot) \ge 0$.

4. Telescoping $P_{X^n} = \prod_{i=1}^n P_{X_i|X^{i-1}}$ and $Q_{X^n} = \prod_{i=1}^n Q_{X_i|X^{i-1}}$.

5. Inequality follows from monotonicity. To get conditions for equality, notice that by the chain rule for D:

$$\begin{aligned} D(P_{XY}\|Q_{XY}) &= D(P_{Y|X}\|Q_{Y|X}|P_X) + \underbrace{D(P_X\|P_X)}_{=0} \\ &= D(P_{X|Y}\|Q_{X|Y}|P_Y) + D(P_Y\|Q_Y) \end{aligned}$$

   and hence we get the claimed result from positivity of D.

6. This again follows from monotonicity. □

Corollary 2.1.

$$\begin{array}{rcl} D(P_{X_1\cdots X_n}\|Q_{X_1}\cdots Q_{X_n}) & \ge & \sum D(P_{X_i}\|Q_{X_i}) \quad \textit{or} \\ & = & \textit{iff } P_{X^n} = \prod_{j=1}^n P_{X_j} \end{array}$$

Note: In general we can have $D(P_{XY}\|Q_{XY}) \lessgtr D(P_X\|Q_X) + D(P_Y\|Q_Y)$. For example, if $X = Y$ under P and Q, then $D(P_{XY}\|D(Q_{XY}) = D(P_X\|Q_X) < 2D(P_X\|Q_X)$. Conversely, if $P_X = Q_X$ and $P_Y = Q_Y$ but $P_{XY} \ne Q_{XY}$ we have $D(P_{XY}\|Q_{XY}) > 0 = D(P_X\|Q_X) + D(P_Y\|Q_Y)$.

Corollary 2.2. *$Y = f(X) \Rightarrow D(P_Y\|Q_Y) \le D(P_X\|Q_X)$, with equality if f is 1-1.*

Note: $D(P_Y\|Q_Y) = D(P_X\|Q_X) \not\Rightarrow f$ is 1-1. Example: P_X = Gaussian, Q_X = Laplace, $Y = |X|$.

Corollary 2.3 (Large deviations estimate)**.** *For any subset $E \subset \mathcal{X}$ we have*

$$d(P_X[E]\|Q_X[E]) \le D(P_X\|Q_X)$$

Proof. Consider $Y = \mathbf{1}_{\{X\in E\}}$. □

2.3 Mutual information

Definition 2.3 (Mutual information)**.**

$$I(X;Y) = D(P_{XY}\|P_X P_Y)$$

Note:

- Intuition: $I(X;Y)$ measures the dependence between X and Y, or, the information about X (resp. Y) provided by Y (resp. X)
- Defined by Shannon (in a different form), in this form by Fano.
- Note: not restricted to discrete.
- $I(X;Y)$ is a functional of the joint distribution P_{XY}, or equivalently, the pair $(P_X, P_{Y|X})$.

Theorem 2.3 (Properties of I)**.**

1. $I(X;Y) = D(P_{XY}\|P_X P_Y) = D(P_{Y|X}\|P_Y|P_X) = D(P_{X|Y}\|P_X|P_Y)$
2. *Symmetry:* $I(X;Y) = I(Y;X)$
3. *Positivity:* $I(X;Y) \ge 0$*;* $I(X;Y) = 0$ *iff* $X \perp\!\!\!\perp Y$
4. $I(f(X);Y) \le I(X;Y)$*;* f *one-to-one* $\Rightarrow I(f(X);Y) = I(X;Y)$
5. *"More data $\Rightarrow$ More info":* $I(X_1, X_2; Z) \ge I(X_1; Z)$

Proof.

1. $I(X;Y) = \mathbb{E}\log\frac{P_{XY}}{P_X P_Y} = \mathbb{E}\log\frac{P_{Y|X}}{P_Y} = \mathbb{E}\log\frac{P_{X|Y}}{P_X}$.
2. Apply data-processing inequality twice to the map $(x,y) \to (y,x)$ to get $D(P_{X,Y}\|P_X P_Y) = D(P_{Y,X}\|P_Y P_X)$.
3. By definition.
4. We will use the data-processing property of mutual information (to be proved shortly, see Theorem 2.5). Consider the chain of data processing: $(x,y) \mapsto (f(x),y) \mapsto (f^{-1}(f(x)),y)$. Then
   $I(X;Y) \ge I(f(X);Y) \ge I(f^{-1}(f(X));Y) = I(X;Y)$
5. Consider $f(X_1, X_2) = X_1$. □

Theorem 2.4 (I v.s. H)**.**

1. *$I(X;X) = \begin{cases} H(X) & X \text{ discrete} \\ +\infty & \text{otherwise} \end{cases}$*

2. *If X, Y discrete then*
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$
If only X discrete then
$$I(X;Y) = H(X) - H(X|Y)$$

3. *If X, Y are real-valued vectors, have joint pdf and all three differential entropies are finite then*
$$I(X;Y) = h(X) + h(Y) - h(X,Y)$$
If X has marginal pdf p_X and conditional pdf $p_{X|Y}(x|y)$ then
$$I(X;Y) = h(X) - h(X|Y)\,.$$

4. *If X or Y are discrete then $I(X;Y) \le \min(H(X), H(Y))$, with equality iff $H(X|Y) = 0$ or $H(Y|X) = 0$, i.e., one is a deterministic function of the other.*

Proof.

1. By definition, $I(X;X) = D(P_{X|X} \| P_X | P_X) = \mathbb{E}_{x\sim X} D(\delta_x \| P_X)$. If P_X is discrete, then $D(\delta_x \| P_X) = \log \frac{1}{P_X(x)}$ and $I(X;X) = H(X)$. If P_X is not discrete, then let $\mathcal{A} = \{x : P_X(x) > 0\}$ denote the set of atoms of P_X. Let $\Delta = \{(x,x) : x \notin \mathcal{A}\} \subset \mathcal{X} \times \mathcal{X}$. Then $P_{X,X}(\Delta) = P_X(\mathcal{A}^c) > 0$ but since
$$(P_X \times P_X)(E) \triangleq \int_{\mathcal{X}} P_X(dx_1) \int_{\mathcal{X}} P_X(dx_2) 1\{(x_1,x_2) \in E\}$$
we have by taking $E = \Delta$ that $(P_X \times P_X)(\Delta) = 0$. Thus $P_{X,X} \not\ll P_X \times P_X$ and thus
$$I(X;X) = D(P_{X,X} \| P_X P_X) = +\infty\,.$$

2. $\mathbb{E} \log \frac{P_{XY}}{P_X P_Y} = \mathbb{E}\left[\log \frac{1}{P_X} + \log \frac{1}{P_Y} - \log \frac{1}{P_{XY}}\right]$. □

3. Similarly, when $P_{X,Y}$ and $P_X P_Y$ have densities p_{XY} and $p_X p_Y$ we have
$$D(P_{XY} \| P_X P_Y) \triangleq \mathbb{E}\left[\log \frac{p_{XY}}{p_X p_Y}\right] = h(X) + h(Y) - h(X,Y)$$

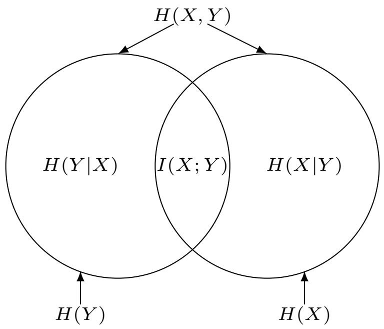
4. Follows from 2.

Corollary 2.4 (Conditioning reduces entropy)**.** *X discrete: $H(X|Y) \le H(X)$, with equality iff $X \perp\!\!\!\perp Y$.*
Intuition: *The amount of entropy reduction = mutual information*

Example: $X = U \,\texttt{OR}\, Y$, where $U, Y \overset{\text{i.i.d.}}{\sim} \text{Bern}(\frac{1}{2})$. Then $X \sim \text{Bern}(\frac{3}{4})$ and $H(X) = h(\frac{1}{4}) < 1\,\texttt{bits} = H(X|Y=0)$, i.e., conditioning on $Y = 0$ increases entropy. But *on average*, $H(X|Y) = \mathbb{P}[Y=0]\,H(X|Y=0) + \mathbb{P}[Y=1]\,H(X|Y=1) = \frac{1}{2}\,\texttt{bits} < H(X)$, by the strong concavity of $h(\cdot)$.

Note: Information, entropy and Venn diagrams:

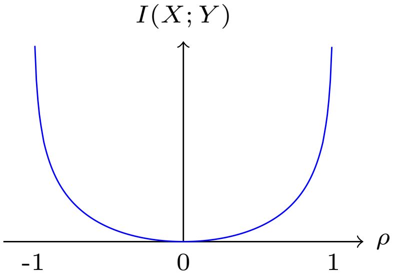
1. The following Venn diagram illustrates the relationship between entropy, conditional entropy, joint entropy, and mutual information.



2. If you do the same for 3 variables, you will discover that the triple intersection corresponds to
$$H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_2, X_3) - H(X_1, X_3) + H(X_1, X_2, X_3) \tag{2.4}$$
which is sometimes denoted $I(X;Y;Z)$. It can be both positive and negative (why?).

3. In general, one can treat random variables as sets (so that r.v. X_i corresponds to set E_i and (X_1, X_2) corresponds to $E_1 \cup E_2$). Then we can define a unique signed measure μ on the finite algebra generated by these sets so that every information quantity is found by replacing
$$I/H \to \mu \quad ; \to \cap \quad , \to \cup \quad | \to \setminus .$$
As an example, we have
$$H(X_1|X_2, X_3) = \mu(E_1 \setminus (E_2 \cup E_3)), \tag{2.5}$$
$$I(X_1, X_2; X_3|X_4) = \mu(((E_1 \cup E_2) \cap E_3) \setminus E_4). \tag{2.6}$$
By inclusion-exclusion, quantity (2.4) corresponds to $\mu(E_1 \cap E_2 \cap E_3)$, which explains why μ is not necessarily a positive measure.

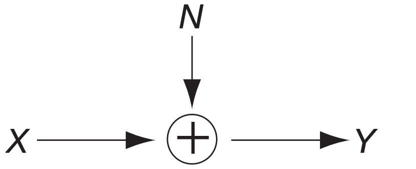
Example: *Bivariate Gaussian.* X, Y — jointly Gaussian
$$I(X;Y) = \frac{1}{2} \log \frac{1}{1 - \rho_{XY}^2}$$
where $\rho_{XY} \triangleq \frac{\mathbb{E}[(X-\mathbb{E}X)(Y-\mathbb{E}Y)]}{\sigma_X \sigma_Y} \in [-1, 1]$ is the correlation coefficient.



Proof. WLOG, by shifting and scaling if necessary, we can assume $\mathbb{E}X = \mathbb{E}Y = 0$ and $\mathbb{E}X^2 = \mathbb{E}Y^2 = 1$. Then $\rho = \mathbb{E}XY$. By joint Gaussianity, $Y = \rho X + Z$ for some $Z \sim \mathcal{N}(0, 1 - \rho^2) \perp\!\!\!\perp X$. Then using the divergence formula for Gaussians (1.16), we get
$$\begin{aligned} I(X;Y) &= D(P_{Y|X} \| P_Y | P_X) \\ &= \mathbb{E} D(\mathcal{N}(\rho X, 1 - \rho^2) \| \mathcal{N}(0, 1)) \\ &= \mathbb{E}\left[\frac{1}{2} \log \frac{1}{1 - \rho^2} + \frac{\log \mathrm{e}}{2} \left((\rho X)^2 + 1 - \rho^2 - 1\right)\right] \\ &= \frac{1}{2} \log \frac{1}{1 - \rho^2} \end{aligned}$$
□

Note: Similar to the role of mutual information, the correlation coefficient also measures the dependency between random variables which are real-valued (more generally, on an inner-product space) in certain sense. However, mutual information is invariant to bijections and more general: it can be defined not just for numerical random variables, but also for apples and oranges.

Example: *Additive white Gaussian noise (AWGN) channel.* $X \perp\!\!\!\perp N$ — independent Gaussian



$$I(X; X+N) = \tfrac{1}{2}\log\Big(1 + \underbrace{\tfrac{\sigma_X^2}{\sigma_N^2}}_{\text{signal-to-noise ratio (SNR)}}\Big)$$

Example: *Gaussian vectors.* $\mathbf{X} \in \mathbb{R}^m, \mathbf{Y} \in \mathbb{R}^n$ — jointly Gaussian

$$I(\mathbf{X}; \mathbf{Y}) = \frac{1}{2}\log\frac{\det\Sigma_{\mathbf{X}}\det\Sigma_{\mathbf{Y}}}{\det\Sigma_{[\mathbf{X},\mathbf{Y}]}}$$

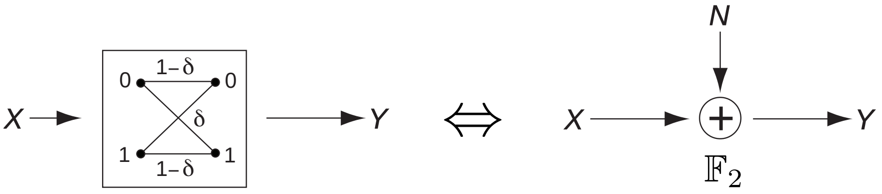
where $\Sigma_{\mathbf{X}} \triangleq \mathbb{E}\left[(\mathbf{X} - \mathbb{E}\mathbf{X})(\mathbf{X} - \mathbb{E}\mathbf{X})'\right]$ denotes the covariance matrix of $\mathbf{X} \in \mathbb{R}^m$, and $\Sigma_{[\mathbf{X},\mathbf{Y}]}$ denotes the the covariance matrix of the random vector $[\mathbf{X}, \mathbf{Y}] \in \mathbb{R}^{m+n}$.

In the special case of additive noise: $\mathbf{Y} = \mathbf{X} + \mathbf{N}$ for $\mathbf{N} \perp\!\!\!\perp \mathbf{X}$, we have

$$I(\mathbf{X}; \mathbf{X} + \mathbf{N}) = \frac{1}{2}\log\frac{\det(\Sigma_{\mathbf{X}} + \Sigma_{\mathbf{N}})}{\det\Sigma_{\mathbf{N}}}$$

since $\det\Sigma_{[\mathbf{X},\mathbf{X}+\mathbf{N}]} = \det\left(\begin{smallmatrix}\Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}} \\ \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}}+\Sigma_{\mathbf{N}}\end{smallmatrix}\right) \overset{\text{why?}}{=} \det\Sigma_{\mathbf{X}}\det\Sigma_{\mathbf{N}}$.

Example: *Binary symmetric channel (BSC).*



$$\begin{aligned} X\ &\sim \mathrm{Bern}\Big(\frac{1}{2}\Big),\ N \sim \mathrm{Bern}(\delta) \\ Y &= X + N \\ I(X;Y) &= \log 2 - h(\delta) \end{aligned}$$

Example: *Addition over finite groups.* X is uniform on G and independent of Z. Then

$$I(X; X+Z) = \log|G| - H(Z)$$

Proof. Show that $X + Z$ is uniform on G regardless of Z. □

2.4 Conditional mutual information and conditional independence

Definition 2.4 (Conditional mutual information)**.**

$$
\begin{aligned}
I(X;Y|Z) &= D(P_{XY|Z} \| P_{X|Z} P_{Y|Z} | P_Z) && (2.7)\\
&= \mathbb{E}_{z\sim P_Z}[I(X;Y|Z=z)]\,. && (2.8)
\end{aligned}
$$

where the product of two random transformations is $(P_{X|Z=z}P_{Y|Z=z})(x,y) \triangleq P_{X|Z}(x|z)P_{Y|Z}(y|z)$, under which X and Y are independent conditioned on Z.

Note: $I(X;Y|Z)$ is a functional of P_{XYZ}.

Remark 2.2 (Conditional independence)**.** A family of distributions can be represented by a directed acyclic graph. A simple example is a Markov chain (line graph), which represents distributions that factor as $\{P_{XYZ} : P_{XYZ} = P_X P_{Y|X} P_{Z|Y}\}$.

$$
\text{Cond. indep. notation}
\left\{
\begin{array}{lll}
X \to Y \to Z & \Leftrightarrow & P_{XZ|Y} = P_{X|Y} \cdot P_{Z|Y} \\
& \Leftrightarrow & P_{Z|XY} = P_{Z|Y} \\
& \Leftrightarrow & P_{XYZ} = P_X \cdot P_{Y|X} \cdot P_{Z|Y} \\
& \Leftrightarrow & X, Y, Z \text{ form a Markov chain} \\
& \Leftrightarrow & X \perp\!\!\!\perp Z|Y \\
& \Leftrightarrow & P_{XYZ} = P_Y \cdot P_{X|Y} \cdot P_{Z|Y} \\
& \Leftrightarrow & Z \to Y \to X
\end{array}
\right.
$$

Theorem 2.5 (Further properties of Mutual Information)**.**

1. *$I(X;Z|Y) \ge 0$, with equality iff $X \to Y \to Z$*

2. *(Kolmogorov identity or small chain rule)*

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

3. *(**Data Processing**) If $X \to Y \to Z$, then*

   a) $I(X;Z) \le I(X;Y)$

   b) $I(X;Y|Z) \le I(X;Y)$

4. *(Full chain rule)*

$$
I(X^n;Y) = \sum_{k=1}^{n} I(X_k;Y|X^{k-1})
$$

Proof. 1. By definition and Theorem 2.3.3.

2.

$$
\frac{P_{XYZ}}{P_{XY}P_Z} = \frac{P_{XZ}}{P_X P_Z} \cdot \frac{P_{Y|XZ}}{P_{Y|X}}
$$

3. Apply Kolmogorov identity to $I(Y,Z;X)$:

$$\begin{aligned} I(Y,Z;X) &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \\ &= I(X;Z) + I(X;Y|Z) \end{aligned}$$

4. Recursive application of Kolmogorov identity. □

Example: 1-to-1 function ⇒ $I(X;Y) = I(X;f(Y))$
Note: In general, $I(X;Y|Z) \gtrless I(X;Y)$. Examples:

a) ">": Conditioning does not always decrease M.I. To find counterexamples when X, Y, Z do not form a Markov chain, notice that there is only one directed acyclic graph non-isomorphic to $X \to Y \to Z$, namely $X \to Y \leftarrow Z$. Then a counterexample is

$$\begin{aligned} X, Z &\overset{\text{i.i.d.}}{\sim} \text{Bern}(\frac{1}{2}) && Y = X \oplus Z \\ I(X;Y) &= 0 && \text{since } X \perp\!\!\!\perp Y \\ I(X;Y|Z) = I(X;X \oplus Z|Z) = H(X) &= \log 2 \end{aligned}$$

b) "<": $Z = Y$. Then $I(X;Y|Y) = 0$.

Note: (Chain rule for I ⇒ Chain rule for H) Set $Y = X^n$. Then $H(X^n) = I(X^n;X^n) = \sum_{k=1}^n I(X_k;X^n|X^{k-1}) = \sum_{k=1}^n H(X_k|X^{k-1})$, since $H(X_k|X^n, X^{k-1}) = 0$.

Remark 2.3 (Data processing for mutual information via data processing of divergence)**.** We proved data processing for mutual information in Theorem 2.5 using Kolmogorov's identity. In fact, data processing for mutual information is *implied by* the data processing for divergence:

$$I(X;Z) = D(P_{Z|X}\|P_Z|P_X) \le D(P_{Y|X}\|P_Y|P_X) = I(X;Y),$$

where note that for each x, we have $P_{Y|X=x} \xrightarrow{P_{Z|Y}} P_{Z|X=x}$ and $P_Y \xrightarrow{P_{Z|Y}} P_Z$. Therefore if we have a bi-variate functional of distributions $\mathcal{D}(P\|Q)$ which satisfies data processing, then we can define an "M.I.-like" quantity via $I_{\mathcal{D}}(X;Y) \triangleq \mathcal{D}(P_{Y|X}\|P_Y|P_X) \triangleq \mathbb{E}_{x\sim P_X}\mathcal{D}(P_{Y|X=x}\|P_Y)$ which will satisfy data processing on Markov chains. A rich class of examples arises by taking $\mathcal{D} = D_f$ (an f-divergence, defined in (1.15)). That f-divergence satisfies data-processing is going to be shown in Remark 4.2.

2.5 Strong data-processing inequalities

For many random transformations $P_{Y|X}$, it is possible to improve the data-processing inequality (2.3): For any P_X, Q_X we have

$$D(P_Y\|Q_Y) \le \eta_{KL} D(P_X\|Q_X)\,,$$

where $\eta_{KL} < 1$ and depends on the channel $P_{Y|X}$ only. Similarly, this gives an improvement in the data-processing inequality for mutual information: For any $P_{U,X}$ we have

$$U \to X \to Y \quad \Longrightarrow \quad I(U;Y) \le \eta_{KL} I(U;X)\,.$$

For example, for $P_{Y|X} = BSC(\delta)$ we have $\eta_{KL} = (1-2\delta)^2$. Strong data-processing inequalities quantify the intuitive observation that noise inside the channel $P_{Y|X}$ must reduce the information that Y carries about the data U, regardless of how smart the hook up $U \to X$ is.

This is an active area of research, see [PW15] for a short summary.

2.6* How to avoid measurability problems?

As we mentioned in Remark 2.1 conditions imposed by Definition 2.1 on $P_{Y|X}$ are insufficient. Namely, we get the following two issues:

1. Radon-Nikodym derivatives such as $\frac{dP_{Y|X=x}}{dQ_Y}(y)$ may not be jointly measurable in (x,y)
2. Set $\{x : P_{Y|X=x} \ll Q_Y\}$ may not be measurable.

The easiest way to avoid all such problems is the following:

Agreement A1: All conditional kernels $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ in these notes will be assumed to be defined by choosing a σ-finite measure μ_2 on $\mathcal{Y}$ and measurable function $\rho(y|x) \geq 0$ on $\mathcal{X} \times \mathcal{Y}$ such that

$$P_{Y|X}(A|x) = \int_A \rho(y|x)\mu_2(dy)$$

for all x and measurable sets A and $\int_{\mathcal{Y}} \rho(y|x)\mu_2(dy) = 1$ for all x.

Notes:

1. Given another kernel $Q_{Y|X}$ specified via $\rho'(y|x)$ and μ_2' we may first replace μ_2 and μ_2' via $\mu_2'' = \mu_2 + \mu_2'$ and thus assume that both $P_{Y|X}$ and $Q_{Y|X}$ are specified in terms of the same dominating measure μ_2''. (This modifies $\rho(y|x)$ to $\rho(y|x)\frac{d\mu_2}{d\mu_2''}(y)$.)
2. Given two kernels $P_{Y|X}$ and $Q_{Y|X}$ specified in terms of the same dominating measure μ_2 and functions $\rho_P(y|x)$ and $\rho_Q(y|x)$, respectively, we may set

$$\frac{dP_{Y|X}}{dQ_{Y|X}} \triangleq \frac{\rho_P(y|x)}{\rho_Q(y|x)}$$

   outside of $\rho_Q = 0$. When $P_{Y|X=x} \ll Q_{Y|X=x}$ the above gives a version of the Radon-Nikodym derivative, which is automatically measurable in (x,y).
3. Given Q_Y specified as

$$dQ_Y = q(y)d\mu_2$$

   we may set

$$A_0 = \{x : \int_{\{q=0\}} \rho(y|x)d\mu_2 = 0\}$$

   This set plays a role of $\{x : P_{Y|X=x} \ll Q_Y\}$. Unlike the latter A_0 is guaranteed to be measurable by Fubini [Ç11, Prop. 6.9]. By "plays a role" we mean that it allows to prove statements like: For any P_X

$$P_{X,Y} \ll P_X Q_Y \iff P_X[A_0] = 1\,.$$

So, while our agreement resolves the two measurability problems above, it introduces a new one. Indeed, given a joint distribution $P_{X,Y}$ on standard Borel spaces, it is always true that one can extract a conditional distribution $P_{Y|X}$ satisfying Definition 2.1 (this is called disintegration). However, it is not guaranteed that $P_{Y|X}$ will satisfy Agreement A1. To work around this issue as well, we add another agreement:

Agreement A2: All joint distributions $P_{X,Y}$ are specified by means of data: μ_1,μ_2 – σ-finite measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and measurable function $\lambda(x,y)$ such that

$$P_{X,Y}(E) \triangleq \int_E \lambda(x,y)\mu_1(dx)\mu_2(dy)\,.$$

Notes:

1. Again, given a finite or countable collection of joint distributions $P_{X,Y}, Q_{X,Y}, \ldots$ satisfying A2 we may without loss of generality assume they are defined in terms of a common μ_1, μ_2.

2. Given $P_{X,Y}$ satisfying A2 we can disintegrate it into conditional (satisfying A1) and marginal:

$$P_{Y|X}(A|x) = \int_A \rho(y|x)\mu_2(dy) \qquad \rho(y|x) \triangleq \frac{\lambda(x,y)}{p(x)} \tag{2.9}$$

$$P_X(A) = \int_A p(x)\mu_1(dx) \qquad p(x) \triangleq \int_{\mathcal{Y}} \lambda(x,\eta)\mu_2(d\eta) \tag{2.10}$$

   with $\rho(y|x)$ defined arbitrarily for those x, for which $p(x) = 0$.

Remark 2.4. The first problem can also be resolved with the help of Doob's version of Radon-Nikodym theorem [Ç11, Chapter V.4, Theorem 4.44]: If the σ-algebra on $\mathcal{Y}$ is separable (satisfied whenever $\mathcal{Y}$ is a Polish space, for example) and $P_{Y|X=x} \ll Q_{Y|X=x}$ then there exists a jointly measurable version of Radon-Nikodym derivative

$$(x,y) \mapsto \frac{dP_{Y|X=x}}{dQ_{Y|X=x}}(y)$$

§ 3. Sufficient statistic. Continuity of divergence and mutual information

3.1 Sufficient statistics and data-processing

Definition 3.1 (Sufficient Statistic)**.** Let

- P_X^θ be a collection of distributions of X parameterized by θ
- $P_{T|X}$ be some probability kernel. Let $P_T^\theta \triangleq P_{T|X} \circ P_X^\theta$ be the induced distribution on T for each θ.

We say that T is a *sufficient statistic* (s.s.) $\underline{\text{of } X}$ $\underline{\text{for } \theta}$ if there exists a transition probability kernel $P_{X|T}$ so that $P_X^\theta P_{T|X} = P_T^\theta P_{X|T}$. (I.e.: $P_{X|T}$ can be chosen to not depend on θ).

Note:

- Know T, can forget X (T contains all the information that is sufficient to make inference about θ)
- Obviously any one-to-one transformation of X is sufficient. Therefore the interesting case is when T is a low-dimensional recap of X (dimensionality reduction)
- θ need not be a random variable (the definition does not involve any distribution on θ)

Theorem 3.1. *Let $\theta \to X \to T$. Then the following are equivalent*

1. *T is a s.s. of X for θ.*
2. *$\forall P_\theta$, $\theta \to T \to X$.*
3. *$\forall P_\theta$, $I(\theta; X|T) = 0$.*
4. *$\forall P_\theta$, $I(\theta; X) = I(\theta; T)$, i.e., data processing inequality for M.I. holds with equality.*

Theorem 3.2 (Fisher's factorization criterion)**.** *For all $\theta \in \Theta$, let P_X^θ have a density p_θ with respect to a measure μ (e.g., discrete – pmf, continuous – pdf). Let $T = T(X)$ be a deterministic function of X. Then T is a s.s. of X for θ iff*

$$p_\theta(x) = g_\theta(T(x))h(x)$$

for some measurable functions g_θ and h, $\forall \theta \in \Theta$.

Proof. We only give the proof in the discrete case (continuous case $\sum \to \int d\mu$). Let $t = T(x)$.

"⇒": Suppose T is a s.s. of X for θ. Then $p_\theta(x) = P_\theta(X = x) = P_\theta(X = x, T = t) = P_\theta(X = x|T = t)P_\theta(T = t) = \underbrace{P(X = x|T = T(x))}_{h(x)}\underbrace{P_\theta(T = T(x))}_{g_\theta(T(x))}$

"⇐": Suppose the factorization holds. Then

$$P_\theta(X = x|T = t) = \frac{p_\theta(x)}{\sum_x \mathbf{1}_{\{T(x)=t\}} p_\theta(x)} = \frac{g_\theta(t)h(x)}{\sum_x \mathbf{1}_{\{T(x)=t\}} g_\theta(t)h(x)} = \frac{h(x)}{\sum_x \mathbf{1}_{\{T(x)=t\}} h(x)},$$

free of θ. □

Example:

1. *Normal mean model.* Let $\theta \in \mathbb{R}$ and observations $X_i \overset{\text{indep.}}{\sim} \mathcal{N}(\theta, 1), i \in [n]$. Then the sample mean $\bar{X} = \frac{1}{n}\sum_j X_j$ is a s.s. of X^n for θ.
   Verify: $P^\theta_{X^n}$ factorizes.

2. *Coin flips.* Let $B_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$. Then $\sum_{i=1}^n B_i$ is a s.s. of B^n for θ.

3. *Uniform distribution.* Let $U_i \overset{\text{i.i.d.}}{\sim} \text{uniform}[0, \theta]$. Then $\max_{i\in[n]} U_i$ is a s.s. of U^n for θ.

Example: *Binary hypothesis testing.* $\theta = \{0, 1\}$. Given $\theta = 0$ or 1, $X \sim P_X$ or Q_X. Then Y – the output of $P_{Y|X}$ – is a s.s. of X for θ iff $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$, i.e., $P_{X|Y} = Q_{X|Y}$ holds P_Y-a.s. Indeed, the latter means that for kernel $Q_{X|Y}$ we have

$$P_X P_{Y|X} = P_Y Q_{X|Y} \quad \text{and} \quad Q_X P_{Y|X} = Q_Y Q_{X|Y}\,,$$

which is precisely the definition of s.s. when $\theta \in \{0, 1\}$. This example explains condition for equality in the data-processing for divergence:

$$\begin{tikzcd}
P_X \arrow[dr] & & & & P_Y \\
& X \arrow[rr, "P_{Y|X}"] & & Y \arrow[ur] \arrow[dr, dashed] & \\
Q_X \arrow[ur, dashed] & & & & Q_Y
\end{tikzcd}$$

Then assuming $D(P_Y\|Q_Y) < \infty$ we have:

$$D(P_X\|Q_X) = D(P_Y\|Q_Y) \quad \Longleftrightarrow \quad Y - \text{s.s. for testing } P_X \text{ vs. } Q_X$$

Proof: Let $Q_{XY} = Q_X P_{Y|X}$, $P_{XY} = P_X P_{Y|X}$, then

$$\begin{aligned}
D(P_{XY}\|Q_{XY}) &= \underbrace{D(P_{Y|X}\|Q_{Y|X}|P_X)}_{=0} + D(P_X\|Q_X) \\
&= D(P_{X|Y}\|Q_{X|Y}|P_Y) + D(P_Y\|Q_Y) \\
&\ge D(P_Y\|Q_Y)
\end{aligned}$$

with equality iff $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$, which is equivalent to Y being a s.s. for testing P_X vs Q_X as desired.

3.2 Geometric interpretation of mutual information

Mutual information as "weighted distance":

$$I(X;Y) = D(P_{Y|X}\|P_Y|P_X) = \sum_x D(P_{Y|X=x}\|P_Y)P_X(x)$$

Theorem 3.3 (Golden formula)**.** *$\forall Q_Y$ such that $D(P_Y\|Q_Y) < \infty$*

$$I(X;Y) = D(P_{Y|X}\|Q_Y|P_X) - D(P_Y\|Q_Y)$$

Proof. For discrete case: $I(X;Y) = \mathbb{E}\log\frac{P_{Y|X}Q_Y}{P_Y Q_Y}$, group $P_{Y|X}$ and Q_Y. □

Corollary 3.1 (mutual information as center of gravity)**.**

$$I(X;Y) = \min_Q D(P_{Y|X}\|Q|P_X),$$

achieved at $Q = P_Y$.

Note: This representation is useful to bound mutual information from above.

Theorem 3.4 (mutual information as distance to product distributions)**.**

$$I(X;Y) = \min_{Q_X,Q_Y} D(P_{XY}\|Q_X Q_Y)$$

Proof. $I(X;Y) = \mathbb{E}\log\frac{P_{XY}Q_X Q_Y}{P_X P_Y Q_X Q_Y}$, group P_{XY} and $Q_X Q_Y$ and bound marginal divergences $D(P_X\|Q_X)$ and $D(P_Y\|Q_Y)$ by zero. □

Note: Generalization to conditional mutual information.

$$I(X;Z|Y) = \min_{Q_{XYZ}:X\to Y\to Z} D(P_{XYZ}\|Q_{XYZ})$$

Proof. By chain rule,

$$\begin{aligned}
&D(P_{XYZ}\|Q_X Q_{Y|X}Q_{Z|Y})\\
=&D(P_{XYZ}\|P_X P_{Y|X}P_{Z|Y}) + D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_{Z|Y}\|Q_{Z|Y}|P_Y)\\
=&D(P_{XYZ}\|P_Y P_{X|Y}P_{Z|Y}) + \dots\\
=&\underbrace{D(P_{XZ|Y}\|P_{X|Y}P_{Z|Y}|P_Y)}_{I(X;Z|Y)} + \dots
\end{aligned}$$

□

Interpretation: The most general graphical model for the triplet (X,Y,Z) is a 3-clique. What is the information flow on the edge $X \to Z$? To answer, notice that removing this edge restricts possible joint distributions to a Markov chain $X \to Y \to Z$. Thus, it is natural to ask what is the minimum distance between a given $P_{X,Y,Z}$ and the set of all distributions $Q_{X,Y,Z}$ satisfying the Markov chain constraint. By the above calculation, optimal $Q_{X,Y,Z} = P_Y P_{X|Y}P_{Z|Y}$ and hence the distance is $I(X;Z|Y)$. It is natural to take this number as the information flowing on the edge $X \to Z$.

3.3 Variational characterizations of divergence: Donsker-Varadhan

Why variational characterization (sup- or inf-representation): $F(x) = \sup_{\lambda\in\Lambda} f_\lambda(x)$

1. Regularity, e.g., recall

   a) Pointwise supremum of convex functions is convex

   b) Pointwise supremum of lower semicontinuous (lsc) functions is lsc

2. Give bounds by choosing a (suboptimal) λ

Theorem 3.5 (Donsker-Varadhan). *Let P, Q be probability measures on $\mathcal{X}$ and let $\mathcal{C}$ denote the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_Q[\exp\{f(X)\}] < \infty$. If $D(P\|Q) < \infty$ then for every $f \in \mathcal{C}$ expectation $\mathbb{E}_P[f(X)]$ exists and furthermore*

$$D(P\|Q) = \sup_{f\in\mathcal{C}} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp\{f(X)\}] \,. \tag{3.1}$$

Proof. "≤": take $f = \log \frac{dP}{dQ}$.

"≥": Fix $f \in \mathcal{C}$ and define a probability measure Q^f (*tilted version* of Q) via $Q^f(dx) \triangleq \frac{\exp\{f(x)\}Q(dx)}{\int_{\mathcal{X}} \exp\{f(x)\}Q(dx)}$, or equivalently,

$$Q^f(dx) = \exp\{f(x) - Z_f\}Q(dx)\,, \qquad Z_f \triangleq \log \mathbb{E}_Q[\exp\{f(X)\}]\,.$$

Then, obviously $Q^f \ll Q$ and we have

$$\mathbb{E}_P[f(X)] - Z_f = \mathbb{E}_P\left[\log \frac{dQ^f}{dQ}\right] = \mathbb{E}_P\left[\log \frac{dPdQ^f}{dQdP}\right] = D(P\|Q) - D(P\|Q^f) \le D(P\|Q)\,. \quad \square$$

Notes:

1. What is Donsker-Varadhan good for? By setting $f(x) = \epsilon \cdot g(x)$ with $\epsilon \ll 1$ and linearizing exp and log we can see that when $D(P\|Q)$ is small, expectations under P can be approximated by expectations over Q (change of measure): $\mathbb{E}_P[g(X)] \approx \mathbb{E}_Q[g(X)]$. This holds for all functions g with finite exponential moment under Q. Total variation distance provides a similar bound, but for a narrower class of bounded functions:

$$|\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)]| \le \|g\|_\infty \mathrm{TV}(P, Q)\,.$$

2. More formally, inequality $\mathbb{E}_P[f(X)] \le \log \mathbb{E}_Q[\exp f(X)] + D(P\|Q)$ is useful in estimating $\mathbb{E}_P[f(X)]$ for complicated distribution P (e.g. over large-dimensional vector X^n with lots of weak inter-coordinate dependencies) by making a smart choice of Q (e.g. with iid components).

3. In the next lecture we will show that $P \mapsto D(P\|Q)$ is convex. A general method of obtaining variational formulas like (3.1) is by Young-Fenchel inequality. Indeed, (3.1) is exactly this inequality since the Fenchel-Legendre conjugate of $D(\cdot\|Q)$ is given by a convex map $f \mapsto Z_f$.

Theorem 3.6 (Weak lower-semicontinuity of divergence). *Let $\mathcal{X}$ be a metric space with Borel σ-algebra $\mathcal{H}$. If P_n and Q_n converge weakly (in distribution) to P, Q, then*

$$D(P\|Q) \le \liminf_{n\to\infty} D(P_n\|Q_n)\,. \tag{3.2}$$

Proof. First method: On a metric space $\mathcal{X}$ bounded continuous functions ($\mathcal{C}_b$) are dense in the set of all integrable functions. Then in Donsker-Varadhan (3.1) we can replace $\mathcal{C}$ by $\mathcal{C}_b$ to get

$$D(P_n\|Q_n) = \sup_{f\in\mathcal{C}_b} \mathbb{E}_{P_n}[f(X)] - \log\mathbb{E}_{Q_n}[\exp\{f(X)\}]\,.$$

Recall $P_n \to P$ weakly if and only if $\mathbb{E}_{P_n} f(X) \to \mathbb{E}_P f(X)$ for all $f \in \mathcal{C}_b$. Taking the limit concludes the proof.

Second method (less mysterious): Let $\mathcal{A}$ be the algebra of Borel sets E whose boundary has zero $(P+Q)$ measure, i.e.

$$\mathcal{A} = \{E \in \mathcal{H} : (P+Q)(\partial E) = 0\}\,.$$

By the property of weak convergence P_n and Q_n converge pointwise on $\mathcal{A}$. Thus by (3.8) we have

$$D(P_{\mathcal{A}}\|Q_{\mathcal{A}}) \le \varliminf_{n\to\infty} D(P_{n,\mathcal{A}}\|Q_{n,\mathcal{A}})$$

If we show $\mathcal{A}$ is $(P+Q)$-dense in $\mathcal{H}$, we are done by (3.7). To get an idea, consider $\mathcal{X} = \mathbb{R}$. Then open sets are $(P+Q)$-dense in $\mathcal{H}$ (since finite measures are regular), while the algebra $\mathcal{F}$ generated by open intervals is $(P+Q)$-dense in the open sets. Since there are at most countably many points $a \in \mathcal{X}$ with $P(a) + Q(a) > 0$, we may further approximate each interval (a, b) whose boundary has non-zero $(P+Q)$ measure by a slightly larger interval from $\mathcal{A}$. □

Note: In general, $D(P\|Q)$ is *not* continuous in either P or Q. Example: Let $B_1, \ldots, B_n \overset{\text{i.i.d.}}{\sim} \{\pm 1\}$ equiprobably. Then $S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n B_i \xrightarrow{\text{D}} \mathcal{N}(0,1)$. But $D(\underbrace{P_{S_n}}_{\text{discrete}}\|\underbrace{\mathcal{N}(0,1)}_{\text{cont's}}) = \infty$ for all n. Note that this is an example for strict inequality in (3.2).

Note: Why do we care about continuity of information measures? Let's take divergence as an example.

1. *Computation.* For complicated P and Q direct computation of $D(P\|Q)$ might be hard. Instead, one may want to discretize them then let the computer compute. **Question**: Is this procedure stable, i.e., as the quantization becomes finer, does this procedure guarantee to converge to the true value? Yes! Continuity w.r.t. discretization is guaranteed by the next theorem.

2. *Estimating information measures.* In many statistical setups, oftentimes we do not know P or Q, if we estimate the distribution from data (e.g., estimate P by empirical distribution $\hat{P}_n$ from n samples) and then plug in, does $D(\hat{P}_n\|Q)$ provide a good estimator for $D(P\|Q)$? Well, note from the first example that this is a bad idea if Q is continuous, since $D(\hat{P}_n\|Q) = \infty$ for n. In fact, if one convolves the empirical distribution with a tiny bit of, say, Gaussian distribution, then it will always have a density. If we allow the variance of the Gaussian to vanish with n appropriately, we will have convergence. This leads to the idea of *kernel density estimators*. All these need regularity properties of divergence.

3.4 Variational characterizations of divergence: Gelfand-Yaglom-Perez

The point of the following theorem is that divergence on general alphabets can be defined via divergence on finite alphabets and discretization. Moreover, as the quantization becomes finer, we approach the value of divergence.

Theorem 3.7 (Gelfand-Yaglom-Perez)**.** *Let P, Q be two probability measures on $\mathcal{X}$ with σ-algebra $\mathcal{F}$. Then*

$$D(P\|Q) = \sup_{\{E_1,\dots,E_n\}} \sum_{i=1}^{n} P[E_i] \log \frac{P[E_i]}{Q[E_i]}\,, \tag{3.3}$$

where the supremum is over all finite $\mathcal{F}$-measurable partitions: $\bigcup_{j=1}^{n} E_j = \mathcal{X}, E_j \cap E_i = \varnothing$, and $0 \log \frac{1}{0} = 0$ and $\log \frac{1}{0} = \infty$ per our usual convention.

Remark 3.1. This theorem, in particular, allows us to prove all general identities and inequalities for the cases of discrete random variables.

Proof. "≥": Fix a finite partition $E_1, \dots E_n$. Define a function (quantizer/discretizer) $f : \mathcal{X} \to \{1, \dots, n\}$ as follows: For any x, let $f(x)$ denote the index j of the set E_j to which X belongs. Let X be distributed according to either P or Q and set $Y = f(X)$. Applying data processing inequality for divergence yields

$$\begin{aligned} D(P\|Q) &= D(P_X\|Q_X) \\ &\geq D(P_Y\|Q_Y) \\ &= \sum_i P[E_i] \log \frac{P[E_i]}{Q[E_i]}\,. \end{aligned} \tag{3.4}$$

"≤": To show $D(P\|Q)$ is indeed achievable, first note that if $P \not\ll Q$, then by definition, there exists B such that $Q(B) = 0 < P(B)$. Choosing the partition $E_1 = B$ and $E_2 = B^c$, we have $D(P\|Q) = \infty = \sum_{i=1}^{2} P[E_i] \log \frac{P[E_i]}{Q[E_i]}$. In the sequel we assume that $P \ll Q$, hence the likelihood ratio $\frac{dP}{dQ}$ is well-defined. Let us define a partition of $\mathcal{X}$ by partitioning the range of $\log \frac{dP}{dQ}$: $E_j = \{x : \log \frac{dP}{dQ} \in \epsilon \cdot [j - n/2, j + 1 - n/2)\}$, $j = 1, \dots, n-1$ and $E_n = \{x : \log \frac{dP}{dQ} < 1 - n/2$ or $\log \frac{dP}{dQ} \geq n/2)\}$.[1]Note that on E_j, $\log \frac{dP}{dQ} \leq \epsilon(j + 1 - n/2) \leq \log \frac{P(E_j)}{Q(E_j)} + \epsilon$. Hence $\sum_{j=1}^{n-1} \int_{E_j} dP \log \frac{dP}{dQ} \leq \sum_{j=1}^{n-1} \epsilon P(E_j) + P(E_j) \log \frac{P(E_j)}{Q(E_j)} \leq \epsilon + \sum_{j=1}^{n} \epsilon P(E_j) + P(E_j) \log \frac{P(E_j)}{Q(E_j)} + P(E_n) \log \frac{1}{P(E_n)}$. In other words, $\sum_{j=1}^{n} P(E_j) \log \frac{P(E_j)}{Q(E_j)} \geq \int_{E_n^c} dP \log \frac{dP}{dQ} - \epsilon - P(E_n) \log \frac{1}{P(E_n)}$. Let $n \to \infty$ and $\epsilon \to 0$ be such that $n\epsilon \to \infty$ (e.g., $\epsilon = 1/\sqrt{n}$). The proof is complete by noting that $P(E_n) \to 0$ and $\int \mathbf{1}_{\left\{|\log \frac{dP}{dQ}| \leq \epsilon n\right\}} dP \log \frac{dP}{dQ} \xrightarrow{\epsilon n \uparrow \infty} \int dP \log \frac{dP}{dQ} = D(P\|Q)$. □

3.5 Continuity of divergence. Dependence on σ-algebra.

For finite alphabet $\mathcal{X}$ it is easy to establish continuity of entropy and divergence:

Proposition 3.1. *Let $\mathcal{X}$ be finite, fix distribution Q on $\mathcal{X}$ with $Q(x) > 0$ for all $x \in \mathcal{X}$. Then map*

$$P \mapsto D(P\|Q)$$

is continuous. In particular,

$$P \mapsto H(P) \tag{3.5}$$

is continuous.

[1] *Intuition*: The main idea is to note that the loss in the inequality (3.4) is in fact $D(P_X\|Q_X) = D(P_Y\|Q_Y) + D(P_{X|Y}\|Q_{X|Y}|P_Y)$, and we want to show that the conditional divergence is small. Note that $P_{X|Y=j} = P_{X|X\in E_j}$ and $Q_{X|Y=j} = Q_{X|X\in E_j}$. Hence $\frac{dP_{X|Y=j}}{dQ_{X|Y=j}} = \frac{dP}{dQ}\frac{Q(E_j)}{P(E_j)}\mathbf{1}_{E_j}$. Once we partitioned the likelihood ratio sufficiently finely, these two conditional distribution are very close to each other.

Warning: Divergence is never continuous in the pair, even for finite alphabets: $d(\frac{1}{n}\|2^{-n}) \not\to 0$.

Proof. Notice that

$$D(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}$$

and each term is a continuous function of $P(x)$. □

Our next goal is to study continuity properties of divergence for general alphabets. First, however, we need to understand dependence on the σ-algebra of the space. Indeed, divergence $D(P\|Q)$ implicitly depends on the σ-algebra $\mathcal{F}$ defining the measurable space $(\mathcal{X},\mathcal{F})$. To emphasize the dependence on $\mathcal{F}$ we will write

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}})\,.$$

We want to understand how does $D(P_{\mathcal{F}}\|Q_{\mathcal{F}})$ depend upon refining $\mathcal{F}$. Notice that we can even define $D(P_{\mathcal{F}}\|Q_{\mathcal{F}})$ for any *algebra* of sets $\mathcal{F}$ and two positive additive set-functions P, Q on $\mathcal{F}$. For this we take (3.3) as the definition. Note that when $\mathcal{F}$ is not a σ-algebra or P, Q are not σ-additive, we do not have Radon-Nikodym theorem and thus our original definition is not applicable.

Corollary 3.2 (Measure-theoretic properties of divergence)**.** *Let P, Q be probability measures on the measurable space $(\mathcal{X},\mathcal{H})$. Assume all algebras below are sub-algebras of $\mathcal{H}$. Then:*

- *(Monotonicity) If $\mathcal{F}\subseteq\mathcal{G}$ then*

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) \le D(P_{\mathcal{G}}\|Q_{\mathcal{G}})\,. \tag{3.6}$$

- *Let $\mathcal{F}_1\subseteq\mathcal{F}_2\ldots$ be an increasing sequence of algebras and let $\mathcal{F}=\bigcup_n\mathcal{F}_n$ be their limit, then*

$$D(P_{\mathcal{F}_n}\|Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}\|Q_{\mathcal{F}})\,.$$

- *If $\mathcal{F}$ is $(P+Q)$-dense in $\mathcal{G}$ then*[2]

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) = D(P_{\mathcal{G}}\|Q_{\mathcal{G}})\,. \tag{3.7}$$

- *(Monotone convergence theorem) Let $\mathcal{F}_1\subseteq\mathcal{F}_2\ldots$ be an increasing sequence of algebras and let $\mathcal{F}=\bigvee_n\mathcal{F}_n$ be the σ-algebra generated by them, then*

$$D(P_{\mathcal{F}_n}\|Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}\|Q_{\mathcal{F}})\,.$$

  In particular,

$$D(P_{X^\infty}\|Q_{X^\infty}) = \lim_{n\to\infty} D(P_{X^n}\|Q_{X^n})\,.$$

- *(Lower-semicontinuity of divergence) If $P_n\to P$ and $Q_n\to Q$ pointwise on the algebra $\mathcal{F}$, then*[3]

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) \le \liminf_{n\to\infty} D(P_{n,\mathcal{F}}\|Q_{n,\mathcal{F}})\,. \tag{3.8}$$

Proof. Straightforward applications of (3.3) and the observation that any algebra $\mathcal{F}$ is μ-dense in the σ-algebra $\sigma\{\mathcal{F}\}$ it generates, for any μ on $(\mathcal{X},\mathcal{H})$.[4] □

Note: Pointwise convergence on $\mathcal{H}$ is weaker than convergence in total variation and stronger than convergence in distribution (aka "weak convergence"). However, (3.8) can be extended to this mode of convergence (see Theorem 3.6).

[2] Note: $\mathcal{F}$ is μ-dense in $\mathcal{G}$ if $\forall E\in\mathcal{G}, \epsilon>0 \exists E'\in\mathcal{F}$ s.t. $\mu[E\Delta E']\le\epsilon$.

[3] $P_n\to P$ pointwise on some algebra $\mathcal{F}$ if $\forall E\in\mathcal{F}: P_n[E]\to P[E]$.

[4] This may be shown by transfinite induction: to each ordinal ω associate an algebra $\mathcal{F}_\omega$ generated by monotone limits of sets from $\mathcal{F}_{\omega'}$ with $\omega'<\omega$. Then $\sigma\{\mathcal{F}\}=\mathcal{F}_{\omega_0}$, where ω_0 is the first ordinal for which $\mathcal{F}_\omega$ is a monotone class. But $\mathcal{F}$ is μ-dense in each $\mathcal{F}_\omega$ by transfinite induction.

3.6 Variational characterizations and continuity of mutual information

Again, similarly to Proposition 3.1, it is easy to show that in the case of finite alphabets mutual information is continuous in the distribution:

Proposition 3.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite alphabets. Then*

$$P_{X,Y} \mapsto I(X;Y)$$

is continuous.

Proof. Apply representation

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

and (3.5). □

Further properties of mutual information follow from $I(X;Y) = D(P_{XY} \| P_X P_Y)$ and corresponding properties of divergence, e.g.

1. $$I(X;Y) = \sup_f \mathbb{E}[f(X,Y)] - \log \mathbb{E}[\exp\{f(X,\bar{Y})\}]\,,$$

   where $\bar{Y}$ is a copy of Y, independent of X and supremum is over bounded, or even bounded continuous functions.

2. If $(X_n, Y_n) \overset{d}{\to} (X,Y)$ converge in distribution, then

   $$I(X;Y) \le \liminf_{n\to\infty} I(X_n;Y_n)\,. \tag{3.9}$$

   Good example of strict inequality: $X_n = Y_n = \frac{1}{n}Z$. In this case $(X_n,Y_n) \overset{d}{\to} (0,0)$ but $I(X_n;Y_n) = H(Z) > 0 = I(0;0)$.

3. $$I(X;Y) = \sup_{\{E_i\}\times\{F_j\}} \sum_{i,j} P_{XY}[E_i \times F_j] \log \frac{P_{XY}[E_i \times F_j]}{P_X[E_i]P_Y[F_j]}\,,$$

   where supremum is over finite partitions of spaces $\mathcal{X}$ and $\mathcal{Y}$.[5]

4. (Monotone convergence):

   $$I(X^\infty;Y) = \lim_{n\to\infty} I(X^n;Y) \tag{3.10}$$

   $$I(X^\infty;Y^\infty) = \lim_{n\to\infty} I(X^n;Y^n) \tag{3.11}$$

   This implies that all mutual information between two-processes X^∞ and Y^∞ is contained in their finite-dimensional projections, leaving nothing for the tail σ-algebra.

[5] To prove this from (3.3) one needs to notice that algebra of measurable rectangles is dense in the product σ-algebra.

4.1 Convexity of information measures

Theorem 4.1. *$(P,Q) \mapsto D(P\|Q)$ is convex.*

Proof. First proof: Let $X \in \{0,1\}$, $P_X = [\lambda, 1-\lambda]$. Select two conditional kernels:

$$\begin{aligned} P_{Y|X=0} &= P_0, \quad P_{Y|X=1} = P_1 && (4.1)\\ Q_{Y|X=0} &= Q_0, \quad Q_{Y|X=1} = Q_1 && (4.2)\end{aligned}$$

Conditioning increases divergence, hence

$$D(P_{Y|X}\|Q_{Y|X}|P_X) \ge D(P_Y\|Q_Y)$$

Second proof: $(p,q) \to p\log\frac{p}{q}$ is convex on $\mathbb{R}_+^2$ [Verify by computing the Hessian matrix and showing that it is positive semidefinite][1]

Third proof: By the Donsker-Varadhan variational representation,

$$D(P\|Q) = \sup_{f\in\mathcal{C}} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp\{f(X)\}]\,.$$

where for fixed f, $P \to \mathbb{E}_P[f(X)]$ is affine (hence convex), $Q \mapsto \log\mathbb{E}_Q[\exp\{f(X)\}]$ is concave. Therefore $(P,Q) \mapsto D(P\|Q)$ is pointwise supremum of convex functions, hence convex. □

Remark 4.1. The first proof shows that for an arbitrary measure of similarity $\mathcal{D}(P\|Q)$ convexity of $(P,Q) \mapsto \mathcal{D}(P\|Q)$ is *equivalent* to "conditioning increases divergence" property of $\mathcal{D}$. Convexity can also be understood as "mixing decreases divergence".

Remark 4.2 (f-divergences)**.** Any f-divergence, cf. (1.15), satisfies all the key properties of the usual divergence: positivity, monotonicity, data processing (DP), conditioning increases divergence (CID) and convexity in the pair. Indeed, by previous remark the last two are equivalent. Furthermore, proof of Theorem 2.2 showed that DP and CID are implied by monotonicity. Thus, consider P_{XY} and Q_{XY} and note

$$\begin{aligned} D_f(P_{XY}\|Q_{XY}) &= \mathbb{E}_{Q_{XY}}\left[f\left(\frac{P_{XY}}{Q_{XY}}\right)\right] && (4.3)\\ &= \mathbb{E}_{Q_Y}\mathbb{E}_{Q_{X|Y}}\left[f\left(\frac{P_Y}{Q_Y}\cdot\frac{P_{X|Y}}{Q_{X|Y}}\right)\right] && (4.4)\\ &\ge \mathbb{E}_{Q_Y}\left[f\left(\frac{P_Y}{Q_Y}\right)\right], && (4.5)\end{aligned}$$

where inequality follows by applying Jensen's inequality to convex function f. Finally, positivity follows from monotonicity by taking Y to be a constant and recalling that $f(1) = 0$.

[1] This is a general phenomenon: for a convex $f(\cdot)$ the *perspective* function $(p,q) \mapsto qf\left(\frac{p}{q}\right)$ is convex too.

Theorem 4.2 (Entropy)**.** *$P_X \mapsto H(P_X)$ is concave.*

Proof. If P_X is on a finite alphabet, then proof is complete by $H(X) = \log|\mathcal{X}| - D(P_X\|U_X)$. Otherwise, set

$$P_{X|Y} = \begin{cases} P_0 & Y = 0 \\ P_1 & Y = 1 \end{cases}, \; P_Y(Y=0) = \lambda$$

Then apply $H(X|Y) \le H(X)$. □

Recall that $I(X,Y)$ is a function of P_{XY}, or equivalently, $(P_X, P_{Y|X})$. Denote $I(P_X, P_{Y|X}) = I(X;Y)$.

Theorem 4.3 (Mutual Information)**.**

- *For fixed $P_{Y|X}$, $P_X \mapsto I(P_X, P_{Y|X})$ is concave.*
- *For fixed P_X, $P_{Y|X} \mapsto I(P_X, P_{Y|X})$ is convex.*

Proof.

- *First proof*: Introduce $\theta \in \text{Bern}(\lambda)$. Define $P_{X|\theta=0} = P_X^0$ and $P_{X|\theta=1} = P_X^1$. Then $\theta \to X \to Y$. Then $P_X = \bar{\lambda}P_X^0 + \lambda P_X^1$. $I(X;Y) = I(X,\theta;Y) = I(\theta;Y) + I(X;Y|\theta) \ge I(X;Y|\theta)$, which is our desired $I(\bar{\lambda}P_X^0 + \lambda P_X^1, P_{Y|X}) \ge \bar{\lambda}I(P_X^0, P_{Y|X}) + \lambda I(P_X^0, P_{Y|X})$.

  Second proof: $I(X;Y) = \min_Q D(P_{Y|X}\|Q|P_X)$ – pointwise minimum of affine functions is concave.

  Third proof: Pick a Q and use the golden formula: $I(X;Y) = D(P_{Y|X}\|Q|P_X) - D(P_Y\|Q)$, where $P_X \mapsto D(P_Y\|Q)$ is convex, as the composition of the $P_X \mapsto P_Y$ (affine) and $P_Y \mapsto D(P_Y\|Q)$ (convex).
- $I(X;Y) = D(P_{Y|X}\|P_Y|P_X)$ □

4.2* Local behavior of divergence

Due to smoothness of the function $(p,q) \mapsto p\log\frac{p}{q}$ at $(1,1)$ it is natural to expect that the functional

$$P \mapsto D(P\|Q)$$

should also be smooth as $P \to Q$. Due to non-negativity and convexity, it is then also natural to expect that this functional decays quadratically. In this section, we show that generally decay is sublinear and it is quadratic in the special case when $\chi^2(P\|Q) < \infty$ (see below).

Proposition 4.1. *When $D(P\|Q) < \infty$, the one-sided derivative in $\lambda = 0$ vanishes:*

$$\frac{d}{d\lambda}\Big|_{\lambda=0} D(\lambda P + \bar{\lambda}Q\|Q) = 0$$

Proof.

$$\frac{1}{\lambda}D(\lambda P+\bar{\lambda}Q\|Q)=\mathbb{E}_Q\left[\frac{1}{\lambda}(\lambda f+\bar{\lambda})\log(\lambda f+\bar{\lambda})\right]$$

where $f=\dfrac{dP}{dQ}$. As $\lambda\to 0$ the function under expectation decreases to $(f-1)\log e$ monotonically. Indeed, the function

$$\lambda\mapsto g(\lambda)\triangleq(\lambda f+\bar{\lambda})\log(\lambda f+\bar{\lambda})$$

is convex and equals zero at $\lambda=0$. Thus $\frac{g(\lambda)}{\lambda}$ is increasing in λ. Moreover, by convexity of $x\mapsto x\log x$

$$\frac{1}{\lambda}(\lambda f+\bar{\lambda})(\log(\lambda f+\bar{\lambda}))\le\frac{1}{\lambda}(\lambda f\log f+\bar{\lambda}1\log 1)=f\log f$$

and by assumption $f\log f$ is Q-integrable. Thus the Monotone Convergence Theorem applies. □

Note: More generally, under suitable technical conditions,

$$\frac{d}{d\lambda}\Big|_{\lambda=0}D(\lambda P+\bar{\lambda}Q\|R)=\mathbb{E}_P\left[\log\frac{dQ}{dR}\right]-D(Q\|R)\,.$$

and

$$\frac{d}{d\lambda}\Big|_{\lambda=0}D(\bar{\lambda}P_1+\lambda Q_1\|\bar{\lambda}P_0+\lambda Q_0)=\mathbb{E}_{Q_1}\left[\log\frac{dP_1}{dP_0}\right]-D(P_1\|P_0)+\mathbb{E}_{P_1}\left[1-\frac{dQ_0}{dP_0}\right]\log e$$

The message of Proposition 4.1 is that the function

$$\lambda\mapsto D(\lambda P+\bar{\lambda}Q\|Q)\,,$$

is $o(\lambda)$ as $\lambda\to 0$. In fact, in most cases it is quadratic in λ. To state a precise version, we need to define the concept of χ^2-divergence – a version of f-divergence (1.15):

$$\chi^2(P\|Q)\triangleq\int dQ\left(\frac{dP}{dQ}-1\right)^2\,.$$

This is a very popular measure of distance between P and Q, frequently used in statistics. It has many important properties, but we will only mention that χ^2 dominates KL-divergence:

$$D(P\|Q)\le\log(1+\chi^2(P\|Q))\,.$$

Our second result about local properties of KL-divergence is the following:

Proposition 4.2 (KL is locally χ^2-like)**.** *If* $\chi^2(P\|Q)<\infty$ *then*

$$D(\lambda P+\bar{\lambda}Q\|Q)=\frac{\lambda^2\log e}{2}\chi^2(P\|Q)+o(\lambda^2)\,,\qquad\lambda\to 0\,.$$

Proof. First, notice that

$$D(P\|Q)=\mathbb{E}_Q\left[g\left(\frac{dP}{dQ}\right)\right],$$

where

$$g(x)\triangleq x\log x-(x-1)\log e\,.$$

Note that $x\mapsto\frac{g(x)}{(x-1)^2\log e}=\int_0^1\frac{s ds}{x(1-s)+s}$ is decreasing in x on $(0,\infty)$. Therefore

$$0\le g(x)\le(x-1)^2\log e\,,$$

and hence

$$0 \le \frac{1}{\lambda^2} g\left(\bar{\lambda} + \lambda\frac{dP}{dQ}\right) \le \left(\frac{dP}{dQ} - 1\right)^2 \log e.$$

By the dominated convergence theorem (which is applicable since $\chi^2(P\|Q) < \infty$) we have

$$\lim_{\lambda\to 0} \frac{1}{\lambda^2}\mathbb{E}_Q\left[g\left(\bar{\lambda} + \lambda\frac{dP}{dQ}\right)\right] = \frac{g''(1)}{2}\mathbb{E}_Q\left[\left(\frac{dP}{dQ} - 1\right)^2\right] = \frac{\log e}{2}\chi^2(P\|Q)\,.$$

□

4.3* Local behavior of divergence and Fisher information

Consider a parameterized set of distributions $\{P_\theta, \theta \in \Theta\}$ and assume Θ is an open subset of $\mathbb{R}^d$. Furthermore, suppose that distribution P_θ are all given in the form of

$$P_\theta(dx) = f(x|\theta)\mu(dx)\,,$$

where μ is some common dominating measure (e.g. Lebesgue or counting). If for a fixed x functions $\theta \to f(x|\theta)$ are smooth, one can define Fisher information matrix with respect to parameter θ as

$$J_F(\theta) \triangleq \mathbb{E}_{X\sim P_\theta}\left[VV^T\right], \quad V \triangleq \nabla_\theta \log f(X|\theta)\,. \tag{4.6}$$

Under suitable regularity conditions, Fisher information matrix has several equivalent expressions:

$$\begin{aligned} J_F(\theta) &= \mathrm{cov}_{X\sim P_\theta}\left[\nabla_\theta \log f(X|\theta)\right] && (4.7)\\ &= (4\log e)\int \mu(dx)(\nabla_\theta\sqrt{f(x|\theta)})(\nabla_\theta\sqrt{f(x|\theta)})^T && (4.8)\\ &= -(\log e)\,\mathbb{E}_\theta[\mathrm{Hess}_\theta(\log f(X|\theta))]\,, && (4.9)\end{aligned}$$

where the latter is obtained by differentiating

$$0 = \int \mu(dx) f(x|\theta)\frac{\partial}{\partial\theta_i}\log f(x|\theta)$$

in θ_j.

Trace of this matrix is called Fisher information and similarly can be expressed in a variety of forms:

$$\begin{aligned} \mathrm{tr}\, J_F(\theta) &= \int \mu(dx)\frac{\|\nabla_\theta f(x|\theta)\|^2}{f(x|\theta)} && (4.10)\\ &= 4\int \mu(dx)\|\nabla_\theta\sqrt{f(x|\theta)}\|^2 && (4.11)\\ &= -(\log e)\cdot\mathbb{E}_{X\sim P_\theta}\left[\sum_{i=1}^d \frac{\partial^2}{\partial\theta_i\partial\theta_i}\log f(X|\theta)\right], && (4.12)\end{aligned}$$

Significance of Fisher information matrix arises from the fact that it gauges the local behaviour of divergence for smooth parametric families. Namely, we have (again under suitable technical conditions):

$$D(P_{\theta_0}\|P_{\theta_0+\xi}) = \frac{1}{2\log e}\xi^T J_F(\theta_0)\xi + o(\|\xi\|^2)\,, \tag{4.13}$$

which is obtained by integrating the Taylor expansion:

$$\log f(x|\theta_0+\xi) = \log f(x|\theta_0) + \xi^T \nabla_\theta \log f(x|\theta_0) + \frac{1}{2}\xi^T \mathrm{Hess}_\theta(\log f(x|\theta_0))\xi + o(\|\xi\|^2)\,.$$

Property (4.13) is of paramount importance in statistics. We should remember it as: *Divergence is locally quadratic on the parameter space, with Hessian given by the Fisher information matrix.*

Remark 4.3. It can be seen that if one introduces another parametrization $\tilde\theta \in \tilde\Theta$ by means of a smooth invertible map $\tilde\Theta \to \Theta$, then Fisher information matrix changes as

$$J_F(\tilde\theta) = A^T J_F(\theta) A\,, \tag{4.14}$$

where $A = \frac{d\theta}{d\tilde\theta}$ is the Jacobian of the map. So we can see that J_F transforms similarly to the metric tensor in Riemannian geometry. This idea can be used to define a Riemannian metric on the space of parameters Θ, called Fisher-Rao metric. This is explored in a field known as information geometry [AN07].

Example: Consider Θ to be the interior of a simplex of all distributions on a finite alphabet $\{0,\ldots,d\}$. We will take $\theta_1,\ldots,\theta_d$ as free parameters and set $\theta_0 = 1 - \sum_{i=1}^d \theta_i$. So all derivatives are with respect to $\theta_1,\ldots,\theta_d$ only. Then we have

$$P_\theta(x) = f(x|\theta) = \begin{cases} \theta_x, & x = 1,\ldots,d \\ 1 - \sum_{x\neq 0}\theta_x, & x = 0 \end{cases}$$

and for Fisher information matrix we get

$$J_F(\theta) = (\log^2 e)\left\{\mathrm{diag}(\frac{1}{\theta_1},\ldots,\frac{1}{\theta_d}) + \frac{1}{1-\sum_{i=1}^d \theta_i} 1\cdot 1^T\right\}\,, \tag{4.15}$$

where $1\cdot 1^T$ is the $d\times d$ matrix of all ones. For future reference, we also compute determinant of $J_F(\theta)$. To that end notice that $\det(A + xy^T) = \det A \cdot \det(I + A^{-1}xy^T) = \det A\cdot(1 + y^T A^{-1}x)$, where we used the identity $\det(I+AB) = \det(I+BA)$. Thus, we have

$$\det J_F(\theta) = (\log e)^{2d} \prod_{x=0}^d \frac{1}{\theta_x} = (\log e)^{2d} \frac{1}{1-\sum_{x=1}^d \theta_x} \prod_{x=1}^d \frac{1}{\theta_x}\,. \tag{4.16}$$

4.4 Extremization of mutual information

Two problems of interest

- Fix $P_{Y|X} \to \max_{P_X} I(X;Y)$ — channel coding

  Note: This maximum is called "capacity" of a set of distributions $\{P_{Y|X=x}, x\in\mathcal{X}\}$.
- Fix $P_X \to \min_{P_{Y|X}} I(X;Y)$ — lossy compression

Theorem 4.4 (Saddle point)**.** *Let $\mathcal{P}$ be a convex set of distributions on $\mathcal{X}$. Suppose there exists $P_X^* \in \mathcal{P}$ such that*

$$\sup_{P_X\in\mathcal{P}} I(P_X, P_{Y|X}) = I(P_X^*, P_{Y|X}) \triangleq C$$

and let $P_X^* \xrightarrow{P_{Y|X}} P_Y^*$. Then for all $P_X \in \mathcal{P}$ and for all Q_Y, we have

$$D(P_{Y|X}\|P_Y^*|P_X) \le D(P_{Y|X}\|P_Y^*|P_X^*) \le D(P_{Y|X}\|Q_Y|P_X^*). \tag{4.17}$$

Note: P_X^* (resp., P_Y^*) is called a capacity-achieving input (resp., output) distribution, or a *caid* (resp., the *caod*).

Proof. Right inequality: obvious from $C = I(P_X^*, P_{Y|X}) = \min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X^*)$.

Left inequality: If $C = \infty$, then trivial. In the sequel assume that $C < \infty$, hence $I(P_X, P_{Y|X}) < \infty$ for all $P_X \in \mathcal{P}$. Let $P_{X_\lambda} = \lambda P_X + \bar{\lambda} P_X^* \in \mathcal{P}$ by convexity of $\mathcal{P}$, and introduce $\theta \sim \mathrm{Bern}(\lambda)$, so that $P_{X_\lambda|\theta=0} = P_X^*$, $P_{X_\lambda|\theta=1} = P_X$, and $\theta \to X_\lambda \to Y_\lambda$. Then

$$\begin{aligned} C \ge I(X_\lambda; Y_\lambda) &= I(\theta, X_\lambda; Y_\lambda) = I(\theta; Y_\lambda) + I(X_\lambda; Y_\lambda|\theta) \\ &= D(P_{Y_\lambda|\theta}\|P_{Y_\lambda}|P_\theta) + \lambda I(P_X, P_{Y|X}) + \bar{\lambda} C \\ &= \lambda D(P_Y\|P_{Y_\lambda}) + \bar{\lambda} D(P_Y^*\|P_{Y_\lambda}) + \lambda I(P_X, P_{Y|X}) + \bar{\lambda} C \\ &\ge \lambda D(P_Y\|P_{Y_\lambda}) + \lambda I(P_X, P_{Y|X}) + \bar{\lambda} C. \end{aligned}$$

Since $I(P_X, P_{Y|X}) < \infty$, we can subtract it to obtain

$$\lambda(C - I(P_X, P_{Y|X})) \ge \lambda D(P_Y\|P_{Y_\lambda}).$$

Dividing both sides by λ, taking the lim inf and using lower semicontinuity of D, we have

$$\begin{aligned} & C - I(P_X, P_{Y|X}) \ge \liminf_{\lambda\to 0} D(P_Y\|P_{Y_\lambda}) \ge D(P_Y\|P_Y^*) \\ \Longrightarrow\ & C \ge I(P_X, P_{Y|X}) + D(P_Y\|P_Y^*) = D(P_{Y|X}\|P_Y|P_X) + D(P_Y\|P_Y^*) = D(P_{Y|X}\|P_Y^*|P_X). \end{aligned}$$

Here is an even shorter proof:

$$C \ge I(X_\lambda; Y_\lambda) = D(P_{Y|X}\|P_{Y_\lambda}|P_{X_\lambda}) \tag{4.18}$$
$$= \lambda D(P_{Y|X}\|P_{Y_\lambda}|P_X) + \bar{\lambda} D(P_{Y|X}\|P_{Y_\lambda}|P_X^*) \tag{4.19}$$
$$\ge \lambda D(P_{Y|X}\|P_{Y_\lambda}|P_X) + \bar{\lambda} C \tag{4.20}$$
$$= \lambda D(P_{X,Y}\|P_X P_{Y_\lambda}) + \bar{\lambda} C, \tag{4.21}$$

where inequality is by the right part of (4.17) (already shown). Thus, subtracting $\bar{\lambda} C$ and dividing by λ we get

$$D(P_{X,Y}\|P_X P_{Y_\lambda}) \le C$$

and the proof is completed by taking $\liminf_{\lambda\to 0}$ and applying lower semincontinuity of divergence. □

Corollary 4.1. *In addition to the assumptions of Theorem 4.4, suppose $C < \infty$. Then caod P_Y^* is unique. It satisfies the property that for any P_Y induced by some $P_X \in \mathcal{P}$ (i.e. $P_Y = P_{Y|X} \circ P_X$) we have*

$$D(P_Y\|P_Y^*) \le C < \infty \tag{4.22}$$

and in particular $P_Y \ll P_Y^*$.

Proof. The statement is: $I(P_X, P_{Y|X}) = C \Rightarrow P_Y = P_Y^*$. Indeed:

$$\begin{aligned} C = D(P_{Y|X}\|P_Y|P_X) &= D(P_{Y|X}\|P_Y^*|P_X) - D(P_Y\|P_Y^*) \\ &\le D(P_{Y|X}\|P_Y^*|P_X^*) - D(P_Y\|P_Y^*) \\ &= C - D(P_Y\|P_Y^*) \Rightarrow P_Y = P_Y^* \end{aligned}$$

Statement (4.22) follows from the left inequality in (4.17) and "conditioning increases divergence". □

Notes:

- Finiteness of C is necessary. Counterexample: The identity channel $Y = X$, where X takes values on integers. Then any distribution with infinite entropy is caid or caod.
- *Non-uniqueness of caid.* Unlike the caod, caid does not need to be unique. Let $Z_1 \sim \text{Bern}(\frac{1}{2})$. Consider $Y_1 = X_1 \oplus Z_1$ and $Y_2 = X_2$. Then $\max_{P_{X_1X_2}} I(X_1, X_2; Y_1, Y_2) = \log 4$, achieved by $P_{X_1X_2} = \text{Bern}(p) \times \text{Bern}(\frac{1}{2})$ for any p. Note that the *caod* is unique: $P^*_{Y_1Y_2} = \text{Bern}(\frac{1}{2}) \times \text{Bern}(\frac{1}{2})$.

Review: Minimax and saddlepoint

Suppose we have a bivariate function f. Then we always have the *minimax inequality*:

$$\inf_y \sup_x f(x,y) \geq \sup_x \inf_y f(x,y).$$

When does it hold with equality?

1. It turns out minimax equality is implied by the existence of a saddle point (x^*, y^*), i.e.,
$$f(x, y^*) \leq f(x^*, y^*) \leq f(x^*, y) \qquad \forall x, y$$
Furthermore, minimax equality also implies existence of saddle point if inf and sup are achieved c.f. [BNO03, Section 2.6]) for all x, y [Straightforward to check. See proof of corollary below].
2. There are a number of known criteria establishing
$$\inf_y \sup_x f(x,y) = \sup_x \inf_y f(x,y)$$
They usually require some continuity of f, compactness of domains and convexity in x and concavity in y. One of the most general version is due to M. Sion [Sio58].
3. The mother result of all this minimax theory is a theorem of von Neumann on bilinear functions: Let A and B have finite alphabets, and $g(a,b)$ be arbitrary, then
$$\min_{P_A} \max_{P_B} \mathbb{E}[g(A,B)] = \max_{P_B} \min_{P_A} \mathbb{E}[g(A,B)]$$
Here $(x,y) \leftrightarrow (P_A, P_B)$ and $f(x,y) \leftrightarrow \sum_{a,b} P_A(a) P_B(b) g(a,b)$.
4. A more general version is: if $\mathcal{X}$ and $\mathcal{Y}$ are compact convex domains in $\mathbb{R}^n$, $f(x,y)$ continuous in (x,y), concave in x and convex in y then
$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f(x,y) = \min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} f(x,y)$$

Applying Theorem 4.4 to conditional divergence gives the following result.

Corollary 4.2 (Minimax)**.** *Under assumptions of Theorem 4.4, we have*

$$\begin{aligned}\max_{P_X \in \mathcal{P}} I(X;Y) &= \max_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) \\ &= \min_{Q_Y} \max_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X)\end{aligned}$$

Proof. This follows from saddle-point trivially: Maximizing/minimizing the leftmost/rightmost sides of (4.17) gives

$$\min_{Q_Y} \max_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) \le \max_{P_X \in \mathcal{P}} D(P_{Y|X} \| P_Y^* | P_X) \le D(P_{Y|X} \| P_Y^* | P_X^*)$$

$$\le \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X^*) \le \max_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X).$$

but by definition $\min \max \ge \max \min$. □

4.5 Capacity = information radius

Review: Radius and diameter

Let (X, d) be a metric space. Let A be a bounded subset.

1. *Radius* (aka Chebyshev radius) of A: the radius of the smallest ball that covers A, i.e., $\mathrm{rad}\,(A) = \inf_{y \in X} \sup_{x \in A} d(x, y)$.
2. *Diameter* of A: $\mathrm{diam}\,(A) = \sup_{x,y \in A} d(x, y)$.
3. Note that the radius and the diameter both measure how big/rich a set is.
4. From definition and triangle inequality we have
$$\frac{1}{2}\mathrm{diam}\,(A) \le \mathrm{rad}\,(A) \le \mathrm{diam}\,(A)$$
5. In fact, the rightmost upper bound can frequently be improved. A result of Bohnenblust [Boh38] shows that in $\mathbb{R}^n$ equipped with any norm we always have $\mathrm{rad}\,(A) \le \frac{n}{n+1}\mathrm{diam}\,(A)$. For $\mathbb{R}^n$ with ℓ_∞-norm the situation is even simpler: $\mathrm{rad}\,(A) = \frac{1}{2}\mathrm{diam}\,(A)$ (such spaces are called centrable).

The next simple corollary shows that capacity is just the radius of the set of distributions $\{P_{Y|X=x}, x \in \mathcal{X}\}$ when distances are measured by divergence (although, we remind, divergence is not a metric).

Corollary 4.3. *For fixed kernel $P_{Y|X}$, let $\mathcal{P} = \{$all dist. on $\mathcal{X}\}$ and $\mathcal{X}$ is finite, then*

$$\begin{aligned} \max_{P_X} I(X;Y) &= \max_x D(P_{Y|X=x} \| P_Y^*) \\ &= D(P_{Y|X=x} \| P_Y^*) \qquad \forall x : P_X^*(x) > 0\ . \end{aligned}$$

The last corollary gives a geometric interpretation to capacity: it equals the radius of the smallest divergence-"ball" that encompasses all distributions $\{P_{Y|X=x} : x \in \mathcal{X}\}$. Moreover, P_Y^* is a convex combination of some $P_{Y|X=x}$ and it is **equidistant** to those.

4.6 Existence of caod (general case)

We have shown above that the solution to

$$C = \sup_{P_X \in \mathcal{P}} I(X;Y)$$

can be a) interpreted as a saddle point; b) written in the minimax form and c) that caod P_Y^* is unique. This was all done under the extra assumption that supremum over P_X is attainable. It turns out, properties b) and c) can be shown without that extra assumption.

Theorem 4.5 (Kemperman)**.** *For any $P_{Y|X}$ and a convex set of distributions $\mathcal{P}$ such that*

$$C = \sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) < \infty \tag{4.23}$$

there exists a unique P_Y^* with the property that

$$C = \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| P_Y^* | P_X) . \tag{4.24}$$

Furthermore,

$$\begin{aligned} C &= \sup_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) && (4.25)\\ &= \min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) && (4.26)\\ &= \min_{Q_Y} \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| Q_Y) , \qquad \textit{(if } \mathcal{P} = \{\textit{all } P_X\}\textit{.)} && (4.27) \end{aligned}$$

Note: Condition (4.23) is automatically satisfied if there is any Q_Y such that

$$\sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) < \infty . \tag{4.28}$$

Example: *Non-existence of caid.* Let $Z \sim \mathcal{N}(0,1)$ and consider the problem

$$C = \sup_{\substack{P_X : \mathbb{E}[X]=0, \mathbb{E}[X^2]=P \\ \mathbb{E}[X^4]=s}} I(X; X+Z) . \tag{4.29}$$

If we remove the constraint $\mathbb{E}[X^4] = s$ the unique caid is $P_X = \mathcal{N}(0,P)$, see Theorem 4.6. When $s \neq 3P^2$ then such P_X is no longer inside the constraint set $\mathcal{P}$. However, for $s > 3P^2$ the maximum

$$C = \frac{1}{2}\log(1+P)$$

is still attainable. Indeed, we can add a small "bump" to the gaussian distribution as follows:

$$P_X = (1-p)\mathcal{N}(0,P) + p\delta_x ,$$

where $p \to 0$, $px^2 \to 0$ but $px^4 \to s - 3P^2 > 0$. This shows that for the problem (4.29) with $s > 3P^2$ the caid does not exist, the caod $P_Y^* = \mathcal{N}(0, 1+P)$ exists and unique as Theorem 4.5 postulates.

Proof of Theorem 4.5. Let P'_{X_n} be a sequence of input distributions achieving C, i.e., $I(P'_{X_n}, P_{Y|X}) \to C$. Let $\mathcal{P}_n$ be the convex hull of $\{P'_{X_1}, \ldots, P'_{X_n}\}$. Since $\mathcal{P}_n$ is a finite-dimensional simplex, the concave function $P_X \mapsto I(P_X, P_{Y|X})$ attains its maximum at some point $P_{X_n} \in \mathcal{P}_n$, i.e.,

$$I_n \triangleq I(P_{X_n}, P_{Y|X}) = \max_{P_X \in \mathcal{P}_n} I(P_X, P_{Y|X}) .$$

Denote by P_{Y_n} be the sequence of output distributions corresponding to P_{X_n}. We have then:

$$\begin{aligned} D(P_{Y_n} \| P_{Y_{n+k}}) &= D(P_{Y|X} \| P_{Y_{n+k}} | P_{X_n}) - D(P_{Y|X} \| P_{Y_n} | P_{X_n}) && (4.30)\\ &\le I(P_{X_{n+k}}, P_{Y|X}) - I(P_{X_n}, P_{Y|X}) && (4.31)\\ &\le C - I_n , && (4.32) \end{aligned}$$

where in (4.31) we applied Theorem 4.4 to $(\mathcal{P}_{n+k}, P_{Y_{n+k}})$. By the Pinsker-Csiszár inequality (1.14) and since $I_n \nearrow C$, we conclude that the sequence P_{Y_n} is Cauchy in total variation:

$$\sup_{k\ge1} \mathrm{TV}(P_{Y_n}, P_{Y_{n+k}}) \to 0\,, \qquad n \to \infty\,.$$

Since the space of probability distributions is complete in total variation, the sequence must have a limit point $P_{Y_n} \to P_Y^*$. By taking a limit as $k \to \infty$ in (4.32) and applying the lower semi-continuity of divergence (Theorem 3.6) we get

$$D(P_{Y_n} \| P_Y^*) \le \lim_{k\to\infty} D(P_{Y_n} \| P_{Y_{n+k}}) \le C - I_n\,,$$

and therefore, $P_{Y_n} \to P_Y^*$ in the (stronger) sense of $D(P_{Y_n} \| P_Y^*) \to 0$. Therefore,

$$D(P_{Y|X} \| P_Y^* | P_{X_n}) = I_n + D(P_{Y_n} \| P_Y^*) \to C\,. \tag{4.33}$$

Take any $P_X \in \bigcup_{k\ge1} \mathcal{P}_k$. Then $P_X \in \mathcal{P}_n$ for all sufficiently large n and thus by Theorem 4.4

$$D(P_{Y|X} \| P_{Y_n} | P_X) \le I_n \le C\,, \tag{4.34}$$

which by lower semi-continuity of divergence implies

$$D(P_{Y|X} \| P_Y^* | P_X) \le C\,. \tag{4.35}$$

Finally, to prove that (4.35) holds for arbitrary $P_X \in \mathcal{P}$, we may repeat the argument above with $\mathcal{P}_n$ replaced by $\tilde{\mathcal{P}}_n = \mathrm{conv}(P_X \cup \mathcal{P}_n)$, denoting the resulting sequences by $\tilde{P}_{X_n}, \tilde{P}_{Y_n}$ and the limit point by $\tilde{P}_Y^*$ we have:

$$\begin{aligned} D(P_{Y_n} \| \tilde{P}_{Y_n}) &= D(P_{Y|X} \| \tilde{P}_{Y_n} | P_{X_n}) - D(P_{Y|X} \| P_{Y_n} | P_{X_n}) & (4.36)\\ &\le C - I_n\,, & (4.37) \end{aligned}$$

where (4.37) follows from (4.35) since $P_{X_n} \in \tilde{\mathcal{P}}_n$. Hence taking limit as $n \to \infty$ we have $\tilde{P}_Y^* = P_Y^*$ and therefore (4.35) holds.

Finally, to see (4.26), note that by definition capacity as a max-min is at most the min-max, i.e.,

$$C = \sup_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) \le \min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) \le \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| P_Y^* | P_X) = C$$

in view of (4.34). □

Corollary 4.4. *Let $\mathcal{X}$ be countable and $\mathcal{P}$ a convex set of distributions on $\mathcal{X}$. If $\sup_{P_X \in \mathcal{P}} H(X) < \infty$ then*

$$\sup_{P_X \in \mathcal{P}} H(X) = \min_{Q_X} \sup_{P_X \in \mathcal{P}} \sum_x P_X(x) \log \frac{1}{Q_X(x)} < \infty$$

and the optimizer Q_X^* exist and is unique. If $Q_X^* \in \mathcal{P}$ then it is also a unique maximizer of $H(X)$.

Proof. Just apply Kemperman's result to channel $Y = X$. □

Example: Assume that $f : \mathbb{Z} \to \mathbb{R}$ is such that $\sum_{n\in\mathbb{Z}} \exp\{-\lambda f(n)\} < \infty$ for all $\lambda > 0$. Then

$$\max_{X:\mathbb{E}[f(X)]\le a} H(X) \le \inf_{\lambda>0} \lambda a + \log \sum_n \exp\{-\lambda f(n)\}\,.$$

This follows from taking $Q(n) = c\exp\{-\lambda f(n)\}$. This bound is often tight and achieved by $P_X(n) = c\exp\{-\lambda f(n)\}$, known as the Gibbs distribution for energy function f.

4.7 Gaussian saddle point

For additive noise, there is also a different kind of saddle point between P_X and the distribution of noise:

Theorem 4.6. *Let* $X_g \sim \mathcal{N}(0,\sigma_X^2)$, $N_g \sim \mathcal{N}(0,\sigma_N^2)$, $X_g \perp\!\!\!\perp N_g$. *Then:*

1. *"Gaussian capacity":*

$$C = I(X_g; X_g + N_g) = \frac{1}{2}\log\Big(1 + \frac{\sigma_X^2}{\sigma_N^2}\Big)$$

2. *"Gaussian input is the best": For all* $X \perp\!\!\!\perp N_g$ *and* $\mathsf{var}X \le \sigma_X^2$,

$$I(X; X + N_g) \le I(X_g; X_g + N_g),$$

   with equality iff $X \overset{\mathrm{D}}{=} X_g$.

3. *"Gaussian noise is the worst": For for all* N *s.t.* $\mathbb{E}[X_g N] = 0$ *and* $\mathbb{E}N^2 \le \sigma_N^2$,

$$I(X_g; X_g + N) \ge I(X_g; X_g + N_g),$$

   with equality iff $N \overset{\mathrm{D}}{=} N_g$ *and independent of* X_g.

Note: Intuitive remarks

1. For AWGN channel, Gaussian input is the most favorable. Indeed, immediately from the second statement we have

$$\max_{X:\mathsf{var}X \le \sigma_X^2} I(X; X + N_g) = \frac{1}{2}\log\Big(1 + \frac{\sigma_X^2}{\sigma_N^2}\Big)$$

   which is the capacity formula for the AWGN channel.

2. For Gaussian source, additive Gaussian noise is the worst in the sense that it minimizes the mutual information provided by the noisy version.

Proof. WLOG, assume all random variables have zero mean. Let $Y_g = X_g + N_g$. Define

$$g(x) = D(P_{Y_g|X_g=x} \| P_{Y_g}) = D(\mathcal{N}(x,\sigma_N^2) \| \mathcal{N}(0,\sigma_X^2 + \sigma_N^2)) = \underbrace{\frac{1}{2}\log\Big(1 + \frac{\sigma_X^2}{\sigma_N^2}\Big)}_{=C} + \frac{\log e}{2}\frac{x^2 - \sigma_X^2}{\sigma_X^2 + \sigma_N^2}$$

1. Compute $I(X_g; X_g + N_g) = \mathbb{E}[g(X_g)] = C$

2. Recall the inf-representation $I(X;Y) = \min_Q D(P_{Y|X} \| Q | P_X)$. Then

$$I(X; X + N_g) \le D(P_{Y_g|X_g} \| P_{Y_g} | P_X) = \mathbb{E}[g(X)] \le C < \infty\,.$$

   Furthermore, if $I(X; X + N_g)$ then uniqueness of caod, cf. Corollary 4.1, implies $P_Y = P_{Y_g}$. But $P_Y = P_X * \mathcal{N}(0,\sigma_N^2)$. Then it must be that $X \sim \mathcal{N}(0,\sigma_X^2)$ simply by considering characteristic functions:

$$\Psi_X(t) \cdot \mathrm{e}^{-\frac{1}{2}\sigma_N^2 t^2} = \mathrm{e}^{-\frac{1}{2}(\sigma_X^2 + \sigma_N^2)t^2} \Rightarrow \Psi_X(t) = \mathrm{e}^{-\frac{1}{2}\sigma_X^2 t^2} \Longrightarrow X \sim \mathcal{N}(0,\sigma_X^2)$$

3. Let $Y = X_g + N$ and let $P_{Y|X_g}$ be the respective kernel. [Note that here we only assume that N is *uncorrelated* with X_g, i.e., $\mathbb{E}[NX_g] = 0$, not necessarily independent.] Then

$$\begin{aligned}
I(X_g; X_g + N) &= D(P_{X_g|Y} \| P_{X_g} | P_Y) \\
&= D(P_{X_g|Y} \| P_{X_g|Y_g} | P_Y) + \mathbb{E} \log \frac{P_{X_g|Y_g}(X_g|Y)}{P_{X_g}(X_g)} \\
&\geq \mathbb{E} \log \frac{P_{X_g|Y_g}(X_g|Y)}{P_{X_g}(X_g)} && (4.38) \\
&= \mathbb{E} \log \frac{P_{Y_g|X_g}(Y|X_g)}{P_{Y_g}(Y)} && (4.39) \\
&= C + \frac{\log e}{2} \mathbb{E}\Big[\frac{Y^2}{\sigma_X^2 + \sigma_N^2} - \frac{N^2}{\sigma_N^2}\Big] && (4.40) \\
&= C + \frac{\log e}{2} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2} \Big(1 - \frac{\mathbb{E}N^2}{\sigma_N^2}\Big) && (4.41) \\
&\geq C\,, && (4.42)
\end{aligned}$$

where

- (4.39): $\frac{P_{X_g|Y_g}}{P_{X_g}} = \frac{P_{Y_g|X_g}}{P_{Y_g}}$
- (4.41): $\mathbb{E}[X_g N] = 0$ and $\mathbb{E}[Y^2] = \mathbb{E}[N^2] + \mathbb{E}[X_g^2]$.
- (4.42): $\mathbb{E}N^2 \leq \sigma_N^2$.

Finally, the conditions for equality in (4.38) say

$$D(P_{X_g|Y} \| P_{X_g|Y_g} | P_Y) = 0$$

Thus, $P_{X_g|Y} = P_{X_g|Y_g}$, i.e., X_g is conditionally Gaussian: $P_{X_g|Y=y} = \mathcal{N}(by, c^2)$ for some constant b, c. In other words, under P_{X_gY}, we have

$$X_g = bY + cZ \quad , \quad Z \sim \text{Gaussian} \perp\!\!\!\perp Y.$$

But then Y must be Gaussian itself by Cramer's Theorem or simply by considering characteristic functions:

$$\Psi_Y(t) \cdot e^{ct^2} = e^{c't^2} \Rightarrow \Psi_Y(t) = e^{c''t^2} \implies \text{Y- Gaussian}$$

Therefore, (X_g, Y) must be jointly Gaussian and hence $N = Y - X_g$ is Gaussian. Thus we conclude that it is only possible to attain $I(X_g; X_g + N) = C$ if N is Gaussian of variance σ_N^2 and independent of X_g. □

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 $X_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp X_n$ 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 $\{0,1\} = \mathbb{F}_2$:

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

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 $Z_G^n \sim U Z_G^n$, 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 $P = Q$.

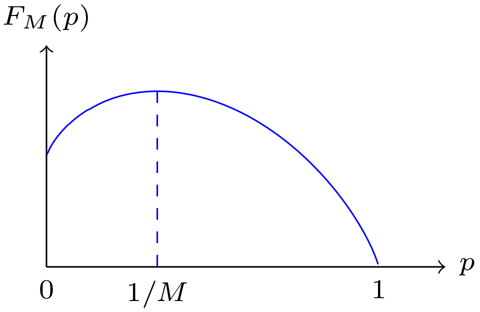
Second proof: Given any $P = (P_{\max}, P_2, \ldots, P_M)$, apply a random permutation π to the last $M-1$ atoms to obtain the distribution P_π. Then averaging P_π over all permutation π gives Q. Then use concavity of entropy or "conditioning reduces entropy": $H(Q) \ge H(P_\pi|\pi) = H(P)$.

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*

$$\begin{aligned} P_e \triangleq \mathbb{P}[W \ne \hat{W}] &\ge 1 - \frac{I(X;Y) + h(P_e)}{\log M} \qquad (5.8) \\ &\ge 1 - \frac{I(X;Y) + \log 2}{\log M}. \qquad (5.9) \end{aligned}$$

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:

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

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

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

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

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

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

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

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

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

Together with (5.10), we have

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

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

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

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.

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

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

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

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*

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

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

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

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. □

Part II

Lossless data compression

§ 6. Variable-length Lossless Compression

The principal engineering goal of compression is to represent a given sequence $a_1, a_2, \dots, a_n$ produced by a source as a sequence of bits of minimal possible length. Of course, reducing the number of bits is generally impossible, unless the source imposes certain restrictions. That is if only a small subset of all sequences actually occur in practice. Is it so for real world sources?

As a simple demonstration, one may take two English novels and compute empirical frequencies of each letter. It will turn out to be the same for both novels (approximately). Thus, we can see that there is some underlying structure in English texts restricting possible output sequences. The structure goes beyond empirical frequencies of course, as further experimentation (involving digrams, word frequencies etc) may reveal. Thus, the main reason for the possibility of data compression is the *experimental (empirical) law: real-world sources produce very restricted sets of sequences.*

How do we model these restrictions? Further experimentation (with language, music, images) reveals that frequently, the structure may be well described if we assume that sequences are generated probabilistically. This is one of the main contributions of Shannon: *another empirical law states that real-world sources may be described probabilistically with increasing precision starting from i.i.d., 1-st order Markov, 2-nd order Markov etc.* Note that sometimes one needs to find an appropriate basis in which this "law" holds – this is the case of images (i.e. rasterized sequence of pixels won't appear to have local probabilistic laws, because of forgetting the 2-D constraints; wavelets and local Fourier transform provide much better bases).[1]

So our initial investigation will be about representing one random variable $X \sim P_X$ in terms of bits efficiently. Types of compression:

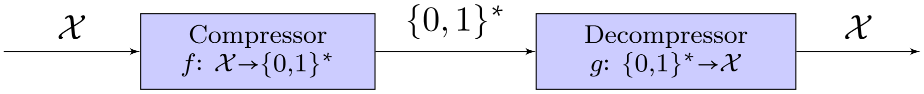
- Lossy
  $X \to W \to \hat{X}$ s.t. $\mathbb{E}[(X - \hat{X})^2] \le$ distortion.

- Lossless
  $P(X \ne \hat{X}) = 0$. variable-length code, uniquely decodable codes, prefix codes, Huffman codes

- Almost lossless
  $P(X \ne \hat{X}) \le \epsilon$. fixed-length codes

6.1 Variable-length, lossless, optimal compressor

Coding paradigm:


$\mathcal{X}$
Compressor
f: $\mathcal{X} \to \{0,1\}^*$
$\{0,1\}^*$
Decompressor
g: $\{0,1\}^* \to \mathcal{X}$
$\mathcal{X}$

[1] Of course, one should not take these "laws" too far. In regards to language modeling, (finite-state) Markov assumption is too simplistic to truly generate all proper sentences, cf. Chomsky [Cho56].

Remark 6.1.

- Codeword: $f(x) \in \{0,1\}^*$; Codebook: $\{f(x) : x \in \mathcal{X}\} \subset \{0,1\}^*$
- Since $\{0,1\}^* = \{\varnothing, 0, 1, 00, 01, \dots\}$ is countable, lossless compression is only possible for discrete R.V.;
- if we want $g \circ f = 1_{\mathcal{X}}$ (lossless), then f must be injective;
- relabel $\mathcal{X}$ such that $\mathcal{X} = \mathbb{N} = \{1, 2, \dots\}$ and order the pmf decreasingly: $P_X(i) \geq P_X(i+1)$.

Length function:

$$l : \{0,1\}^* \to \mathbb{N}$$

e.g., $l(01001) = 5$.

Objectives: Find the best compressor f to minimize

$$\mathbb{E}[l(f(X))]$$

$$\sup l(f(X))$$

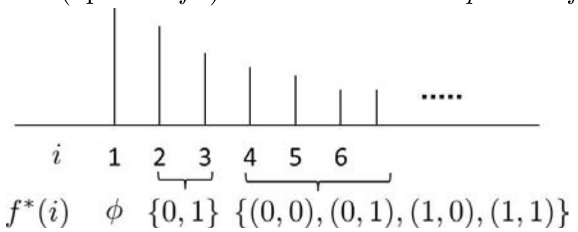
$$\text{median } l(f(X))$$

It turns out that there is a compressor f^* that minimizes all together!

Main idea: Assign longer codewords to less likely symbols, and reserve the shorter codewords for more probable symbols.

Aside: It is useful to introduce the partial order of *stochastic dominance*: For real-valued RV X and Y, we say Y stochastically dominates (or, is stochastically larger than) X, denoted by $X \overset{\text{st.}}{\leq} Y$, if $\mathbb{P}[Y \leq t] \leq \mathbb{P}[X \leq t]$ for all $t \in \mathbb{R}$. In other words, $X \overset{\text{st.}}{\leq} Y$ iff the CDF of X is larger than the CDF of Y pointwise. In particular, if X is dominated by Y stochastically, so are their means, medians, supremum, etc.

Theorem 6.1 (optimal f^*). *Consider the compressor f^* defined by*



Then

1. *length of codeword:*

$$l(f^*(i)) = \lfloor \log_2 i \rfloor$$

2. *$l(f^*(X))$ is stochastically the smallest: for any lossless f,*

$$l(f^*(X)) \overset{\text{st.}}{\leq} l(f(X))$$

i.e., for any k, $\mathbb{P}[l(f(X)) \leq k] \leq \mathbb{P}[l(f^*(X)) \leq k]$.

Proof. Note that

$$|A_k| \triangleq |\{x : l(f(x)) \le k| \le \sum_{i=0}^{k} 2^i = 2^{k+1} - 1 = |\{x : l(f^*(x)) \le k| \triangleq |A_k^*|.$$

where the inequality is because of f is lossless and $|A_k|$ exceeds the total number of binary strings of length less than k. Then

$$\mathbb{P}[l(f(X)) \le k] = \sum_{x \in A_k} P_X(x) \le \sum_{x \in A_k^*} P_X(x) = \mathbb{P}[l(f^*(X)) \le k],$$

since $|A_k| \le |A_k^*|$ and A_k^* contains all $2^{k+1} - 1$ most likely symbols. □

The following lemma is useful in bounding the expected code length of f^*. It says if the random variable is integer-valued, then its entropy can be controlled using its mean.

Lemma 6.1. *For any $Z \in \mathbb{N}$ s.t. $\mathbb{E}[Z] < \infty$, $H(Z) \le \mathbb{E}[Z] h(\frac{1}{\mathbb{E}[Z]})$, where $h(\cdot)$ is the binary entropy function.*

Theorem 6.2 (Optimal average code length: exact expression)**.** *Suppose $X \in \mathbb{N}$ and $P_X(1) \ge P_X(2) \ldots$. Then*

$$\mathbb{E}[l(f^*(X))] = \sum_{k=1}^{\infty} \mathbb{P}[X \ge 2^k].$$

Proof. Recall that expectation of $U \in \mathbb{Z}_+$ can be written as $\mathbb{E}[U] = \sum_{k \ge 1} \mathbb{P}[U \ge k]$. Then by Theorem 6.1, $\mathbb{E}[l(f^*(X))] = \mathbb{E}[\lfloor \log_2 X \rfloor] = \sum_{k \ge 1} \mathbb{P}[\lfloor \log_2 X \rfloor \ge k] = \sum_{k \ge 1} \mathbb{P}[\log_2 X \ge k]$. □

Theorem 6.3 (Optimal average code length v.s. entropy)**.**

$$H(X) \texttt{ bits} - \log_2[e(H(X) + 1)] \le \mathbb{E}[l(f^*(X))] \le H(X) \texttt{ bits}$$

Note: Theorem 6.3 is the first example of a coding theorem, which relates the fundamental limit $\mathbb{E}[l(f^*(X))]$ (operational quantity) to the entropy $H(X)$ (information measure).

Proof. Define $L(X) = l(f^*(X)))$.

RHS: observe that since the pmf are ordered decreasingly by assumption, $P_X(m) \le 1/m$, so $L(m) \le \log_2 m \le \log_2(1/P_X(m))$, take exp., $\mathbb{E}[L(X)] \le H(X)$.

LHS:

$$\begin{aligned}
H(X) = H(X, L) &= H(X|L) + H(L) \\
&\le \mathbb{E}[L] + h\left(\frac{1}{1 + \mathbb{E}[L]}\right)(1 + \mathbb{E}[L]) && \text{(Lemma 6.1)} \\
&= \mathbb{E}[L] + \log_2(1 + \mathbb{E}[L]) + \mathbb{E}[L] \log\left(1 + \frac{1}{\mathbb{E}[L]}\right) \\
&\le \mathbb{E}[L] + \log_2(1 + \mathbb{E}[L]) + \log_2 e && (x \log(1 + 1/x) \le \log e, \forall x > 0) \\
&\le \mathbb{E}[L] + \log(e(1 + H(X))) && \text{(by RHS)}
\end{aligned}$$

where we have used $H(X|L = k) \le k \texttt{ bits}$, since given $l(f^*(X))) = k$, X has at most 2^k choices. □

Note: (Memoryless source) If $X = S^n$ is an i.i.d. sequence, then

$$nH(S) \geq \mathbb{E}[l(f^*(S^n))] \geq nH(S) - \log n + O(1).$$

For iid sources, the exact behavior is found in [SV11, Theorem 4] as:

$$\mathbb{E}[\ell(f^*(S^n))] = nH(S) - \frac{1}{2}\log n + O(1)\,,$$

unless the source is uniform (in which case it is $nH(S) + O(1)$.

Theorem 6.3 relates the *mean* of $l(f^*(X)) \leq k$ to that of $\log_2 \frac{1}{P_X(X)}$ (entropy). The next result relates their *CDFs*.

Theorem 6.4 (Code length distribution of f^*)**.** $\forall \tau > 0, k \in \mathbb{Z}_+$,

$$\mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \leq k\right] \leq \mathbb{P}\left[l(f^*(X)) \leq k\right] \leq \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \leq k + \tau\right] + 2^{-\tau+1}$$

Proof. LHS: easy, use $P_X(m) \leq 1/m$. Then similarly as in Theorem 6.3, $L(m) = \lfloor \log_2 m \rfloor \leq \log_2 m \leq \log_2 \frac{1}{P_X(m)}$. Hence $L(X) \leq \log_2 \frac{1}{P_X(X)}$ a.s.
RHS: (truncation)

$$\begin{aligned}\mathbb{P}\left[L \leq k\right] &= \mathbb{P}\left[L \leq k, \log_2 \frac{1}{P_X(X)} \leq k + \tau\right] + \mathbb{P}\left[L \leq k, \log_2 \frac{1}{P_X(X)} > k + \tau\right]\\ &\leq \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \leq k + \tau\right] + \sum_{x \in \mathcal{X}} P_X(x) \mathbf{1}_{\{l(f^*(x)) \leq k\}} \mathbf{1}_{\{P_X(x) \leq 2^{-k-\tau}\}}\\ &\leq \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \leq k + \tau\right] + (2^{k+1} - 1) \cdot 2^{-k-\tau} \end{aligned}$$

□

So far our discussion applies to an arbitrary random variable X. Next we consider the source as a random process $(S_1, S_2, \ldots)$ and introduce blocklength. We apply our results to $X = S^n$: the first n symbols. The following corollary states that the limiting behavior of $l(f^*(S^n))$ and $\log \frac{1}{P_{S^n}(S^n)}$ always coincide.

Corollary 6.1. *Let $(S_1, S_2, \ldots)$ be some random process and U be some random variable. Then*

$$\frac{1}{n}\log_2 \frac{1}{P_{S^n}(S^n)} \xrightarrow{\text{D}} U \quad \Leftrightarrow \quad \frac{1}{n} l(f^*(S^n)) \xrightarrow{\text{D}} U \tag{6.1}$$

and

$$\frac{1}{\sqrt{n}}\left(\log_2 \frac{1}{P_{S^n}(S^n)} - H(S^n)\right) \xrightarrow{\text{D}} V \quad \Leftrightarrow \quad \frac{1}{\sqrt{n}}(l(f^*(S^n)) - H(S^n)) \xrightarrow{\text{D}} V \tag{6.2}$$

Proof. The proof is simply logic. First recall: convergence in distribution is equivalent to convergence of CDF at any continuity point. $U_n \xrightarrow{\text{D}} U \Leftrightarrow \mathbb{P}[U_n \leq u] \to \mathbb{P}[U \leq u]$ for all u at which point the CDF of U is continuous (i.e., not an atom of U).

Apply Theorem 6.4 with $k = un$ and $\tau = \sqrt{n}$:

$$\mathbb{P}\left[\frac{1}{n}\log_2 \frac{1}{P_X(X)} \leq u\right] \leq \mathbb{P}\left[\frac{1}{n} l(f^*(X)) \leq u\right] \leq \mathbb{P}\left[\frac{1}{n}\log_2 \frac{1}{P_X(X)} \leq u + \frac{1}{\sqrt{n}}\right] + 2^{-\sqrt{n}+1}.$$

Apply Theorem 6.4 with $k = H(S^n) + \sqrt{n}u$ and $\tau = n^{1/4}$:

$$\mathbb{P}\left[\frac{1}{\sqrt{n}}\left(\log\frac{1}{P_{S^n}(S^n)} - H(S^n)\right) \le u\right] \le \mathbb{P}\left[\frac{l(f^*(S^n)) - H(S^n)}{\sqrt{n}} \le u\right]$$
$$\le \mathbb{P}\left[\frac{1}{\sqrt{n}}\left(\log\frac{1}{P_{S^n}(S^n)} - H(S^n)\right) \le u + n^{-1/4}\right] + 2^{-n^{1/4}+1} \quad \square$$

Remark 6.2 (Memoryless source)**.** Now let us consider S^n that are i.i.d. Then $\log\frac{1}{P_{S^n}(S^n)} = \sum_{i=1}^n \log\frac{1}{P_S(S_i)}$.

1. By the Law of Large Numbers (LLN), we know that $\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} \xrightarrow{\mathbb{P}} \mathbb{E}\log\frac{1}{P_S(S)} = H(S)$. Therefore in (6.1) the limiting distribution U is degenerate, i.e., $U = H(S)$, and we have $\frac{1}{n}l(f^*(S^n)) \xrightarrow{\mathbb{P}} \mathbb{E}\log\frac{1}{P_S(S)} = H(S)$. [Note: convergence in distribution to a constant $\Leftrightarrow$ convergence in probability to a constant]

2. By the Central Limit Theorem (CLT), if $V(S) \triangleq \mathrm{Var}\left[\log\frac{1}{P_S(S)}\right] < \infty$,[2] then we know that V in (6.2) is Gaussian, i.e.,

$$\frac{1}{\sqrt{nV(S)}}\left(\log\frac{1}{P_{S^n}(S^n)} - nH(S)\right) \xrightarrow{\mathrm{D}} \mathcal{N}(0,1).$$

   Consequently, we have the following Gaussian approximation for the probability law of the optimal code length

$$\frac{1}{\sqrt{nV(S)}}(l(f^*(S^n)) - nH(S)) \xrightarrow{\mathrm{D}} \mathcal{N}(0,1),$$

   or, in shorthand,

$$l(f^*(S^n)) \sim nH(S) + \sqrt{nV(S)}\mathcal{N}(0,1) \quad \text{in distribution.}$$

   Gaussian approximation tells us the speed of $\frac{1}{n}l(f^*(S^n))$ to entropy and give us a good approximation at finite n. In the next section we apply our bounds to approximate the distribution of $\ell(f^*(S^n))$ in a concrete example:

6.1.1 Compressing iid ternary source

Consider the source outputing n ternary letters each independent and distributed as

$$P_X = [.445 \quad .445 \quad .11].$$

For iid source it can be shown

$$\mathbb{E}[\ell(f^*(X^n))] = nH(X) - \frac{1}{2}\log(2\pi eVn) + O(1),$$

where we denoted the *varentropy* of X by

$$V(X) \triangleq \mathrm{Var}\left[\log\frac{1}{P_X(X)}\right].$$

[2] V is often known as the *varentropy* of S.

The Gaussian approximation to $\ell(f^*(X))$ is defined as

$$nH(X) - \frac{1}{2}\log 2\pi eVn + \sqrt{nV}Z\,,$$

where $Z \sim \mathcal{N}(0,1)$.

On Fig. 6.1, 6.2, 6.3 we plot the distribution of the length of the optimal compressor for different values of n and compare with the Gaussian approximation.

Upper/lower bounds on the expectation:

$$H(X^n) - \log(H(X^n)+1) - \log e \le \mathbb{E}[\ell(f^*(X^n))] \le H(X^n)$$

Here are the numbers for different n

$$\begin{array}{rrcrcr} n = 20 & 21.5 & \le & 24.3 & \le & 27.8 \\ n = 100 & 130.4 & \le & 134.4 & \le & 139.0 \\ n = 500 & 684.1 & \le & 689.2 & \le & 695.0 \end{array}$$

In all cases above $\mathbb{E}[\ell(f^*(X))]$ is close to a midpoint between the two.

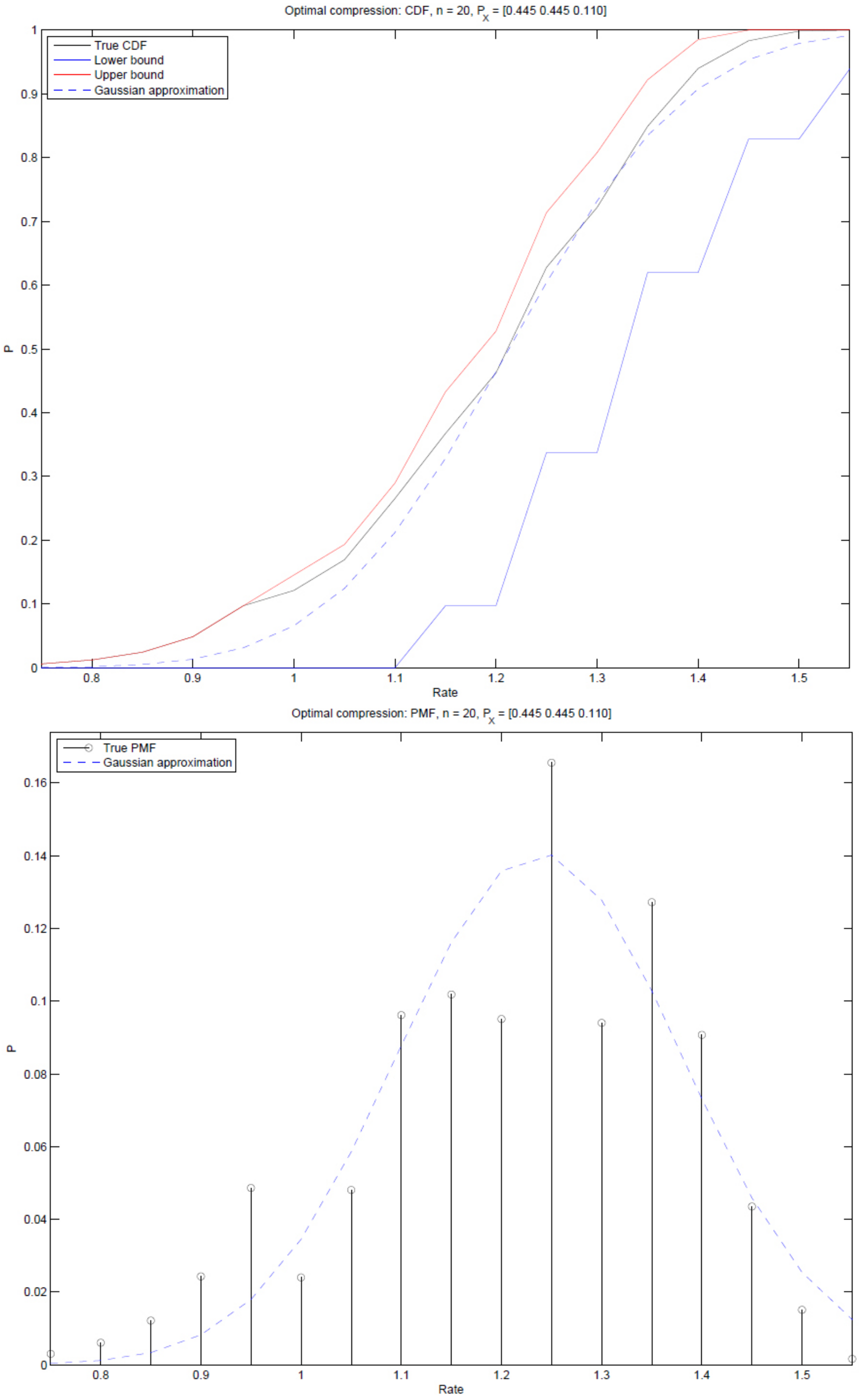


Figure 6.1: CDF and PMF of optimal compressor

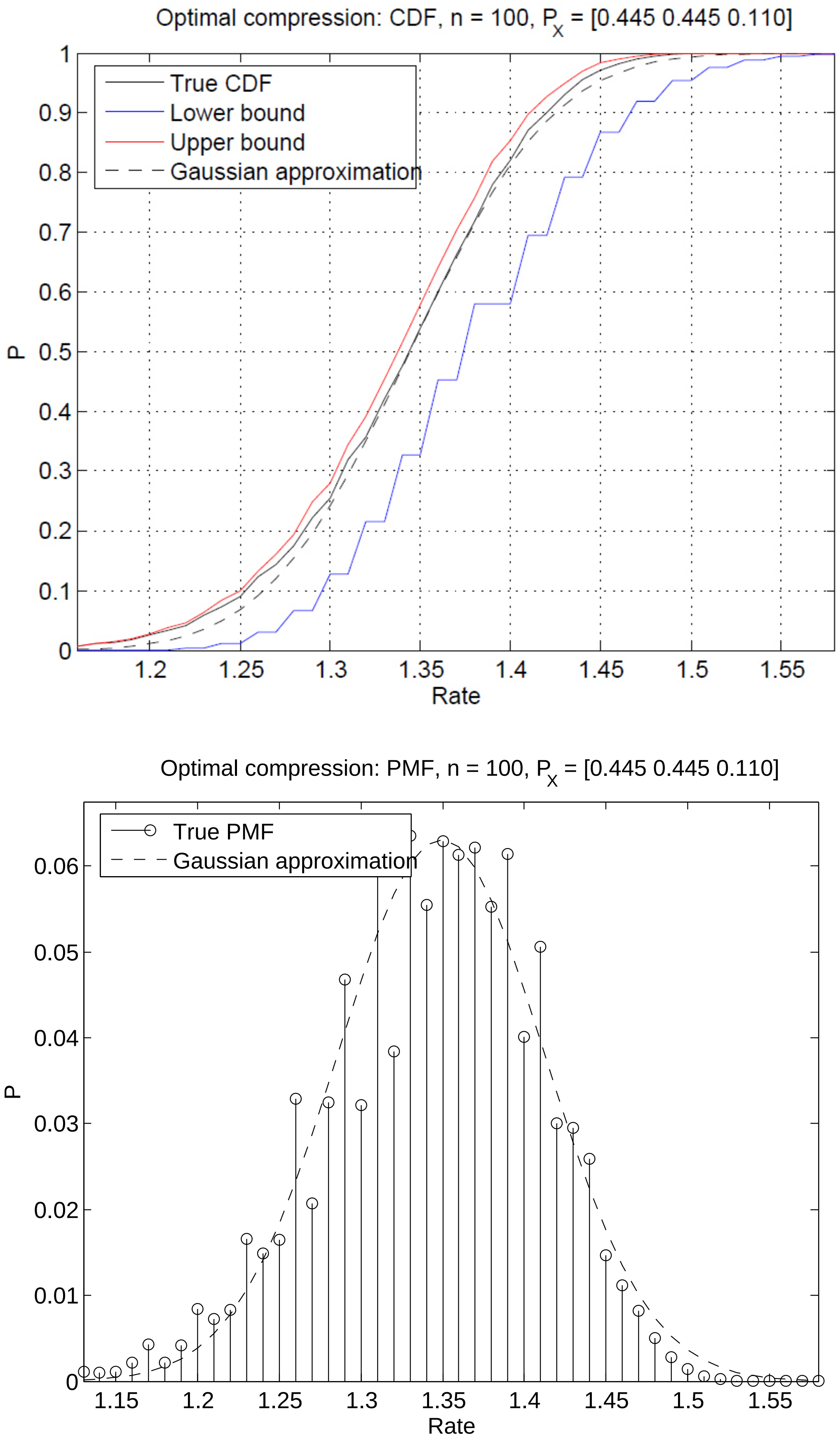


Figure 6.2: CDF and PMF, **Gaussian is shifted** to the true $\mathbb{E}[\ell(f^*(X))]$

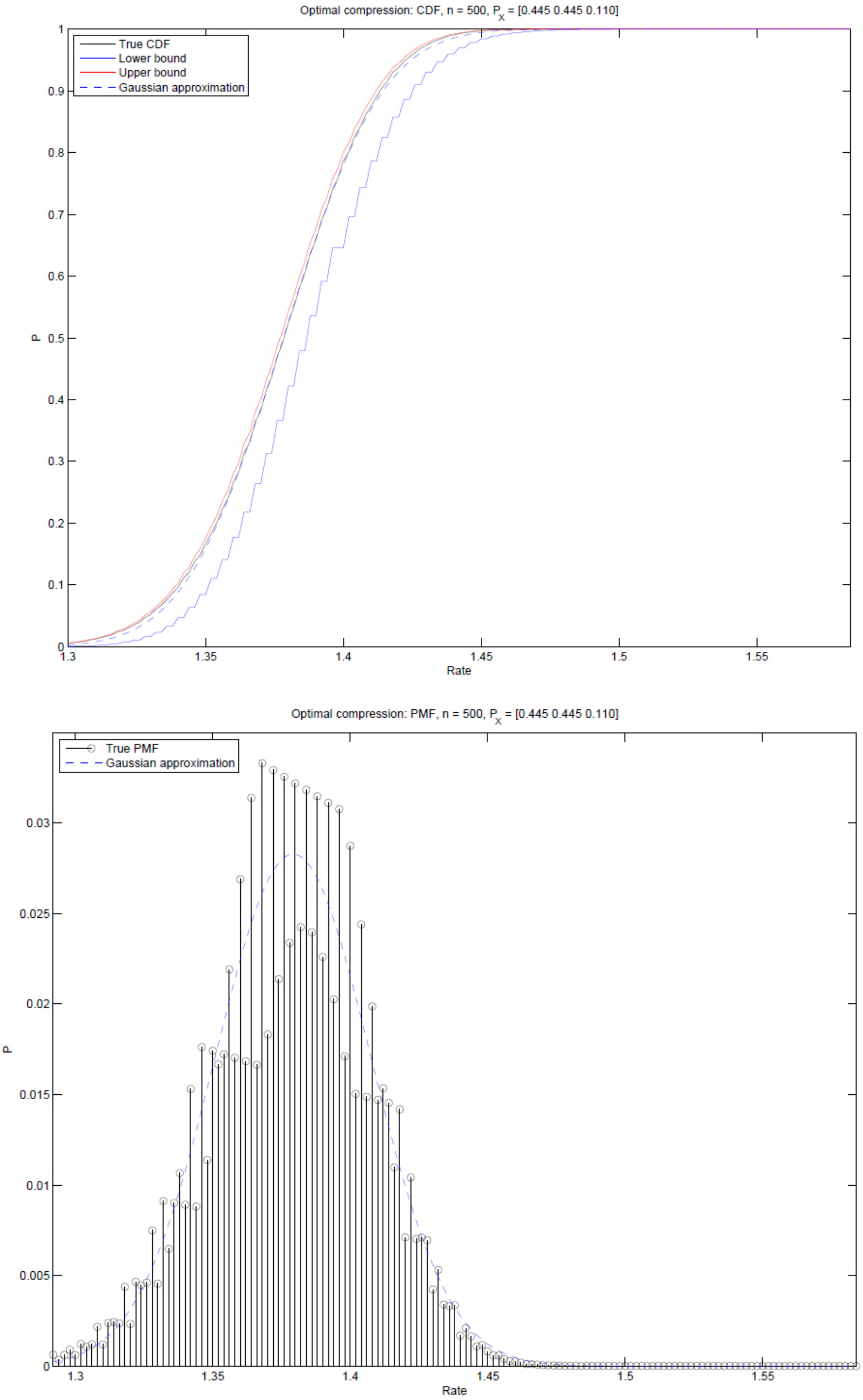
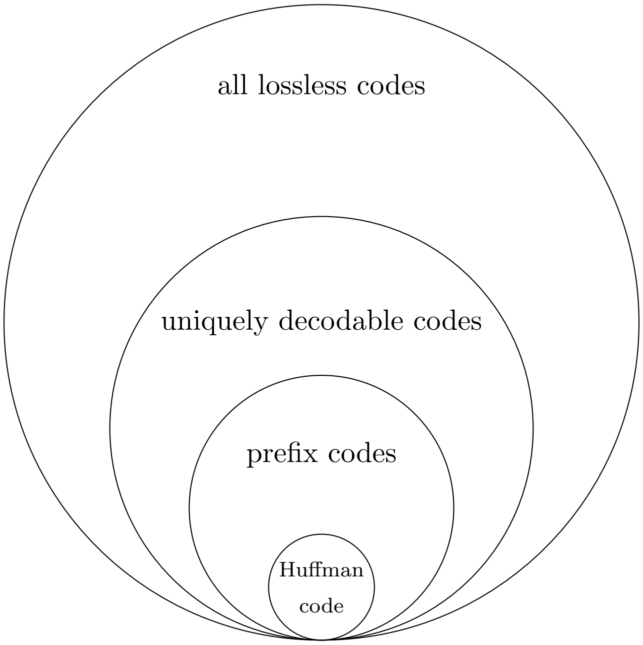


Figure 6.3: CDF and PMF of optimal compressor

6.2 Uniquely decodable codes, prefix codes and Huffman codes



We have studied f^*, which achieves the stochastically smallest code length among all variable-length compressors. Note that f^* is obtained by ordering the pmf and assigning shorter codewords to more likely symbols. In this section we focus on a specific class of compressors with good properties which lead to low complexity and short delay when decoding from a stream of compressed bits. This part is more combinatorial in nature.

We start with a few definition. Let $\mathcal{A}^+ = \bigcup_{n\geq 1} \mathcal{A}^n$ denotes all non-empty finite-length strings consisting of symbols from alphabet $\mathcal{A}$

Definition 6.1 (Extension of a code)**.** The extension of $f : \mathcal{A} \to \{0,1\}^*$ is $f : \mathcal{A}^+ \to \{0,1\}^*$ where $f(a_1,\ldots,a_n) = (f(a_1),\ldots,f(a_n))$ is defined by concatenating the bits.

Definition 6.2 (Uniquely decodable codes)**.** $f : \mathcal{A} \to \{0,1\}^*$ is *uniquely decodable* if its extension $f : \mathcal{A}^+ \to \{0,1\}^*$ is injective.

Definition 6.3 (Prefix codes)**.** $f : \mathcal{A} \to \{0,1\}^*$ is a *prefix code*[3] if no codeword is a prefix of another (e.g., 010 is a prefix of 0101).

Example:

- $f(a) = 0, f(b) = 1, f(c) = 10$ – not uniquely decodable, since $f(ba) = f(c) = 10$.
- $f(a) = 0, f(b) = 10, f(c) = 11$ – uniquely decodable and prefix.
- $f(a) = 0, f(b) = 01, f(c) = 011, f(d) = 0111$ – uniquely decodable but not prefix, since as long as 0 appears, we know that the last codeword has terminated.

Remark 6.3.

1. Prefix codes are uniquely decodable.

[3] Also known as prefix-free/comma-free/instantaneous code.

2. Similar to prefix-free codes, one can define suffix-free codes. Those are also uniquely decodable (one should start decoding in reverse direction).

3. By definition, any uniquely decodable code does not have the empty string as a codeword. Hence $f : \mathcal{X} \to \{0,1\}^+$ in both Definition 6.2 and Definition 6.3.

4. Unique decodability means that one can decode from a stream of bits without ambiguity, but one might need to look ahead in order to decide the termination of a codeword. (Think of the last example). In contrast, prefix codes allow the decoder to decode instantaneously without looking ahead.

5. Prefix code ↔ binary tree (codewords are leaves) ↔ strategy to ask "yes/no" questions

Theorem 6.5 (Kraft-McMillan)**.**

1. *Let $f : \mathcal{A} \to \{0,1\}^*$ be uniquely decodable. Set $l_a = l(f(a))$. Then f satisfies the* Kraft inequality

$$\sum_{a\in\mathcal{A}} 2^{-l_a} \le 1. \tag{6.3}$$

2. *Conversely, for any set of code length $\{l_a : a \in \mathcal{A}\}$ satisfying (6.3), there exists a prefix code f, such that $l_a = l(f(a))$.*

Note: The consequence of Theorem 6.5 is that as far as compression efficiency is concerned, we can forget about uniquely decodable codes that are not prefix codes.

Proof. We prove the Kraft inequality for prefix codes and uniquely decodable codes separately. The purpose for doing a separate proof for prefix codes is to illustrate the powerful technique of *probabilistic method*. The idea is from [AS08, Exercise 1.8, p. 12].

Let f be a prefix code. Let us construct a probability space such that the LHS of (6.3) is the probability of some event, which cannot exceed one. To this end, consider the following scenario: Generate independent Bern$(\frac{1}{2})$ bits. Stop if a codeword has been written, otherwise continue. This process terminates with probability $\sum_{a\in\mathcal{A}} 2^{-l_a}$. The summation makes sense because the events that a given codeword is written are mutually exclusive, thanks to the prefix condition.

Now let f be a uniquely decodable code. The proof uses *generating function* as a device for counting. (The analogy in coding theory is the weight enumerator function.) First assume $\mathcal{A}$ is finite. Then $L = \max_{a\in\mathcal{A}} l_a$ is finite. Let $G_f(z) = \sum_{a\in\mathcal{A}} z^{l_a} = \sum_{l=0}^{L} A_l(f) z^l$, where $A_l(f)$ denotes the number of codewords of length l in f. For $k \ge 1$, define $f^k : \mathcal{A}^k \to \{0,1\}^*$ as the symbol-by-symbol extension of f. Then $G_{f^k}(z) = \sum_{a^k\in\mathcal{A}^k} z^{l(f^k(a^k))} = \sum_{a_1} \cdots \sum_{a_k} z^{l_{a_1}+\cdots+l_{a_k}} = [G_f(z)]^k = \sum_{l=0}^{kL} A_l(f^k) z^l$. By unique decodability of f, f^k is lossless. Hence $A_l(f^k) \le 2^l$. Therefore we have $G_f(1/2)^k = G_{f^k}(1/2) \le kL$ for all k. Then $\sum_{a\in\mathcal{A}} 2^{-l_a} = G_f(1/2) \le \lim_{k\to\infty} (kL)^{1/k} \to 1$. If $\mathcal{A}$ is countably infinite, for any finite subset $\mathcal{A}' \subset \mathcal{A}$, repeating the same argument gives $\sum_{a\in\mathcal{A}'} 2^{-l_a} \le 1$. The proof is complete by the arbitrariness of $\mathcal{A}'$.

Conversely, given a set of code lengths $\{l_a : a \in \mathcal{A}\}$ s.t. $\sum_{a\in\mathcal{A}} 2^{-l_a} \le 1$, construct a prefix code f as follows: First relabel $\mathcal{A}$ to $\mathbb{N}$ and assume that $l_1 \le l_2 \le \ldots$. For each i, $a_i \triangleq \sum_{k=1}^{i-1} 2^{-l_k} < 1$ by Kraft inequality. Thus we define the codeword $f(i) \in \{0,1\}^*$ as the first l_i bits in the binary expansion of a_i. Prove that f is a prefix code by contradiction: Suppose for some $j > i$, $f(i)$ is the prefix of $f(j)$, since $l_j \ge l_i$. Then $a_j - a_i \le 2^{-l_i}$. But $a_j - a_i = 2^{-l_i} + 2^{-l_{i+1}} + \ldots > 2^{-l_i}$, which is a contradiction. □

Open problems:

1. Find a probabilistic proof of Kraft inequality for uniquely decodable codes.
2. There is a conjecture of Ahslwede that for any sets of lengths for which $\sum 2^{-l_a} \le \frac{3}{4}$ there exists a fix-free code (i.e. one which is simultaneously prefix-free and suffix-free). So far, existence has only been shown when the Kraft sum is $\le \frac{5}{8}$, cf. [Yek04].

In view of Theorem 6.5, the optimal average code length among all prefix (or uniquely decodable) codes is given by the following optimization problem

$$\begin{aligned} L^*(X) \triangleq \min & \sum_{a\in\mathcal{A}} P_X(a) l_a \\ \text{s.t.} & \sum_{a\in\mathcal{A}} 2^{-l_a} \le 1 \\ & l_a \in \mathbb{N} \end{aligned} \tag{6.4}$$

This is an *integer programming* (IP) problem, which in general is hard to solve computationally. It is remarkable that this particular IP problem can be solved in *near-linear* time, thanks to the Huffman algorithm. Before describing the construction of Huffman codes, let us give bounds to $L^*(X)$ in terms of entropy:

Theorem 6.6.

$$H(X) \le L^*(X) \le H(X) + 1\,\mathtt{bit}. \tag{6.5}$$

Proof. "$\le$" Consider the following length assignment $l_a = \left\lceil \log_2 \frac{1}{P_X(a)} \right\rceil$,[4] which satisfies Kraft since $\sum_{a\in\mathcal{A}} 2^{-l_a} \le \sum_{a\in\mathcal{A}} P_X(a) = 1$. By Theorem 6.5, there exists a prefix code f such that $l(f(a)) = \left\lceil \log_2 \frac{1}{P_X(a)} \right\rceil$ and $\mathbb{E} l(f(X)) \le H(X) + 1$.

"$\ge$" We give two proofs for the converse. One of the commonly used ideas to deal with combinatorial optimization is *relaxation*. Our first idea is to drop the integer constraints in (6.4) and *relax* it into the following optimization problem, which obviously provides a lower bound

$$L^*(X) \triangleq \min \sum_{a\in\mathcal{A}} P_X(a) l_a \tag{6.6}$$

$$\text{s.t.} \sum_{a\in\mathcal{A}} 2^{-l_a} \le 1 \tag{6.7}$$

This is a nice *convex programming* problem, since the objective function is affine and the feasible set is convex. Solving (6.6) by Lagrange multipliers (Exercise!) yields the minimum is equal to $H(X)$ (achieved at $l_a = \log_2 \frac{1}{P_X(a)}$).

Another proof is the following: For any f satisfying Kraft inequality, define a probability measure $Q(a) = \frac{2^{-l_a}}{\sum_{a\in\mathcal{A}} 2^{-l_a}}$. Then

$$\begin{aligned} \mathbb{E} l(f(X)) - H(X) &= D(P\|Q) - \log \sum_{a\in\mathcal{A}} 2^{-l_a} \\ &\ge 0 \end{aligned}$$

□

Next we describe the Huffman code, which achieves the optimum in (6.4). In view of the fact that prefix codes and binary trees are one-to-one, the main idea of Huffman code is to build the binary tree bottom-up: Given a pmf $\{P_X(a) : a \in \mathcal{A}\}$,

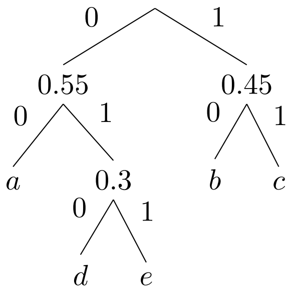
[4] Such a code is called a Shannon code.

1. Choose the two least-probable symbols in the alphabet
2. Delete the two symbols and add a new symbol (with combined weights). Add the new symbol as the parent node of the previous two symbols in the binary tree.

The algorithm terminates in $|\mathcal{A}| - 1$ steps. Given the binary tree, the code assignment can be obtained by assigning 0/1 to the branches. Therefore the time complexity is $O(|\mathcal{A}|)$ (sorted pmf) or $O(|\mathcal{A}| \log |\mathcal{A}|)$ (unsorted pmf).

Example: $\mathcal{A} = \{a, b, c, d, e\}$, $P_X = \{0.25, 0.25, 0.2, 0.15, 0.15\}$.

Huffman tree: codebook:



$$f(a) = 00$$
$$f(b) = 10$$
$$f(c) = 11$$
$$f(d) = 010$$
$$f(e) = 011$$

Theorem 6.7 (Optimality of Huffman codes)**.** *The Huffman code achieves the minimal average code length (6.4) among all prefix (or uniquely decodable) codes.*

Proof. [CT06, Sec. 5.8]. □

Remark 6.4 (Drawbacks of Huffman codes)**.**

1. Does not exploit memory. Solution: block Huffman coding. Shannon's original idea from 1948 paper: in compressing English text, instead of dealing with letters and exploiting the nonequiprobability of the English alphabet, working with pairs of letters to achieve more compression (more generally, n-grams). Indeed, compressing the block $(S_1, \dots, S_n)$ using its Huffman code achieves $H(S_1, \dots, S_n)$ within one bit, but the complexity is $|\mathcal{A}|^n$!
2. Non-universal (constructing the Huffman code needs to know the source distribution). This brings us the question: Is it possible to design universal compressor which achieves entropy for a class of source distributions? And what is the price to pay? – Homework!

There are much more elegant solutions, e.g.,

1. Arithmetic coding: sequential encoding, linear complexity in compressing $(S_1, \dots, S_n)$ (see later).
2. Lempel-Ziv algorithm: low-complexity, universal, provably optimal in a very strong sense.

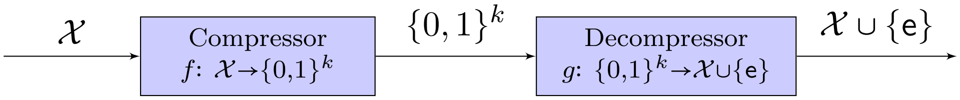
To sum up: Comparison of average code length (in bits):

$$H(X) - \log_2[e(H(X)+1)] \le \mathbb{E}[l(f^*(X))] \le H(X) \le \mathbb{E}[l(f_{\text{Huffman}}(X))] \le H(X) + 1.$$

§ 7. Fixed-length (almost lossless) compression. Slepian-Wolf problem.

7.1 Fixed-length code, almost lossless

Coding paradigm:


$\mathcal{X}$
Compressor
f: $\mathcal{X}\to\{0,1\}^k$
$\{0,1\}^k$
Decompressor
g: $\{0,1\}^k\to\mathcal{X}\cup\{\mathsf{e}\}$
$\mathcal{X}\cup\{\mathsf{e}\}$

Note: If we want $g \circ f = \mathbf{1}_{\mathcal{X}}$, then $k \ge \log_2 |\mathcal{X}|$. But, the transmission link is erroneous anyway... and it turns out that by tolerating a little error probability ϵ, we gain a lot in terms of code length!

Indeed, the key idea is to **allow errors**: Instead of insisting on $g(f(x)) = x$ for all $x \in \mathcal{X}$, consider only lossless decompression for a subset $\mathcal{S} \subset \mathcal{X}$:

$$g(f(x)) = \begin{cases} x & x \in \mathcal{S} \\ \mathsf{e} & x \notin \mathcal{S} \end{cases}$$

and the probability of error: $\mathbb{P}[g(f(X)) \ne X] = \mathbb{P}[g(f(X)) = \mathsf{e}]$.

Definition 7.1. A compressor-decompressor pair (f, g) is called a (k, ϵ)-code if:

$$f: \ \mathcal{X} \to \{0,1\}^k$$
$$g: \ \{0,1\}^k \to \mathcal{X} \cup \{\mathsf{e}\}$$

such that $g(f(x)) \in \{x, \mathsf{e}\}$ and $\mathbb{P}[g(f(X)) = \mathsf{e}] \le \epsilon$.

Fundamental limit:

$$\epsilon^*(X, k) \triangleq \inf\{\epsilon : \exists (k,\epsilon)\text{-code for } X\}$$

The following result connects the respective fundamental limits of fixed-length almost lossless compression and variable-length lossless compression (Lecture 6):

Theorem 7.1 (Fundamental limit of error probabiliy)**.**

$$\epsilon^*(X, k) = \mathbb{P}[l(f^*(X)) \ge k] = 1 - \text{sum of } 2^k - 1 \text{ largest masses of } X.$$

Proof. The proof is essentially tautological. Note $1 + 2 + \cdots + 2^{k-1} = 2^k - 1$. Let $\mathcal{S} = \{2^k - 1 \text{ most likely realizations of } X\}$. Then

$$\epsilon^*(X, k) = \mathbb{P}[X \notin \mathcal{S}] = \mathbb{P}[l(f^*(X)) \ge k].$$

Optimal codes:

- Variable-length: f^* encodes the $2^k - 1$ symbols with the highest probabilities to $\{\phi, 0, 1, 00, \ldots, 1^{k-1}\}$.
- Fixed-length: The optimal compressor f maps the elements of $\mathcal{S}$ into $(00\ldots00), \ldots, (11\ldots10)$ and the rest to $(11\ldots11)$. The decompressor g decodes perfectly except for outputting e upon receipt of $(11\ldots11)$. □

Note: In Definition 7.1 we require that the errors are always *detectable*, i.e., $g(f(x)) = x$ or e. Alternatively, we can drop this requirement and allow *undetectable* errors, in which case we can of course do better since we have more freedom in designing codes. It turns out that we do not gain much by this relaxation. Indeed, if we define

$$\tilde{\epsilon}^*(X,k) = \inf\{\mathbb{P}[g(f(X)) \neq X] : f : \mathcal{X} \to \{0,1\}^k, g : \{0,1\}^k \to \mathcal{X} \cup \{\mathsf{e}\}\},$$

then $\tilde{\epsilon}^*(X,k) = 1 -$sum of 2^k largest masses of X. This follows immediately from $\mathbb{P}[g(f(X)) = X] = \sum_{x\in C} P_X(x)$ where $C \triangleq \{x : g(f(x)) = x\}$ satisfies $|C| \le 2^k$, because f takes no more than 2^k values. Compared to Theorem 7.1, we see that $\tilde{\epsilon}^*(X,k)$ and $\tilde{\epsilon}^*(X,k)$ do not differ much. In particular, $\epsilon^*(X,k+1) \le \tilde{\epsilon}^*(X,k) \le \epsilon^*(X,k)$.

Corollary 7.1 (Shannon). *Let S^n be i.i.d. Then*

$$\lim_{n\to\infty} \epsilon^*(S^n, nR) = \begin{cases} 0 & R > H(S) \\ 1 & R < H(S) \end{cases}$$

$$\lim_{n\to\infty} \epsilon^*(S^n, nH(S) + \sqrt{nV(S)}\gamma) = 1 - \Phi(\gamma).$$

where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$, $H(S) = \mathbb{E}\log\frac{1}{P_S(S)}$ – entropy, $V(S) = \operatorname{Var}\log\frac{1}{P_S(S)}$ – varentropy is assumed to be finite.

Proof. Combine Theorem 7.1 with Theorem 6.1. □

Theorem 7.2 (Converse)**.**

$$\epsilon^*(X,k) \ge \tilde{\epsilon}^*(X,k) \ge \mathbb{P}\left[\log_2\frac{1}{P_X(X)} > k + \tau\right] - 2^{-\tau}, \quad \forall\tau > 0.$$

Proof. Identical to the converse of Theorem 6.4. Let $C = \{x : g(f(x)) = x\}$. Then $|C| \le 2^k$ and $\mathbb{P}[X \in C] \le \mathbb{P}\left[\log_2\frac{1}{P_X(X)} \le k + \tau\right] + \underbrace{\mathbb{P}\left[X \in C, \log_2\frac{1}{P_X(X)} > k + \tau\right]}_{\le 2^{-\tau}}$ □

Two achievability bounds

Theorem 7.3.

$$\epsilon^*(X,k) \le \mathbb{P}\left[\log_2\frac{1}{P_X(X)} \ge k\right] \tag{7.1}$$

and there exists a compressor-decompressor pair that achieves the upper bound.

Proof. Construction: use those $2^k - 1$ symbols with the highest probabilities.

This is essentially the same as the lower bound in Theorem 6.3 from Lecture 6. Note that the m^{th} largest mass $P_X(m) \le \frac{1}{m}$. Therefore

$$\epsilon^*(X,k) = \sum_{m\ge 2^k} P_X(m) = \sum \mathbf{1}_{\{m\ge 2^k\}} P_X(m) \le \sum \mathbf{1}_{\left\{\frac{1}{P_X(m)} \ge 2^k\right\}} P_X(m) = \mathbb{E}\mathbf{1}_{\left\{\log_2\frac{1}{P_X(X)} \ge k\right\}}.$$

□

Theorem 7.4.

$$\epsilon^*(X,k) \le \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} > k-\tau\right] + 2^{-\tau}, \quad \forall \tau > 0 \tag{7.2}$$

and there exists a compressor-decompressor pair that achieves the upper bound.

Note: In fact, Theorem 7.3 is always stronger than Theorem 7.4. Still, we present the proof of Theorem 7.4 and the technology behind it – *random coding* – a powerful technique for proving existence (achievability) which we heavily rely on in this course. To see that Theorem 7.3 gives a better bound, note that even the first term in (7.2) exceeds (7.1). Nevertheless, the method of proof for this weaker bound will be useful for generalizations.

Proof. Construction: **random coding** (Shannon's magic). For a given compressor f, the optimal decompressor which minimizes the error probability is the maximum a posteriori (MAP) decoder, i.e.,

$$g^*(w) = \operatorname*{argmax}_x P_{X|f(X)}(x|w) = \operatorname*{argmax}_{x:f(x)=w} P_X(x),$$

which can be hard to analyze. Instead, let us consider the following (suboptimal) decompressor g:

$$g(w) = \begin{cases} x, & \exists!\ x \in \mathcal{X} \text{ s.t. } f(x) = w \text{ and } \log_2 \frac{1}{P_X(x)} \le k - \tau, \\ & \text{(exists unique high-probability } x \text{ that is mapped to } w) \\ \mathsf{e}, & \text{o.w.} \end{cases}$$

Denote $f(x) = c_x$ and the codebook $\mathcal{C} = \{c_x : x \in \mathcal{X}\} \subset \{0,1\}^k$. It is instructive to think of $\mathcal{C}$ as a hashing table.

Error probability analysis: There are two ways to make an error ⇒ apply union bound. Before proceeding, define

$$J(x, \mathcal{C}) \triangleq \left\{x' \in \mathcal{X} : c_{x'} = c_x, x' \ne x, \log_2 \frac{1}{P_X(x')} < k - \tau\right\}$$

to be the set of high-probability inputs whose hashes collide with that of x. Then we have the following estimate for probability of error:

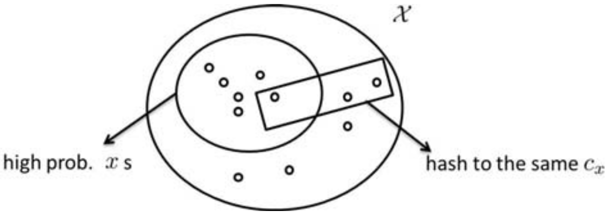
$$\begin{aligned}\mathbb{P}[g(f(X)) = \mathsf{e}] &= \mathbb{P}\left[\left\{\log_2 \frac{1}{P_X(X)} \ge k - \tau\right\} \cup \{J(X,\mathcal{C}) \ne \varnothing\}\right] \\ &\le \mathbb{P}\left[\log_2 \frac{1}{P_X(X)} \ge k - \tau\right] + \mathbb{P}[J(X,\mathcal{C}) \ne \phi]\end{aligned}$$

The first term does not depend on the codebook $\mathcal{C}$, while the second term does. The idea now is to randomize over $\mathcal{C}$ and show that when we average over all possible choices of codebook, the second term is smaller than $2^{-\tau}$. Therefore there exists at least one codebook that achieves the desired bound. Specifically, let us consider $\mathcal{C}$ which is uniformly distributed over all codebooks and independently of X. Equivalently, since $\mathcal{C}$ can be represented by a $|\mathcal{X}| \times k$ binary matrix, whose rows correspond to codewords, we choose each entry to be independent fair coin flips.

Averaging the error probability (over $\mathcal{C}$ and over X), we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{C}}[\mathbb{P}[J(X,\mathcal{C}) \neq \phi]] &= \mathbb{E}_{\mathcal{C},X}\left[\mathbf{1}_{\left\{\exists x' \neq X : \log_2 \frac{1}{P_X(x')} < k-\tau, c_{x'} = c_X\right\}}\right] \\
&\leq \mathbb{E}_{\mathcal{C},X}\left[\sum_{x' \neq X} \mathbf{1}_{\left\{\log_2 \frac{1}{P_X(x')} < k-\tau\right\}} \mathbf{1}_{\{c_{x'} = c_X\}}\right] \qquad \text{(union bound)} \\
&= 2^{-k}\mathbb{E}_X\left[\sum_{x' \neq X} \mathbf{1}_{\{P_X(x') > 2^{-k+\tau}\}}\right] \\
&\leq 2^{-k} \sum_{x' \in \mathcal{X}} \mathbf{1}_{\{P_X(x') > 2^{-k+\tau}\}} \\
&\leq 2^{-k} 2^{k-\tau} = 2^{-\tau}. \qquad \square
\end{aligned}$$

Note: Why the proof works: Compressor $f(x) = c_x$, hashing $x \in \mathcal{X}$ to a random k-bit string $c_x \in \{0,1\}^k$.



high-probability $x \Leftrightarrow \log_2 \frac{1}{P_X(x)} \leq k - \tau \Leftrightarrow P_X(x) \geq 2^{-k+\tau}$.

Therefore the cardinality of high-probability x's is at most $2^{k-\tau} \ll 2^k$ = number of strings. Hence the chance of collision is small.

Note: The random coding argument is a canonical example of *probabilistic method*: To prove the existence of something with certain property, we construct a probability distribution (randomize) and show that on average the property is satisfied. Hence there exists at least one realization with the desired property. The downside of this argument is that it is not constructive, i.e., does not give us an algorithm to find the object.

Note: This is a subtle point: Notice that in the proof we choose the random codebook to be uniform over all possible codebooks. In other words, $C = \{c_x : x \in \mathcal{X}\}$ consists of iid k-bit strings. In fact, in the proof we only need pairwise independence, i.e., $c_x \perp\!\!\!\perp c_{x'}$ for any $x \neq x'$ (Why?). Now, why should we care about this? In fact, having access to external randomness is also a lot of resources. It is more desirable to use less randomness in the random coding argument. Indeed, if we use zero randomness, then it is a deterministic construction, which is the best situation! Using pairwise independent codebook requires significantly less randomness than complete random coding which needs $|\mathcal{X}|k$ bits. To see this intuitively, note that one can use 2 independent random bits to generate 3 random bits that is pairwise independent but not mutually independent, e.g., $\{b_1, b_2, b_1 \oplus b_2\}$. This observation is related to linear compression studied in the next section, where the codeword we generated are not iid, but related through a linear mapping.

Remark 7.1 (AEP for memoryless sources)**.** Consider iid S^n. By WLLN,

$$\frac{1}{n} \log \frac{1}{P_{S^n}(S^n)} \xrightarrow{\mathbb{P}} H(S). \tag{7.3}$$

For any $\delta > 0$, define the set

$$T_n^\delta = \left\{s^n : \left|\frac{1}{n} \log \frac{1}{P_{S^n}(s^n)} - H(S)\right| \leq \delta\right\}.$$

As a consequence of (7.3),

1. $\mathbb{P}\left[S^n \in T_n^\delta\right] \to 1$ as $n \to \infty$.
2. $|T_n^\delta| \le 2^{(H(S)+\delta)n} \ll |\mathcal{S}|^n$.

In other words, S^n is concentrated on the set T_n^δ which is exponentially smaller than the whole space. In almost compression we can simply encode this set losslessly. Although this is different than the optimal encoding, Corollary 7.1 indicates that in the large-n limit the optimal compressor is no better.

The property (7.3) is often referred as the *Asymptotic Equipartition Property* (AEP). Note that for any $s^n \in T_n^\delta$, its likelihood is concentrated around $P_{S^n}(s^n) \in 2^{-(H(S)\pm\delta)n}$, called δ-typical sequences.

Next we study fixed-blocklength code, fundamental limit of error probability $\epsilon^*(X,k)$ for the following coding paradigms:

- Linear Compression
- Compression with Side Information
  - side info available at both sides
  - side info available only at decompressor
  - multi-terminal compressor, single decompressor

7.2 Linear Compression

From Shannon's theorem:

$$\epsilon^*(X,nR) \longrightarrow 0 \text{ or } 1 \qquad\qquad R \gtrless H(S)$$

Our goal is to find compressor with structures. The simplest one can think of is probably linear operation, which is also highly desired for its simplicity (low complexity). But of course, we have to be on a vector space where we can define linear operations. In this part, we assume $X = S^n$, where each coordinate takes values in a finite field (Galois Field), i.e., $S_i \in \mathbb{F}_q$, where q is the cardinality of $\mathbb{F}_q$. This is only possible if $q = p^n$ for some prime p and $n \in \mathbb{N}$. So $\mathbb{F}_q = \mathbb{F}_{p^n}$.

Definition 7.2 (Galois Field)**.** F is a finite set with operations $(+,\cdot)$ where

- $a+b$ associative and commutative
- $a \cdot b$ associative and commutative
- $0, 1 \in F$ s.t. $0 + a = 1 \cdot a = a$.
- $\forall a, \exists - a$, s.t. $a + (-a) = 0$
- $\forall a \neq 0, \exists a^{-1}$, s.t. $a^{-1}a = 1$
- distributive: $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$

Example:

- $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, where p is prime
- $\mathbb{F}_4 = \{0, 1, x, x+1\}$ with addition and multiplication as polynomials $\mod (x^2+x+1)$ over $\mathbb{F}_2[x]$.

Linear Compression Problem: $x \in \mathbb{F}_q^n$, $w = Hx$ where $H : \mathbb{F}_q^n \to \mathbb{F}_q^k$ is linear represented by a matrix $H \in \mathbb{F}_q^{k \times n}$.

$$\begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix} = \begin{bmatrix} h_{11} & \dots & h_{1n} \\ \vdots & & \vdots \\ h_{k1} & \dots & h_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Compression is achieved if $k \leq n$, i.e., H is a fat matrix. Of course, we have to tolerate some error (almost lossless). Otherwise, lossless compression is only possible with $k \geq n$, which not interesting.

Theorem 7.5 (Achievability). *Let $X \in \mathbb{F}_q^n$ be a random vector. $\forall \tau > 0, \exists$ linear compressor $H : \mathbb{F}_q^n \to \mathbb{F}_q^k$ and decompressor $g : \mathbb{F}_q^k \to \mathbb{F}_q^n \cup \{\mathsf{e}\}$, s.t.*

$$\mathbb{P}\left[g(HX) \neq X\right] \leq \mathbb{P}\left[\log_q \frac{1}{P_X(X)} > k - \tau\right] + q^{-\tau}$$

Proof. Fix τ. As pointed in the proof of Shannon's random coding theorem (Theorem 7.4), given the compressor H, the optimal decompressor is the MAP decoder, i.e., $g(w) = \text{argmax}_{x:Hx=w} P_X(x)$, which outputs the most likely symbol that is compatible with the codeword received. Instead, let us consider the following (suboptimal) decoder for its ease of analysis:

$$g(w) = \begin{cases} x & \exists! x \in \mathbb{F}_q^n : w = Hx,\ x - h.p. \\ \mathsf{e} & \text{otherwise} \end{cases}$$

where we used the short-hand:

$$x - h.p. \text{ (high probability)} \Leftrightarrow \log_q \frac{1}{P_X(x)} < k - \tau \Leftrightarrow P_X(x) \geq q^{-k+\tau}.$$

Note that this decoder is the same as in the proof of Theorem 7.4. The proof is also mostly the same, except now hash collisions occur under the linear map H. By union bound,

$$\begin{aligned} \mathbb{P}\left[g(f(X)) = \mathsf{e}\right] &\leq \mathbb{P}\left[\log_q \frac{1}{P_X(x)} > k - \tau\right] + \mathbb{P}\left[\exists x' - h.p. : x' \neq X, Hx' = HX\right] \\ \text{(union bound)} &\leq \mathbb{P}\left[\log_q \frac{1}{P_X(x)} > k - \tau\right] + \sum_x P_X(x) \sum_{x'-h.p., x' \neq x} \mathbf{1}\{Hx' = Hx\} \end{aligned}$$

Now we use random coding to average the second term over all possible choices of H. Specifically, choose H as a matrix independent of X where each entry is iid and uniform on $\mathbb{F}_q$. For distinct x_0 and x_1, the collision probability is

$$\begin{aligned} \mathbb{P}_H[Hx_1 = Hx_0] &= \mathbb{P}_H[Hx_2 = 0] && (x_2 \triangleq x_1 - x_0 \neq 0) \\ &= \mathbb{P}_H[H_1 \cdot x_2 = 0]^k && \text{(iid rows)} \end{aligned}$$

where H_1 is the first row of the matrix H, and each row of H is independent. This is the probability that H_i is in the orthogonal complement of x_2. On $\mathbb{F}_q^n$, the orthogonal complement of a given non-zero vector has cardinality q^{n-1}. So the probability for the first row to lie in this subspace is $q^{n-1}/q^n = 1/q$, hence the collision probability $1/q^k$. Averaging over H gives

$$\mathbb{E}_H \sum_{x'-h.p., x' \neq x} \mathbf{1}\{Hx' = Hx\} = \sum_{x'-h.p., x' \neq x} \mathbb{P}_H[Hx' = H_x] = |\{x' : x' - h.p., x' \neq x\}| q^{-k} \leq q^{k-\tau} q^{-k} = q^{-\tau}$$

Thus the bound holds. □

Notes:

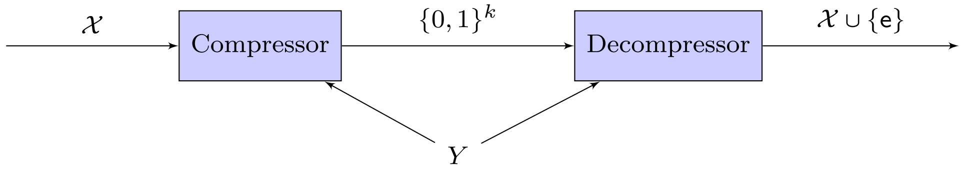
1. Compared to Theorem 7.4, which is obtained by randomizing over all possible compressors, Theorem 7.5 is obtained by randomizing over only linear compressors, and the bound we obtained is identical. Therefore restricting on linear compression almost does not lose anything.

2. Note that in this case it is not possible to make all errors detectable.
3. Can we loosen the requirement on $\mathbb{F}_q$ to instead be a commutative ring? In general, no, since zero divisors in the commutative ring ruin the key proof item of low collision probability in the random hashing. E.g. in $\mathbb{Z}/6\mathbb{Z}$

$$\mathbb{P}\left[H\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix}=0\right]=6^{-k}\qquad\text{but}\qquad\mathbb{P}\left[H\begin{bmatrix}2\\0\\\vdots\\0\end{bmatrix}=0\right]=3^{-k},$$

since $0\cdot 2=3\cdot 2=0$ in $\mathbb{Z}/6\mathbb{Z}$.

7.3 Compression with Side Information at both compressor and decompressor



Definition 7.3 (Compression wih Side Information)**.** Given P_{XY},

- $f:\mathcal{X}\times\mathcal{Y}\to\{0,1\}^k$
- $g:\{0,1\}^k\times\mathcal{Y}\to\mathcal{X}\cup\{\mathsf{e}\}$
- $\mathbb{P}[g(f(X,Y),Y)\neq X]<\epsilon$
- Fundamental Limit: $\epsilon^*(X|Y,k)=\inf\{\epsilon:\ \exists(k,\epsilon)-S.I.\ code\}$

Note: The side information Y need not be discrete. The source X is, of course, discrete.

Note that conditioned on $Y=y$, the problem reduces to compression without side information where the source X is distributed according to $P_{X|Y=y}$. Since Y is known to both the compressor and decompressor, they can use the best code tailored for this distribution. Recall $\epsilon^*(X,k)$ defined in Definition 7.1, the optimal probability of error for compressing X using k bits, which can also be denoted by $\epsilon^*(P_X,k)$. Then we have the following relationship

$$\epsilon^*(X|Y,k)=\mathbb{E}_{y\sim P_Y}[\epsilon^*(P_{X|Y=y},k)],$$

which allows us to apply various bounds developed before.

Theorem 7.6.

$$\mathbb{P}\left[\log\frac{1}{P_{X|Y}(X|Y)}>k+\tau\right]-2^{-\tau}\leq\epsilon^*(X|Y,k)\leq\mathbb{P}\left[\log_2\frac{1}{P_{X|Y}(X|Y)}>k-\tau\right]+2^{-\tau},\quad\forall\tau>0$$

Corollary 7.2. *$(X,Y) = (S^n,T^n)$ where $(S_1,T_1),(S_2,T_2),\dots$ are iid pairs $\sim P_{ST}$*

$$\lim_{n\to\infty} \epsilon^*(S^n|T^n, nR) = \begin{cases} 0 & R > H(S|T) \\ 1 & R < H(S|T) \end{cases}$$

Proof. Using the converse Theorem 7.2 and achievability Theorem 7.4 (or Theorem 7.3) for compression without side information, we have

$$\mathbb{P}\left[\log\frac{1}{P_{X|Y}(X|y)} > k+\tau|Y=y\right] - 2^{-\tau} \le \epsilon^*(P_{X|Y=y},k) \le \mathbb{P}\left[\log\frac{1}{P_{X|Y}(X|y)} > k|Y=y\right]$$

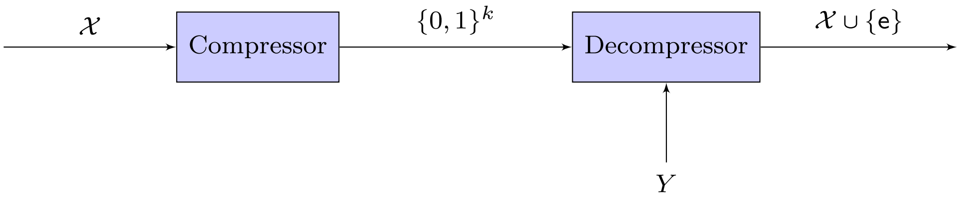
By taking the average over all $y \sim P_Y$, we get the theorem. For the corollary

$$\frac{1}{n}\log\frac{1}{P_{S^n|T^n}(S^n|T^n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{P_{S|T}(S_i|T_i)} \longrightarrow H(S|T) \text{ (in probability)}$$

as $n \to \infty$, using the WLLN. □

7.4 Slepian-Wolf (Compression with Side Information at Decompressor only)

Consider the compression with side information problem, except now the compressor has no access to the side information.



Definition 7.4 (S.W. code)**.** Given P_{XY},

- $f:\mathcal{X}\to\{0,1\}^k$
- $g:\{0,1\}^k\times\mathcal{Y}\to\mathcal{X}\cup\{\mathsf{e}\}$
- $\mathbb{P}[g(f(X),Y)\ne X]\le\epsilon$
- Fundamental Limit: $\epsilon^*_{\text{SW}} = \inf\{\epsilon : \exists(k,\epsilon)\text{-S.W. code}\}$

Now the very surprising result: Even without side information at the compressor, we can still compress down to the conditional entropy!

Theorem 7.7 (Slepian-Wolf, '73)**.**

$$\epsilon^*(X|Y,k) \le \epsilon^*_{\text{SW}}(X|Y,k) \le \mathbb{P}\left[\log\frac{1}{P_{X|Y}(X|Y)} \ge k-\tau\right] + 2^{-\tau}$$

Corollary 7.3.

$$\lim_{n\to\infty} \epsilon^*_{\mathrm{SW}}(S^n|T^n, nR) = \begin{cases} 0 & R > H(S|T) \\ 1 & R < H(S|T) \end{cases}$$

Note: Definition 7.4 does not include the zero-undected-error condition (that is $g(f(x), y) = x$ or e). In other words, we allow for the possibility of undetected errors. Indeed, if we require this condition, the side-information savings will be mostly gone. Indeed, assuming $P_{X,Y}(x,y) > 0$ for all (x, y) it is clear that under zero-undetected-error condition, if $f(x_1) = f(x_2) = c$ then $g(c) = \mathsf{e}$. Thus except for c all other elements in $\{0,1\}^k$ must have unique preimages. Similarly, one can show that Slepian-Wolf theorem does not hold if one uses the setting of variable-length lossless compression (i.e. average length is $H(X)$ not $H(X|Y)$.)

Proof. LHS is obvious, since side information at the compressor and decoder is better than only at the decoder.

For the RHS, first generate a random codebook with iid uniform codewords: $C = \{c_x \in \{0,1\}^k : x \in \mathcal{X}\}$ independently of (X,Y), then define the compressor and decoder as

$$f(x) = C_x$$

$$g(w, y) = \begin{cases} x & \exists! x : C_x = w, x - h.p.|y \\ 0 & \text{o.w.} \end{cases}$$

where we used the shorthand $x - h.p.|y \Leftrightarrow \log_2 \frac{1}{P_{X|Y}(x|y)} < k - \tau$. The error probability of this scheme is

$$\begin{aligned} \mathcal{E}(C) &= \mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau \text{ or } J(X, C|Y) \neq \varnothing\right] \\ &\leq \mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau\right] + \mathbb{P}[J(X, C|Y) \neq \varnothing] \\ &= \mathbb{P}\left[\log \frac{1}{P_{X|Y}(X|Y)} \geq k - \tau\right] + \sum_{x,y} P_{XY}(x,y) \mathbf{1}_{\{J(x,C|y)\neq\varnothing\}}. \end{aligned}$$

where $J(x, C|y) \triangleq \{x' \neq x : x' - h.p.|y, c_x = c_{x'}\}$.

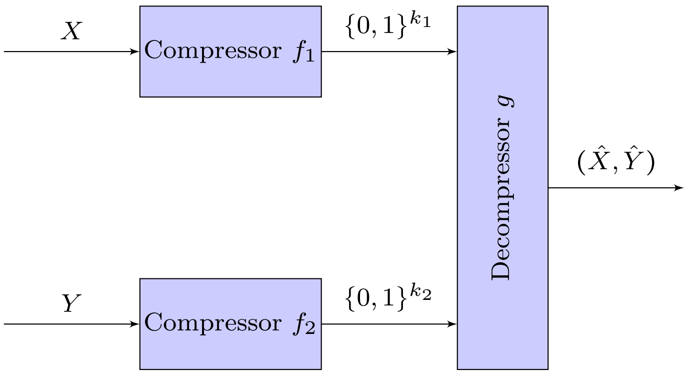
Now averaging over C and applying the union bound: use $|\{x' : x' - h.p.|y\}| \leq 2^{k-\tau}$ and $\mathbb{P}[C_{x'} = C_x] = 2^{-k}$ for any $x \neq x'$,

$$\begin{aligned} \mathbb{P}_C[J(x, C|y) \neq \varnothing] &\leq \mathbb{E}_C\left[\sum_{x'\neq x} \mathbf{1}_{\{x' - h.p.|y\}} \mathbf{1}_{\{C_{x'} = C_x\}}\right] \\ &= 2^{k-\tau} \mathbb{P}[C_{x'} = C_x] \\ &= 2^{-\tau} \end{aligned}$$

Hence the theorem follows as usual from two terms in the union bound. □

7.5 Multi-terminal Slepian Wolf

Distributed compression: Two sources are correlated. Compress individually, decompress jointly. What are those rate pairs that guarantee successful reconstruction?



Definition 7.5. Given P_{XY},

- (f_1,f_2,g) is (k_1,k_2,ϵ)-code if $f_1:\mathcal{X}\to\{0,1\}^{k_1}$, $f_2:\mathcal{Y}\to\{0,1\}^{k_2}$, $g:\{0,1\}^{k_1}\times\{0,1\}^{k_2}\to\mathcal{X}\times\mathcal{Y}$, s.t. $\mathbb{P}[(\hat{X},\hat{Y})\neq(X,Y)]\le\epsilon$, where $(\hat{X},\hat{Y})=g(f_1(X),f_2(Y))$.
- Fundamental limit: $\epsilon^*_{\text{SW}}(X,Y,k_1,k_2)=\inf\{\epsilon:\exists(k_1,k_2,\epsilon)\text{-code}\}$.

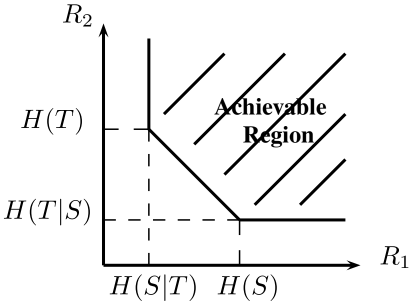
Theorem 7.8. $(X,Y)=(S^n,T^n)$ *- iid pairs*

$$\lim_{n\to\infty}\epsilon^*_{\text{SW}}(S^n,T^n,nR_1,nR_2)=\begin{cases}0 & (R_1,R_2)\in\mathcal{R}_{\text{SW}}\\ 1 & (R_1,R_2)\notin\mathcal{R}_{\text{SW}}\end{cases}$$

where $\mathcal{R}_{\text{SW}}$ *denotes the Slepian-Wolf rate region*

$$\mathcal{R}_{\text{SW}}=\left\{(a,b):\begin{array}{c}a\ge H(S|T)\\ b\ge H(T|S)\\ a+b\ge H(S,T)\end{array}\right.$$

Note: The rate region $\mathcal{R}_{\text{SW}}$ typically looks like:

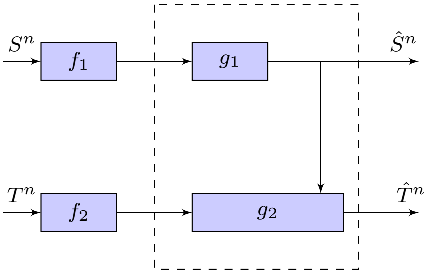


Since $H(T)-H(T|S)=H(S)-H(S|T)=I(S;T)$, the slope is -1.

Proof. Converse: Take $(R_1,R_2)\notin\mathcal{R}_{\text{SW}}$. Then one of three cases must occur:

1. $R_1<H(S|T)$. Then even if encoder and decoder had full T^n, still can't achieve this (from compression with side info result – Corollary 7.2).
2. $R_2<H(T|S)$ (same).
3. $R_1+R_2<H(S,T)$. Can't compress below the joint entropy of the pair (S,T).

Achievability: First note that we can achieve the two corner points. The point $(H(S), H(T|S))$ can be approached by almost lossless compressing S at entropy and compressing T with side information S at the decoder. To make this rigorous, let $k_1 = n(H(S)+\delta)$ and $k_2 = n(H(T|S)+\delta)$. By Corollary 7.1, there exist $f_1 : \mathcal{S}^n \to \{0,1\}^{k_1}$ and $g_1 : \{0,1\}^{k_1} \to \mathcal{S}^n$ s.t. $\mathbb{P}[g_1(f_1(S^n)) \neq S^n] \le \epsilon_n \to 0$. By Theorem 7.7, there exist $f_2 : \mathcal{T}^n \to \{0,1\}^{k_2}$ and $g_2 : \{0,1\}^{k_1} \times \mathcal{S}^n \to \mathcal{T}^n$ s.t. $\mathbb{P}[g_2(f_2(T^n), S^n) \neq T^n] \le \epsilon_n \to 0$. Now that S^n is not available, feed the S.W. decompressor with $g(f(S^n))$ and define the joint decompressor by $g(w_1, w_2) = (g_1(w_1), g_2(w_2, g_1(w_1)))$ (see below):



Apply union bound:

$$\begin{aligned}
&\mathbb{P}[g(f_1(S^n), f_2(T^n)) \neq (S^n, T^n)] \\
&= \mathbb{P}[g(f_1(S^n)) \neq S^n] + \mathbb{P}[g_2(f_2(T^n), g(f_1(S^n))) \neq T^n, g(f_1(S^n)) = S^n] \\
&\le \mathbb{P}[g(f_1(S^n)) \neq S^n] + \mathbb{P}[g_2(f_2(T^n), S^n) \neq T^n] \\
&\le 2\epsilon_n \to 0.
\end{aligned}$$

Similarly, the point $(H(S), H(T|S))$ can be approached.

To achieve other points in the region, use the idea of **time sharing**: If you can achieve with vanishing error probability any two points (R_1, R_2) and (R_1', R_2'), then you can achieve for $\lambda \in [0,1]$, $(\lambda R_1 + \bar{\lambda} R_1', \lambda R_2 + \bar{\lambda} R_2')$ by dividing the block of length n into two blocks of length λn and $\bar{\lambda} n$ and apply the two codes respectively

$$\begin{aligned}
(S_1^{\lambda n}, T_1^{\lambda n}) &\to \begin{bmatrix} \lambda n R_1 \\ \lambda n R_2 \end{bmatrix} \quad \text{using } (R_1, R_2) \text{ code} \\
(S_{\lambda n+1}^{n}, T_{\lambda n+1}^{n}) &\to \begin{bmatrix} \bar{\lambda} n R_1' \\ \bar{\lambda} n R_2' \end{bmatrix} \quad \text{using } (R_1', R_2') \text{ code}
\end{aligned}$$

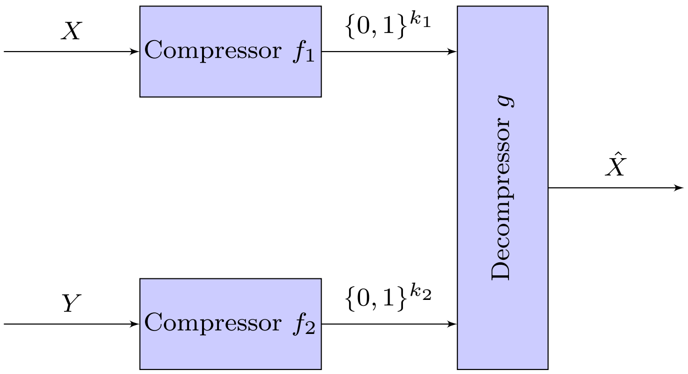
(Exercise: Write down the details rigorously yourself!) Therefore, all convex combinations of points in the achievable regions are also achievable, so the achievable region must be convex. □

7.6* Source-coding with a helper (Ahlswede-Körner-Wyner)

Yet another variation of distributed compression problem is compressing X with a helper, see figure below. Note that the main difference from the previous section is that decompressor is only required to produce the estimate of X, using rate-limited help from an observer who has access to Y. Characterization of rate pairs R_1, R_2 is harder than in the previous section.

Theorem 7.9 (Ahlswede-Körner-Wyner). *Consider i.i.d. source $(X^n, Y^n) \sim P_{X,Y}$ with X discrete. If rate pair (R_1, R_2) is achievable with vanishing probability of error $\mathbb{P}[\hat{X}^n \neq X^n] \to 0$, then there exists an auxiliary random variable U taking values on alphabet of cardinality $|\mathcal{Y}| + 1$ such that $P_{X,Y,U} = P_{X,Y} P_{U|X,Y}$ and*

$$R_1 \ge H(X|U), R_2 \ge I(Y;U). \tag{7.4}$$



Furthermore, for every such random variable U the rate pair $(H(X|U), I(Y;U))$ is achievable with vanishing error.

Proof. We only sketch some crucial details.

First, note that iterating over all possible random variables U (without cardinality constraint) the set of pairs (R_1, R_2) satisfying (7.4) is convex. Next, consider a compressor $W_1 = f_1(X^n)$ and $W_2 = f_2(Y^n)$. Then from Fano's inequality (5.7) assuming $\mathbb{P}[X^n \neq \hat{X}^n] = o(1)$ we have

$$H(X^n|W_1, W_2)) = o(n)\,.$$

Thus, from chain rule and conditioning-decreases-entropy, we get

$$\begin{aligned} nR_1 \ge I(X^n; W_1|W_2) &\ge H(X^n|W_2) - o(n) && (7.5)\\ &= \sum_{k=1}^n H(X_k|W_2, X^{k-1}) - o(n) && (7.6)\\ &\ge \sum_{k=1}^n H(X_k|W_2, X^{k-1}, Y^{k-1}) - o(n) && (7.7)\end{aligned}$$

On the other hand, from (5.2) we have

$$\begin{aligned} nR_2 \ge I(W_2; Y^n) &= \sum_{k=1}^n I(W_2; Y_k|Y^{k-1}) && (7.8)\\ &= \sum_{k=1}^n I(W_2, X^{k-1}; Y_k|Y^{k-1}) && (7.9)\\ &= \sum_{k=1}^n I(W_2, X^{k-1}, Y^{k-1}; Y_k) && (7.10)\end{aligned}$$

where (7.9) follows from $I(W_2, X^{k-1}; Y_k|Y^{k-1}) = I(W_2; Y_k|Y^{k-1}) + I(X^{k-1}; Y_k|W_2, Y^{k-1})$ and the fact that $(W_2, Y_k) \perp\!\!\!\perp X^{k-1}|Y^{k-1}$; and (7.10) from $Y^{k-1} \perp\!\!\!\perp Y_k$. Comparing (7.7) and (7.10) we notice that denoting $U_k = (W_2, X^{k-1}, Y^{k-1})$ we have

$$(R_1, R_2) \ge \frac{1}{n}\sum_{k=1}^n (H(X_k|U_k), I(U_k; Y_k))$$

and thus (from convexity) the rate pair must belong to the region spanned by all pairs $(H(X|U), I(U;Y))$.

To show that without loss of generality the auxiliary random variable U can be taken to be $|\mathcal{Y}| + 1$ valued, one needs to invoke Caratheodory's theorem on convex hulls. We omit the details.

Finally, showing that for each U the mentioned rate-pair is achievable, we first notice that if there were side information at the decompressor in the form of the i.i.d. sequence U^n correlated to X^n, then Slepian-Wolf theorem implies that only rate $R_1 = H(X|U)$ would be sufficient to reconstruct X^n. Thus, the question boils down to creating a correlated sequence U^n at the decompressor by using the minimal rate R_2. This is the content of the so called covering lemma, see Theorem 24.5 below: It is sufficient to use rate $I(U;Y)$ to do so. We omit further details. □

§ 8. Compressing stationary ergodic sources

We have examined the compression of i.i.d. sequence $\{S_i\}$, for which

$$\frac{1}{n} l(f^*(S^n)) \to H(S) \quad \text{in prob.} \tag{8.1}$$

$$\lim_{n\to\infty} \epsilon^*(S^n, nR) = \begin{cases} 0 & R > H(S) \\ 1 & R < H(S) \end{cases} \tag{8.2}$$

In this lecture, we shall examine similar results for ergodic processes and we first state the main theory as follows:

Theorem 8.1 (Shannon-McMillan). *Let $\{S_1, S_2, \dots\}$ be a stationary and ergodic discrete process, then*

$$\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} \xrightarrow{\mathbb{P}} \mathcal{H}, \quad \textit{also a.s. and in } L_1 \tag{8.3}$$

where $\mathcal{H} = \lim_{n\to\infty} \frac{1}{n} H(S^n)$ is the entropy rate.

Corollary 8.1. *For any stationary and ergodic discrete process $\{S_1, S_2, \dots\}$, (8.1) – (8.2) hold with $H(S)$ replaced by $\mathcal{H}$.*

Proof. Shannon-McMillan (we only need convergence in probability) + Theorem 6.4 + Theorem 7.1 which tie together the respective CDF of the random variable $l(f^*(S^n))$ and $\log\frac{1}{P_{S^n}(s^n)}$. □

In Lecture 7 we learned the asymptotic equipartition property (AEP) for iid sources. Here we generalize it to stationary ergodic sources thanks to Shannon-McMillan.

Corollary 8.2 (AEP for stationary ergodic sources). *Let $\{S_1, S_2, \dots\}$ be a stationary and ergodic discrete process. For any $\delta > 0$, define the set*

$$T_n^\delta = \left\{ s^n : \left| \frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \mathcal{H} \right| \le \delta \right\}.$$

Then

1. $\mathbb{P}\left[S^n \in T_n^\delta\right] \to 1$ *as* $n \to \infty$.
2. $2^{n(\mathcal{H}-\delta)}(1 + o(1)) \le |T_n^\delta| \le 2^{(\mathcal{H}+\delta)n}(1 + o(1))$.

Note:

- Convergence in probability for stationary ergodic Markov chains [Shannon 1948]
- Convergence in L_1 for stationary ergodic processes [McMillan 1953]

- Convergence almost surely for stationary ergodic processes [Breiman 1956] (Either of the last two results implies the convergence Theorem 8.1 in probability.)
- For a Markov chain, existence of typical sequences can be understood by thinking of Markov process as sequence of independent decisions regarding which transitions to take. It is then clear that Markov process's trajectory is simply a transformation of trajectories of an i.i.d. process, hence must similarly concentrate similarly on some typical set.

8.1 Bits of ergodic theory

Let's start with a dynamic system view and introduce a few definitions:

Definition 8.1 (Measure preserving transformation). $\tau : \Omega \to \Omega$ is measure preserving (more precisely, probability preserving) if

$$\forall E \in \mathcal{F}, P(E) = P(\tau^{-1}E).$$

The set E is called τ-invariant if $E = \tau^{-1}E$. The set of all τ-invariant sets forms a σ-algrebra (check!) denoted $\mathcal{F}_{inv}$.

Definition 8.2 (stationary process). A process $\{S_n, n = 0, \ldots\}$ is stationary if there exists a measure preserving transformation $\tau : \Omega \to \Omega$ such that:

$$S_j = S_{j-1} \circ \tau = S_0 \circ \tau^j$$

Therefore a stationary process can be described by the tuple $(\Omega, \mathcal{F}, \mathbb{P}, \tau, S_0)$ and $S_k = S_0 \circ \tau^k$.

Notes:

1. Alternatively, a random process $(S_0, S_1, S_2, \ldots)$ is stationary if its joint distribution is invariant with respect to shifts in time, i.e., $P_{S_n^m} = P_{S_{n+t}^{m+t}}$, $\forall n, m, t$. Indeed, given such a process we can define a m.p.t. as follows:
$$(s_0, s_1, \ldots) \xrightarrow{\tau} (s_1, s_2, \ldots) \tag{8.4}$$
So τ is a shift to the right.

2. An event $E \in \mathcal{F}$ is shift-invariant if
$$(s_1, s_2, \ldots) \in E \Rightarrow \forall s_0 (s_0, s_1, s_2, \ldots) \in E$$
or equivalently $E = \tau^{-1}E$ (check!). Thus τ-invariant events are also called shift-invariant, when τ is interpreted as (8.4).

3. Some examples of shift-invariant events are $\{\exists n : x_i = 0 \forall i \geq n\}$, $\{\limsup x_i < 1\}$ etc. A non shift-invariant event is $A = \{x_0 = x_1 = \cdots = 0\}$, since $\tau(1,0,0,\ldots) \in A$ but $(1,0,\ldots) \notin A$.

4. Also recall that the tail σ-algebra is defined as
$$\mathcal{F}_{tail} \triangleq \bigcap_{n \geq 1} \sigma\{S_n, S_{n+1}, \ldots\}.$$
It is easy to check that all shift-invariant events belong to $\mathcal{F}_{tail}$. The inclusion is strict, as for example the event
$$\{\exists n : x_i = 0, \forall \ \underline{\text{odd}} \ i \geq n\}$$
is in $\mathcal{F}_{tail}$ but not shift-invariant.

Proposition 8.1 (Poincare recurrence)**.** *Let τ be measure-preserving for $(\Omega, \mathcal{F}, \mathbb{P})$. Then for any measurable A with $\mathbb{P}[A] > 0$ we have*

$$\mathbb{P}[\bigcup_{k\geq 1} \tau^{-k} A | A] = \mathbb{P}[\tau^k(\omega) \in A - -\textit{infinitely often} | A] = 1\,.$$

Proof. Let $B = \bigcup_{k\geq 1} \tau^{-k} A$. It is sufficient to show that $\mathbb{P}[A \cap B] = \mathbb{P}[A]$ or equivalently

$$\mathbb{P}[A \cup B] = \mathbb{P}[B]\,. \tag{8.5}$$

To that end notice that $\tau^{-1}A \cup \tau^{-1}B = B$ and thus

$$\mathbb{P}[\tau^{-1}(A \cup B)] = \mathbb{P}[B]\,,$$

but the left-hand side equals $\mathbb{P}[A \cup B]$ by the measure-preservation of τ, proving (8.5). □

Note: Consider τ mapping initial state of the conservative (Hamiltonian) mechanical system to its state after passage of a given unit of time. It is known that τ preserves Lebesgue measure in phase space (Liouville's theorem). Thus Poincare recurrence leads to rather counter-intuitive conclusions. For example, opening the barrier separating two gases in a cylinder allows them to mix. Poincare recurrence says that eventually they will return back to the original separated state (with each gas occupying roughly its half of the cylinder).

Definition 8.3 (Ergodicity)**.** A transformation τ is ergodic if $\forall E \in \mathcal{F}_{inv}$ we have $\mathbb{P}[E] = 0$ or 1. A process $\{S_i\}$ is ergodic if all shift invariant events are deterministic, i.e., for any shift invariant event E, $\mathbb{P}\left[S_1^\infty \in E\right] = 0$ or 1.

Example:

- $\{S_k = k^2\}$: ergodic but not stationary
- $\{S_k = S_0\}$: stationary but not ergodic (unless S_0 is a constant). Note that the singleton set $E = \{(s, s, \ldots)\}$ is shift invariant and $\mathbb{P}\left[S_1^\infty \in E\right] = \mathbb{P}\left[S_0 = s\right] \in (0,1)$ – not deterministic.
- $\{S_k\}$ i.i.d. is stationary and ergodic (by Kolmogorov's 0-1 law, tail events have no randomness)
- (Sliding-window construction of ergodic processes)
  If $\{S_i\}$ is ergodic, then $\{X_i = f(S_i, S_{i+1}, \ldots)\}$ is also ergodic. It is called a **B-process** if S_i is i.i.d.
  Example, $S_i \sim \text{Bern}(\frac{1}{2})$ i.i.d., $X_k = \sum_{n=0}^\infty 2^{-n-1} S_{k+n} = 2X_{k-1} \mod 1$. The marginal distribution of X_i is uniform on $[0,1]$. *Note that X_k's behavior is completely deterministic:* given X_0, all the future X_k's are determined exactly. This example shows that certain deterministic maps exhibit ergodic/chaotic behavior under iterative application: although the trajectory is completely deterministic, its time-averages converge to expectations and in general "look random".
- There are also stronger conditions than ergodicity. Namely, we say that τ is mixing (or strong mixing) if

  $$\mathbb{P}[A \cap \tau^{-n} B] \to \mathbb{P}[A]\mathbb{P}[B]\,.$$

  We say that τ is weakly mixing if

  $$\sum_{k=1}^{n} \frac{1}{n} |\mathbb{P}[A \cap \tau^{-n} B] - \mathbb{P}[A]\mathbb{P}[B]| \to 0\,.$$

  Strong mixing implies weak mixing, which implies ergodicity (check!).

- $\{S_i\}$: finite irreducible Markov chain with recurrent states is ergodic (in fact strong mixing), regardless of initial distribution.
  Toy example: kernel $P(0|1) = P(1|0) = 1$ with initial dist. $P(S_0 = 0) = 0.5$. This process only has two sample paths: $\mathbb{P}[S_1^\infty = (010101\ldots)] = \mathbb{P}[S_1^\infty = (101010\ldots)] = \frac{1}{2}$. It is easy to verify this process is ergodic (in the sense defined above!). Note however, that in Markov-chain literature a chain is called ergodic if it is irreducible, aperiodic and recurrent. This example does not satisfy this definition (this clash of terminology is a frequent source of confusion).

- (optional) $\{S_i\}$: stationary zero-mean Gaussian process with autocovariance function $R(n) = \mathbb{E}[S_0 S_n^*]$.

$$\lim_{n\to\infty} \frac{1}{n+1}\sum_{t=0}^{n} R[t] = 0 \Leftrightarrow \{S_i\} \text{ ergodic} \Leftrightarrow \{S_i\} \text{ weakly mixing}$$
$$\lim_{n\to\infty} R[n] = 0 \Leftrightarrow \{S_i\} \text{ mixing}$$

  Intuitively speaking, an ergodic process can have infinite memory in general, but the memory is weak. Indeed, we see that for a stationary Gaussian process ergodicity means the correlation dies (in the Cesaro-mean sense).

  The *spectral measure* is defined as the (discrete time) Fourier transform of the autocovariance sequence $\{R(n)\}$, in the sense that there exists a unique probability measure μ on $[-\frac{1}{2}, \frac{1}{2}]$ such that $R(n) = \mathbb{E}\exp(i2n\pi X)$ where $X \sim \mu$. The spectral criteria can be formulated as follows:

$$\{S_i\} \text{ ergodic} \Leftrightarrow \text{spectral measure has no atoms (CDF is continuous)}$$
$$\{S_i\} \text{ B-process} \Leftrightarrow \text{spectral measure has density}$$

  Detailed exposition on stationary Gaussian processes can be found in [Doo53, Theorem 9.3.2, pp. 474, Theorem 9.7.1, pp. 494–494].[1]

8.2 Proof of Shannon-McMillan

We shall show the convergence in L_1, which implies convergence in probability automatically. In order to prove Shannon-McMillan, let's first introduce the Birkhoff-Khintchine's convergence theorem for ergodic processes, the proof of which is presented in the next subsection.

Theorem 8.2 (Birkhoff-Khintchine's Ergodic Theorem). *If $\{S_i\}$ stationary and ergodic, $\forall$ function $f \in L_1$, i.e., $\mathbb{E}|f(S_1,\ldots)| < \infty$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(S_k,\ldots) = \mathbb{E} f(S_1,\ldots). \quad \textit{a.s. and in } L_1$$

In the special case where f depends on finitely many coordinates, say, $f = f(S_1,\ldots,S_m)$, we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(S_k,\ldots,S_{k+m-1}) = \mathbb{E} f(S_1,\ldots,S_m). \quad \textit{a.s. and in } L_1$$

Interpretation: time average converges to ensemble average.
Example: Consider $f = f(S_1)$

[1] Thanks Prof. Bruce Hajek for the pointer.

- $\{S_i\}$ is iid. Then Theorem 8.2 is SLLN (strong LLN).
- $\{S_i\}$ is such that $S_i = S_1$ for all i – non-ergodic. Then Theorem 8.2 fails unless S_1 is a constant.

Definition 8.4. $\{S_i : i \in \mathbb{N}\}$ is an m^{th} order Markov chain if $P_{S_{t+1}|S_1^t} = P_{S_{t+1}|S_{t-m+1}^t}$ for all $t \geq m$. It is called time homogeneous if $P_{S_{t+1}|S_{t-m+1}^t} = P_{S_{m+1}|S_1^m}$.

Remark 8.1. Showing (8.3) for an m^{th} order time homogeneous Markov chain $\{S_i\}$ is a direct application of Birkhoff-Khintchine.

$$\begin{aligned}\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} &= \frac{1}{n}\sum_{t=1}^{n}\log\frac{1}{P_{S_t|S^{t-1}}(S_t|S^{t-1})}\\ &= \frac{1}{n}\log\frac{1}{P_{S^m}(S^m)} + \frac{1}{n}\sum_{t=m+1}^{n}\log\frac{1}{P_{S_t|S_{t-m}^{t-1}}(S_l|S_{l-m}^{l-1})}\\ &= \underbrace{\frac{1}{n}\log\frac{1}{P_{S_1}(S_1^m)}}_{\to 0} + \underbrace{\frac{1}{n}\sum_{t=m+1}^{n}\log\frac{1}{P_{S_{m+1}|S_1^m}(S_t|S_{t-m}^{t-1})}}_{\to H(S_{m+1}|S_1^m)\text{ by Birkhoff-Khintchine}},\end{aligned} \tag{8.6}$$

where we applied Theorem 8.2 with $f(s_1, s_2, \ldots) = \log\frac{1}{P_{S_{m+1}|S_1^m}(s_{m+1}|s_1^m)}$.

Now let's prove (8.3) for a general stationary ergodic process $\{S_i\}$ which might have infinite memory. The idea is to approximate the distribution of that ergodic process by an m-th order MC (finite memory) and make use of (8.6); then let $m \to \infty$ to make the the approximation accurate (*Markov approximation*).

Proof of Theorem 8.1 in L_1. To show that (8.3) converges in L_1, we want to show that

$$\mathbb{E}\left|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \mathcal{H}\right| \to 0, \quad n \to \infty.$$

To this end, fix an $m \in \mathbb{N}$. Define the following auxiliary distribution for the process:

$$\begin{aligned}Q^{(m)}(S_1^\infty) &= P_{S_1^m}(S_1^m)\prod_{t=m+1}^{\infty}P_{S_t|S_{t-m}^{t-1}}(S_t|S_{t-m}^{t-1})\\ &\overset{\text{stat.}}{=} P_{S_1^m}(S_1^m)\prod_{t=m+1}^{\infty}P_{S_{m+1}|S_1^m}(S_t|S_{t-m}^{t-1})\end{aligned}$$

Note that under $Q^{(m)}$, $\{S_i\}$ is an m^{th}-order time-homogeneous Markov chain.

By triangle inequality,

$$\begin{aligned}\mathbb{E}\left|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \mathcal{H}\right| \leq &\underbrace{\mathbb{E}\left|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \frac{1}{n}\log\frac{1}{Q_{S^n}^{(m)}(S^n)}\right|}_{\triangleq A}\\ &+ \underbrace{\mathbb{E}\left|\frac{1}{n}\log\frac{1}{Q_{S^n}^{(m)}(S^n)} - H_m\right|}_{\triangleq B} + \underbrace{|H_m - \mathcal{H}|}_{\triangleq C}\end{aligned}$$

where $H_m \triangleq H(S_{m+1}|S_1^m)$.

Now

- $C = |H_m - \mathcal{H}| \to 0$ as $m \to \infty$ by Theorem 5.4 (Recall that for stationary processes: $H(S_{m+1}|S_1^m) \to H$ from above).
- As shown in Remark 8.1, for any fixed m, $B \to 0$ in L_1 as $n \to \infty$, as a consequence of Birkhoff-Khintchine. Hence for any fixed m, $\mathbb{E}B \to 0$ as $n \to \infty$.
- For term A,

  $$\mathbb{E}[A] = \frac{1}{n}\mathbb{E}_P\left|\log\frac{dP_{S^n}}{dQ^{(m)}_{S^n}}\right| \le \frac{1}{n}D(P_{S^n}\|Q^{(m)}_{S^n}) + \frac{2\log e}{en}$$

  where

  $$\begin{aligned}\frac{1}{n}D(P_{S^n}\|Q^{(m)}_{S^n}) &= \frac{1}{n}\mathbb{E}\left[\log\frac{P_{S^n}(S^n)}{P_{S^m}(S^m)\prod_{t=m+1}^n P_{S_{m+1}|S_m^1}(S_t|S_{t-m}^{t-1})}\right]\\ &\overset{\text{stat.}}{=}\frac{1}{n}(-H(S^n) + H(S^m) + (n-m)H_m)\\ &\to H_m - \mathcal{H} \text{ as } n \to \infty\end{aligned}$$

  and the next Lemma 8.1.

Combining all three terms and sending $n \to \infty$, we obtain for any m,

$$\limsup_{n\to\infty}\mathbb{E}\left|\frac{1}{n}\log\frac{1}{P_{S^n}(S^n)} - \mathcal{H}\right| \le 2(H_m - \mathcal{H}).$$

Sending $m \to \infty$ completes the proof of L_1-convergence. □

Lemma 8.1.

$$\mathbb{E}_P\left[\left|\log\frac{dP}{dQ}\right|\right] \le D(P\|Q) + \frac{2\log e}{e}.$$

Proof. $|x\log x| - x\log x \le \frac{2\log e}{e}$, $\forall x > 0$, since LHS is zero if $x \ge 1$, and otherwise upper bounded by $2\sup_{0\le x\le 1} x\log\frac{1}{x} = \frac{2\log e}{e}$. □

8.3* Proof of Birkhoff-Khintchine

Proof of Theorem 8.2. $\forall$ function $\tilde{f} \in L_1$, $\forall\epsilon$, there exists a decomposition $\tilde{f} = f + h$ such that f is bounded, and $h \in \mathcal{L}_1$, $\|h\|_1 \le \epsilon$.
Let us first focus on the bounded function f. Note that in the bounded domain $\mathcal{L}_1 \subset \mathcal{L}_2$, thus $f \in \mathcal{L}_2$. Furthermore, $\mathcal{L}_2$ is a Hilbert space with inner product $(f,g) = \mathbb{E}[f(S_1^\infty)\overline{g(S_1^\infty)}]$.

For the measure preserving transformation τ that generates the stationary process $\{S_i\}$, define the operator $T(f) = f \circ \tau$. Since τ is measure preserving, we know that $\|Tf\|_2^2 = \|f\|_2^2$, thus T is a unitary and bounded operator.

Define the operator

$$A_n(f) = \frac{1}{n}\sum_{k=1}^n f \circ \tau^k$$

Intuitively:

$$A_n = \frac{1}{n}\sum_{k=1}^n T^k = \frac{1}{n}(I - T^n)(I - T)^{-1}$$

Then, if $f \perp \ker(I - T)$ we should have $A_n f \to 0$, since only components in the kernel can blow up. This intuition is formalized in the proof below.

Let's further decompose f into two parts $f = f_1 + f_2$, where $f_1 \in \ker(I - T)$ and $f_2 \in \ker(I - T)^\perp$. Observations:

- if $g \in \ker(I - T)$, g must be a constant function. This is due to the ergodicity. Consider indicator function $\mathbf{1}_A$, if $\mathbf{1}_A = \mathbf{1}_A \circ \tau = \mathbf{1}_{\tau^{-1}A}$, then $\mathbb{P}[A] = 0$ or 1. For a general case, suppose $g = Tg$ and g is not constant, then at least some set $\{g \in (a, b)\}$ will be shift-invariant and have non-trivial measure, violating ergodicity.

- $\ker(I - T) = \ker(I - T^*)$. This is due to the fact that T is unitary:

$$g = Tg \Rightarrow \|g\|^2 = (Tg, g) = (g, T^* g) \Rightarrow (T^* g, g) = \|g\| \|T^* g\| \Rightarrow T^* g = g$$

  where in the last step we used the fact that Cauchy-Schwarz $(f, g) \le \|f\| \cdot \|g\|$ only holds with equality for $g = cf$ for some constant c.

- $\ker(I - T)^\perp = \ker(I - T^*)^\perp = [\mathrm{Im}(I - T)]$, where $[\mathrm{Im}(I - T)]$ is an $\mathcal{L}_2$ closure.

- $g \in \ker(I - T)^\perp \iff \mathbb{E}[g] = 0$. Indeed, only zero-mean functions are orthogonal to constants.

With these observations, we know that $f_1 = m$ is a const. Also, $f_2 \in [\mathrm{Im}(I - T)]$ so we further approximate it by $f_2 = f_0 + h_1$, where $f_0 \in \mathrm{Im}(I - T)$, namely $f_0 = g - g \circ \tau$ for some function $g \in \mathcal{L}_2$, and $\|h_1\|_1 \le \|h_1\|_2 < \epsilon$. Therefore we have

$$\begin{aligned}
A_n f_1 &= f_1 = \mathbb{E}[f] \\
A_n f_0 &= \frac{1}{n}(g - g \circ \tau^n) \to 0 \text{ a.s. and } L_1 \\
(\text{since } \mathbb{E}[\sum_{n \ge 1} (\frac{g \circ \tau^n}{n})^2] &= \mathbb{E}[g^2] \sum \frac{1}{n^2} < \infty \quad \Longrightarrow \quad \frac{1}{n} g \circ \tau^n \to 0 \text{ a.s.})
\end{aligned}$$

The proof completes by showing

$$\mathbb{P}[\limsup_n A_n(h + h_1) \ge \delta] \le \frac{2\epsilon}{\delta}. \tag{8.7}$$

Indeed, then by taking $\epsilon \to 0$ we will have shown

$$\mathbb{P}[\limsup_n A_n(f) \ge \mathbb{E}[f] + \delta] = 0$$

as required. □

Proof of (8.7) makes use of the Maximal Ergodic Lemma stated as follows:

Theorem 8.3 (Maximal Ergodic Lemma). *Let $(\mathbb{P}, \tau)$ be a probability measure and a measure-preserving transformation. Then for any $f \in L_1(\mathbb{P})$ we have*

$$\mathbb{P}\left[\sup_{n \ge 1} A_n f > a\right] \le \frac{\mathbb{E}[f \mathbf{1}_{\sup_{n \ge 1} A_n f > a}]}{a} \le \frac{\|f\|_1}{a}$$

where $A_n f = \frac{1}{n} \sum_{k=0}^{n-1} f \circ \tau^k$.

Note: This is a so-called "weak L_1" estimate for a sublinear operator $\sup_n A_n(\cdot)$. In fact, this theorem is exactly equivalent to the following result:

Lemma 8.2 (Estimate for the maximum of averages)**.** *Let $\{Z_n, n = 1, \dots\}$ be a stationary process with $\mathbb{E}[|Z|] < \infty$ then*

$$\mathbb{P}\left[\sup_{n\geq 1} \frac{|Z_1 + \ldots + Z_n|}{n} > a\right] \leq \frac{\mathbb{E}[|Z|]}{a} \qquad \forall a > 0$$

Proof. The argument for this Lemma has originally been quite involved, until a dramatically simple proof (below) was found by A. Garcia.

Define

$$S_n = \sum_{k=1}^{n} Z_k \tag{8.8}$$

$$L_n = \max\{0, Z_1, \dots, Z_1 + \dots + Z_n\} \tag{8.9}$$

$$M_n = \max\{0, Z_2, Z_2 + Z_3, \dots, Z_2 + \dots + Z_n\} \tag{8.10}$$

$$Z^* = \sup_{n\geq 1} \frac{S_n}{n} \tag{8.11}$$

It is sufficient to show that

$$\mathbb{E}[Z_1 1_{\{Z^*>0\}}] \geq 0\,. \tag{8.12}$$

Indeed, applying (8.12) to $\tilde{Z}_1 = Z_1 - a$ and noticing that $\tilde{Z}^* = Z^* - a$ we obtain

$$\mathbb{E}[Z_1 1_{\{Z^*>a\}}] \geq a\mathbb{P}[Z^* > a]\,,$$

from which Lemma follows by upper-bounding the left-hand side with $\mathbb{E}[|Z_1|]$.

In order to show (8.12) we first notice that $\{L_n > 0\} \nearrow \{Z^* > 0\}$. Next we notice that

$$Z_1 + M_n = \max\{S_1, \dots, S_n\}$$

and furthermore

$$Z_1 + M_n = L_n \qquad \text{on } \{L_n > 0\}$$

Thus, we have

$$Z_1 1_{\{L_n>0\}} = L_n - M_n 1_{\{L_n>0\}}$$

where we do not need indicator in the first term since $L_n = 0$ on $\{L_n > 0\}^c$. Taking expectation we get

$$\mathbb{E}[Z_1 1_{\{L_n>0\}}] = \mathbb{E}[L_n] - \mathbb{E}[M_n 1_{\{L_n>0\}}] \tag{8.13}$$

$$\geq \mathbb{E}[L_n] - \mathbb{E}[M_n] \tag{8.14}$$

$$= \mathbb{E}[L_n] - \mathbb{E}[L_{n-1}] = \mathbb{E}[L_n - L_{n-1}] \geq 0\,, \tag{8.15}$$

where we used $M_n \geq 0$, the fact that M_n has the same distribution as L_{n-1}, and $L_n \geq L_{n-1}$, respectively. Taking limit as $n \to \infty$ in (8.15) we obtain (8.12). □

8.4* Sinai's generator theorem

It turns out there is a way to associate to every probability-preserving transformation τ a number, called Kolmogorov-Sinai entropy. This number is invariant to isomorphisms of p.p.t.'s (appropriately defined).

Definition 8.5. Fix a probability-preserving transformation τ acting on probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Kolmogorov-Sinai entropy of τ is defined as

$$\mathcal{H}(\tau) \triangleq \sup_{X_0} \lim_{n\to\infty} \frac{1}{n} H(X_0, X_0 \circ \tau, \ldots, X_0 \circ \tau^{n-1}) ,$$

where supremum is taken over all random variables $X_0 : \Omega \to \mathcal{X}$ with finite range $\mathcal{X}$ and measurable with respect to $\mathcal{F}$.

Note that every random variable X_0 generates a stationary process adapted to τ, that is

$$X_k \triangleq X_0 \circ \tau^k .$$

In this way, Kolmogorov-Sinai entropy of τ equals the maximal entropy rate among all stationary processes adapted to τ. This quantity may be extremely hard to evaluate, however. One help comes in the form of the famous criterion of Y. Sinai. We need to elaborate on some more concepts before:

- σ-algebra $\mathcal{G} \subset \mathcal{F}$ is $\mathbb{P}$-dense in $\mathcal{F}$, or sometimes we also say $\mathcal{G} = \mathcal{F} \mod \mathbb{P}$ or even $\mathcal{G} = \mathcal{F} \mod 0$, if for every $E \in \mathcal{F}$ there exists $E' \in \mathcal{G}$ s.t.

$$\mathbb{P}[E \Delta E'] = 0 .$$

- Partition $\mathcal{A} = \{A_i, i = 1, 2, \ldots\}$ measurable with respect to $\mathcal{F}$ is called generating if

$$\bigvee_{n=0}^{\infty} \sigma\{\tau^{-n}\mathcal{A}\} = \mathcal{F} \mod \mathbb{P} .$$

- Random variable $Y : \Omega \to \mathcal{Y}$ with a *countable* alphabet $\mathcal{Y}$ is called a generator of $(\Omega, \mathcal{F}, \mathbb{P}, \tau)$ if

$$\sigma\{Y, Y \circ \tau, \ldots, Y \circ \tau^n, \ldots\} = \mathcal{F} \mod \mathbb{P}$$

Theorem 8.4 (Sinai's generator theorem). *Let Y be the generator of a p.p.t. $(\Omega, \mathcal{F}, \mathbb{P}, \tau)$. Let $H(\mathbb{Y})$ be the entropy rate of the process $\mathbb{Y} = \{Y_k = Y \circ \tau^k, k = 0, \ldots\}$. If $H(\mathbb{Y})$ is finite, then $\mathcal{H}(\tau) = H(\mathbb{Y})$.*

Proof. Notice that since $H(\mathbb{Y})$ is finite, we must have $H(Y_0^n) < \infty$ and thus $H(Y) < \infty$. First, we argue that $\mathcal{H}(\tau) \ge H(\mathbb{Y})$. If Y has finite alphabet, then it is simply from the definition. Otherwise let Y be $\mathbb{Z}_+$-valued. Define a truncated version $\tilde{Y}_m = \min(Y, m)$, then since $\tilde{Y}_m \to Y$ as $m \to \infty$ we have from lower semicontinuity of mutual information, cf. (3.9), that

$$\lim_{m\to\infty} I(Y; \tilde{Y}_m) \ge H(Y) ,$$

and consequently for arbitrarily small ϵ and sufficiently large m

$$H(Y|\tilde{Y}) \le \epsilon ,$$

Then, consider the chain

$$\begin{aligned} H(Y_0^n) = H(\tilde{Y}_0^n, Y_0^n) &= H(\tilde{Y}_0^n) + H(Y_0^n|\tilde{Y}_0^n) \\ &= H(\tilde{Y}_0^n) + \sum_{i=0}^n H(Y_i|\tilde{Y}_0^n, Y_0^{i-1}) \\ &\le H(\tilde{Y}_0^n) + \sum_{i=0}^n H(Y_i|\tilde{Y}_i) \\ &= H(\tilde{Y}_0^n) + nH(Y|\tilde{Y}) \le H(\tilde{Y}_0^n) + n\epsilon \end{aligned}$$

Thus, entropy rate of $\tilde{\mathbb{Y}}$ (which has finite-alphabet) can be made arbitrarily close to the entropy rate of $\mathbb{Y}$, concluding that $\mathcal{H}(\tau) \ge \mathcal{H}(\mathbb{Y})$.

The main part is showing that for any stationary process $\mathbb{X}$ adapted to τ the entropy rate is upper bounded by $H(\mathbb{Y})$. To that end, consider $X : \Omega \to \mathcal{X}$ with finite $\mathcal{X}$ and define as usual the process $\mathbb{X} = \{X \circ \tau^k, k = 0, 1, \ldots\}$. By generating property of $\mathbb{Y}$ we have that X (perhaps after modification on a set of measure zero) is a function of Y_0^∞. So are all X_k. Thus

$$H(X_0) = I(X_0; Y_0^\infty) = \lim_{n\to\infty} I(X_0; Y_0^n),$$

where we used the continuity-in-σ-algebra property of mutual information, cf. (3.10). Rewriting the latter limit differently, we have

$$\lim_{n\to\infty} H(X_0|Y_0^n) = 0.$$

Fix $\epsilon > 0$ and choose m so that $H(X_0|Y_0^m) \le \epsilon$. Then consider the following chain:

$$\begin{aligned} H(X_0^n) \le H(X_0^n, Y_0^n) &= H(Y_0^n) + H(X_0^n|Y_0^n) \\ &\le H(Y_0^n) + \sum_{i=0}^n H(X_i|Y_i^n) \\ &= H(Y_0^n) + \sum_{i=0}^n H(X_0|Y_0^{n-i}) \\ &\le H(Y_0^n) + m \log|\mathcal{X}| + (n-m)\epsilon, \end{aligned}$$

where we used stationarity of (X_k, Y_k) and the fact that $H(X_0|Y_0^{n-i}) < \epsilon$ for $i \le n-m$. After dividing by n and passing to the limit our argument implies

$$H(\mathbb{X}) \le H(\mathbb{Y}) + \epsilon.$$

Taking here $\epsilon \to 0$ completes the proof.

Alternative proof: Suppose X_0 is taking values on a finite alphabet $\mathcal{X}$ and $X_0 = f(Y_0^\infty)$. Then (this is a measure-theoretic fact) for every $\epsilon > 0$ there exists $m = m(\epsilon)$ and a function $f_\epsilon : \mathcal{Y}^{m+1} \to \mathcal{X}$ s.t.

$$\mathbb{P}[f(Y_0^\infty) \ne f_\epsilon(Y_0^m)] \le \epsilon.$$

(This is just another way to say that $\bigcup_n \sigma\{Y_0^n\}$ is $\mathbb{P}$-dense in $\sigma(Y_0^\infty)$.) Define a stationary process $\tilde{\mathbb{X}}$ as

$$\tilde{X}_j \triangleq f_\epsilon(Y_j^{m+j}).$$

Notice that since $\tilde{X}_0^n$ is a function of Y_0^{n+m} we have

$$H(\tilde{X}_0^n) \le H(Y_0^{n+m}).$$

Dividing by m and passing to the limit we obtain that for entropy rates

$$H(\tilde{\mathbb{X}}) \le H(\mathbb{Y}) .$$

Finally, to relate $\tilde{\mathbb{X}}$ to $\mathbb{X}$ notice that by construction

$$\mathbb{P}[\tilde{X}_j \ne X_j] \le \epsilon .$$

Since both processes take values on a fixed finite alphabet, from Corollary 5.2 we infer that

$$|H(\mathbb{X}) - H(\tilde{\mathbb{X}})| \le \epsilon \log |\mathcal{X}| + h(\epsilon) .$$

Altogether, we have shown that

$$H(\mathbb{X}) \le H(\mathbb{Y}) + \epsilon \log |\mathcal{X}| + h(\epsilon) .$$

Taking $\epsilon \to 0$ we conclude the proof. □

Examples:

- Let $\Omega = [0, 1]$, $\mathcal{F}$–Borel σ-algebra, $\mathbb{P} = \text{Leb}$ and

$$\tau(\omega) = 2\omega \mod 1 = \begin{cases} 2\omega, & \omega < 1/2 \\ 2\omega - 1, & \omega \ge 1/2 \end{cases}$$

  It is easy to show that $Y(\omega) = 1\{\omega < 1/2\}$ is a generator and that $\mathbb{Y}$ is an i.i.d. Bernoulli(1/2) process. Thus, we get that Kolmogorov-Sinai entropy is $\mathcal{H}(\tau) = \log 2$.

- Let Ω be the unit circle $\mathbb{S}^1$, $\mathcal{F}$ – Borel σ-algebra, $\mathbb{P}$ be the normalized length and

$$\tau(\omega) = \omega + \gamma$$

  i.e. τ is a rotation by the angle γ. (When $\frac{\gamma}{2\pi}$ is irrational, this is known to be an ergodic p.p.t.). Here $Y = 1\{|\omega| < 2\pi\epsilon\}$ is a generator for arbitrarily small ϵ and hence

$$\mathcal{H}(\tau) \le H(\mathbb{X}) \le H(Y_0) = h(\epsilon) \to 0 \qquad \text{as } \epsilon \to 0 .$$

  This is an example of a zero-entropy p.p.t.

Remark 8.2. Two p.p.t.'s $(\Omega_1, \tau_1, \mathbb{P}_1)$ and $(\Omega_0, \tau_0, \mathbb{P}_0)$ are called isomorphic if there exists $f_i : \Omega_i \to \Omega_{1-i}$ defined $\mathbb{P}_i$-almost everywhere and such that 1) $\tau_{1-i} \circ f_i = f_{1-i} \circ \tau_i$; 2) $f_i \circ f_{1-i}$ is identity on Ω_i (a.e.); 3) $\mathbb{P}_i[f_{1-i}^{-1} E] = \mathbb{P}_{1-i}[E]$. It is easy to see that Kolmogorov-Sinai entropies of isomorphic p.p.t.s are equal. This observation was made by Kolmogorov in 1958. It was revoluationary, since it allowed to show that p.p.t.s corresponding shifts of iid Bern(1/2) and iid Bern(1/3) procceses are not isomorphic. Before, the only invariants known were those obtained from studying the spectrum of a unitary operator

$$U_\tau : L_2(\Omega, \mathbb{P}) \to L_2(\Omega, \mathbb{P}) \tag{8.16}$$

$$\phi(x) \mapsto \phi(\tau(x)) . \tag{8.17}$$

However, the spectrum of τ corresponding to any non-constant i.i.d. process consists of the entire unit circle, and thus is unable to distinguish Bern(1/2) from Bern(1/3).[2]

[2] To see the statement about the spectrum, let X_i be iid with zero mean and unit variance. Then consider $\phi(x_1^\infty)$ defined as $\frac{1}{\sqrt{m}} \sum_{k=1}^m e^{i\omega k} x_k$. This ϕ has unit energy and as $m \to \infty$ we have $\|U_\tau \phi - e^{i\omega}\phi\|_{L_2} \to 0$. Hence every $e^{i\omega}$ belongs to the spectrum of U_τ.

§ 9. Universal compression

In this lecture we will discuss how to produce compression schemes that do not require apriori knowledge of the distribution. Here, compressor is a map $\mathcal{X}^n \to \{0,1\}^*$. Now, however, there is no one fixed probability distribution P_{X^n} on $\mathcal{X}^n$. The plan for this lecture is as follows:

1. We will start by discussing the earliest example of a universal compression algorithm (of Fitingof). It does not talk about probability distributions at all. However, it turns out to be asymptotically optimal simulatenously for all i.i.d. distributions and with small modifications for all finite-order Markov chains.

2. Next class of universal compressors is based on assuming that a the true distribution P_{X^n} belongs to a given class. These methods proceed by choosing a good model distribution Q_{X^n} serving as the minimax approximation to each distribution in the class. The compression algorithm is designed to work for Q_{X^n} is made.

3. Finally, an entirely different idea are algorithms of Lempel-Ziv type. These automatically adapt to the distribution of the source, without any prior assumptions required.

Throughout this section instead of describing each compression algorithm, we will merely specify some distribution Q_{X^n} and apply one of the following constructions:

- Sort all x^n in the order of decreasing $Q_{X^n}(x^n)$ and assign values from $\{0,1\}^*$ as in Theorem 6.1, this compressor has lengths satisfying

$$\ell(f(x^n)) \le \log \frac{1}{Q_{X^n}(x^n)}\,.$$

- Set lengths to be

$$\ell(f(x^n)) \triangleq \lceil \log \frac{1}{Q_{X^n}(x^n)} \rceil$$

  and apply Kraft's inequality Theorem 6.5 to construct a prefix code.

- Use arithmetic coding (see next section).

The important conclusion is that in all these cases we have

$$\ell(f(x^n)) \le \log \frac{1}{Q_{X^n}(x^n)} + \text{const}\,,$$

and in this way we may and will always replace lengths with $\log \frac{1}{Q_{X^n}(x^n)}$. *In this way, the only job of a universal compression algorithm is to specify Q_{X^n}.*

Remark 9.1. Furthermore, if we only restrict attention to prefix codes, then any code $f : \mathcal{X}^n \to \{0,1\}^*$ defines a distribution $Q_{X^n}(x^n) = 2^{-\ell(f(x^n))}$ (we assume the code's tree is full). In this way, for prefix-free codes results on redundancy, stated in terms of optimizing the choice of Q_{X^n}, imply tight converses too. For one-shot codes without prefix constraints the optimal answers are slightly different, however. (For example, the optimal universal code for all i.i.d. sources satisfies $\mathbb{E}[\ell(f(X^n))] \approx H(X^n) + \frac{|\mathcal{X}|-3}{2}\log n$ in contrast with $\frac{|\mathcal{X}|-1}{2}\log n$ for prefix-free codes.)

9.1 Arithmetic coding

Constructing an encoder table from Q_{X^n} may require a lot of resources if n is large. Arithmetic coding provides a convenient workaround by allowing to output bits sequentially. *Notice that to do so, it requires that not only Q_{X^n} but also its marginalizations $Q_{X^1}, Q_{X^2}, \cdots$ be easily computable.* (This is not the case, for example, for Shtarkov distributions (9.8)-(9.9), which are not compatible for different n.)

Let us agree upon some ordering on the alphabet of $\mathcal{X}$ (e.g. $\mathsf{a} < \mathsf{b} < \cdots < \mathsf{z}$) and extend this order lexicographically to $\mathcal{X}^n$ (that is for $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, we say $x < y$ if $x_i < y_i$ for the first i such that $x_i \neq y_i$, e.g., $\mathsf{baba} < \mathsf{babb}$). Then let

$$F_n(x^n) = \sum_{y^n < x^n} Q_{X^n}(x^n)\,.$$

Associate to each x^n an interval $I_{x^n} = [F_n(x^n), F_n(x^n) + Q_{X^n}(x^n))$. These intervals are disjoint subintervals of $[0,1)$. Now encode

$$x^n \mapsto \text{largest dyadic interval contained in } I_{x^n}\,.$$

Recall that dyadic intervals are intervals of the type $[m2^{-k}, (m+1)2^{-k}]$ where a is an odd integer. Clearly each dyadic interval can be associated with a binary string in $\{0,1\}^*$. We set $f(x^n)$ to be that string. The resulting code is a prefix code satisfying

$$\ell(f(x^n)) \le \left\lceil \log_2 \frac{1}{Q_{X^n}(x^n)} \right\rceil + 1\,.$$

(This is an exercise.)

Observe that

$$F_n(x^n) = F_{n-1}(x^{n-1}) + Q_{X^{n-1}}(x^{n-1}) \sum_{y < x_n} Q_{X_n|X^{n-1}}(y|x^{n-1})$$

and thus $F_n(x^n)$ can be computed sequentially *if $Q_{X^{n-1}}$ and $Q_{X_n|X^{n-1}}$ are easy to compute.* This method is the method of choice in many modern compression algorithms because it allows to dynamically incorporate the learned information about the stream, in the form of updating $Q_{X_n|X^{n-1}}$ (e.g. if the algorithm detects that an executable file contains a long chunk of English text, it may temporarily switch to $Q_{X_n|X^{n-1}}$ modeling the English language).

9.2 Combinatorial construction of Fitingof

Fitingof suggested that a sequence $x^n \in \mathcal{X}^n$ should be prescribed information $\Phi_0(x^n)$ equal to the logarithm of the number of all possible permutations obtainable from x^n (i.e. log-size of the type-class containing x^n). From Stirling's approximation this can be shown to be

$$\Phi_0(x^n) = nH(x_T) + O(\log n) \quad T \sim \text{Unif}[n] \tag{9.1}$$

$$= nH(\hat{P}_{x^n}) + O(\log n)\,, \tag{9.2}$$

where $\hat{P}_{x^n}$ is the empirical distribution of the sequence x^n:

$$\hat{P}_{x^n}(a) \triangleq \frac{1}{n}\sum_{i=1}^{n} 1\{x_i = a\}\,. \tag{9.3}$$

Then Fitingof argues that it should be possible to produce a prefix code with

$$\ell(f(x^n)) = \Phi_0(x^n) + O(\log n)\,. \tag{9.4}$$

This can be done in many ways. In the spirit of what we will do next, let us define

$$Q_{X^n}(x^n) \triangleq \exp\{-\Phi_0(x^n)\}c_n\,,$$

where c_n is a normalization constant c_n. Counting the number of different possible empirical distributions (types), we get

$$c_n = O(n^{-(|\mathcal{X}|-1)})\,,$$

and thus, by Kraft inequality, there must exist a prefix code with lengths satisfying (9.4). Now taking expectation over $X^n \overset{\text{i.i.d.}}{\sim} P_X$ we get

$$\mathbb{E}[\ell(f(X^n))] = nH(P_X) + (|\mathcal{X}| - 1)\log n + O(1)\,,$$

for every i.i.d. source on $\mathcal{X}$.

9.2.1 Universal compressor for all finite-order Markov chains

Fitingof's idea can be extended as follows. Define now the 1-st order information content $\Phi_1(x^n)$ to be the log of the number of all sequences, obtainable by permuting x^n with extra restriction that the new sequence should have the same statistics on digrams. Asymptotically, Φ_1 is just the conditional entropy

$$\Phi_1(x^n) = nH(x_T | x_{T-1 \mod n}) + O(\log n), \quad T \sim \text{Unif}[n]\,.$$

Again, it can be shown that there exists a code such that lengths

$$\ell(f(x^n)) = \Phi_1(x^n) + O(\log n)\,.$$

This implies that for every 1-st order stationary Markov chain $X_1 \to X_2 \to \cdots \to X_n$ we have

$$\mathbb{E}[\ell(f(X^n))] = nH(X_2|X_1) + O(\log n)\,.$$

This can be further continued to define $\Phi_2(x^n)$ and build a universal code, asymptotically optimal for all 2-nd order Markov chains etc.

9.3 Optimal compressors for a class of sources. Redundancy.

So we have seen that we can construct compressor $f : \mathcal{X}^n \to \{0,1\}^*$ that achieves

$$\mathbb{E}[\ell(f(X^n))] \le H(X^n) + o(n)\,,$$

simultaneously for all i.i.d. sources (or even all r-th order Markov chains). What should we do next? Krichevsky suggested that the next barrier should be to optimize regret, or *redundancy*:

$$\mathbb{E}[\ell(f(X^n))] - H(X^n) \to \min$$

simultaneously for a class of sources. We proceed to rigorous definitions.

Given a collection $\{P_{X^n|\theta}, \theta \in \Theta\}$ of sources, and a compressor $f : \mathcal{X}^n \to \{0,1\}^*$ we define its redundancy as

$$\sup_{\theta_0} \mathbb{E}[\ell(f(X^n))|\theta = \theta_0] - H(X^n|\theta = \theta_0)\,.$$

Replacing here lengths with $\log \frac{1}{Q_{X^n}}$ we define redundancy of the distribution Q_{X^n} as

$$\sup_{\theta_0} D(P_{X^n|\theta=\theta_0} \| Q_{X^n})\,.$$

Thus, the question of designing the best universal compressor (in the sense of optimizing worst-case deviation of the average length from the entropy) becomes the question of finding solution of:

$$Q^*_{X^n} = \operatorname*{argmin}_{Q_{X^n}} \sup_{\theta_0} D(P_{X^n|\theta=\theta_0} \| Q_{X^n})\,.$$

We therefore get to the following definition

Definition 9.1 (Redundancy in universal compression)**.** Given a class of sources $\{P_{X^n|\theta=\theta_0}, \theta_0 \in \Theta, n = 1, \dots\}$ we define its minimax redundancy as

$$R^*_n \triangleq \min_{Q_{X^n}} \sup_{\theta_0} D(P_{X^n|\theta=\theta_0} \| Q_{X^n})\,. \tag{9.5}$$

Note that under condition of finiteness of R^*_n, Theorem 4.5 gives the maximin and capacity representation

$$R^*_n = \sup_{P_\theta} \min_{Q_{X^n}} D(P_{X^n|\theta} \| Q_{X^n} | P_\theta) \tag{9.6}$$

$$= \sup_{P_\theta} I(\theta; X^n)\,. \tag{9.7}$$

Thus redundancy is simply the capacity of the channel $\theta \to X^n$. This result, obvious in hindsight, was rather surprising in the early days of universal compression.

Finding exact Q_{X^n}-minimizer in (9.5) is a daunting task even for the simple class of all i.i.d. Bernoulli sources (i.e. $\Theta = [0,1]$, $P_{X^n|\theta} = \mathrm{Bern}^n(\theta)$). It turns out, however, that frequently the approximate minimizer has a rather nice structure: it matches the Jeffreys prior.

Remark 9.2. (Shtarkov and Fitingof) There is a connection between the combinatorial method of Fitingof and the method of optimality for a class. Indeed, following Shtarkov we may want to choose distribution $Q^{(S)}_{X^n}$ so as to minimize the worst-case redundancy *for each realization* x^n (not average!):

$$\min_{Q_{X^n}(x^n)} \sup_{\theta_0} \log \frac{P_{X^n|\theta}(x^n|\theta_0)}{Q_{X^n}(x^n)}$$

This leads to Shtarkov's distribution:

$$Q^{(S)}_{X^n}(x^n) = c \sup_{\theta_0} P_{X^n|\theta}(x^n|\theta_0)\,, \tag{9.8}$$

where c is the normalization constant. If class $\{P_{X^n|\theta}, \theta \in \Theta\}$ is chosen to be all i.i.d. distributions on $\mathcal{X}$ then

$$\text{i.i.d.} \quad Q^{(S)}_{X^n}(x^n) = c \exp\{-nH(\hat{P}_{x^n})\}\,, \tag{9.9}$$

and thus compressing w.r.t. $Q^{(S)}_{X^n}$ recovers Fitingof's construction Φ_0 up to $O(\log n)$ differences between $nH(\hat{P}_{x^n})$ and $\Phi_0(x^n)$. If we take $P_{X^n|\theta}$ to be all 1-st order Markov chains, then we get construction Φ_1 etc.

9.4* Approximate minimax solution: Jeffreys prior

In this section we will only consider the simple setting of a class of sources consisting of all i.i.d. distributions on a given finite alphabet. We will show that the prior, asymptoticall solving capacity question (9.7), is given by the Dirichlet-distribution with parameters set to 1/2, namely the pdf

$$P^*_\theta = \text{const}\frac{1}{\sqrt{\prod_{j=0}^d \theta_j}}\,.$$

First, we give the formal setting as follows:

- Fix $\mathcal{X}$ – finite alphabet of size $|\mathcal{X}| = d+1$, which we will enumerate as $\mathcal{X} = \{0, \ldots, d\}$.
- $\Theta = \{(\theta_j, j = 1, \ldots, d) : \sum_{j=1}^d \theta_j \le 1, \theta_j \ge 0\}$ – is the collection of all probability distributions on $\mathcal{X}$. Note that Θ is a d-dimensional simplex. We will also define

$$\theta_0 \triangleq 1 - \sum_{j=1}^d \theta_j\,.$$

- The source class is

$$P_{X^n|\theta}(x^n|\theta) \triangleq \prod_{j=1}^n \theta_{x_j} = \exp\left\{-n \sum_{a \in \mathcal{X}} \theta_a \log \frac{1}{\hat{P}_{x^n}(a)}\right\}\,,$$

  where as before $\hat{P}_{x^n}$ is the empirical distribution of x^n, cf. (9.3).

In order to derive the caod $Q^*_{X^n}$ we first propose a guess that the caid P_θ in (9.7) is some distribution with smooth density on Θ (this can only be justified by an apriori belief that the caid in such a natural problem should be something that employs all θ's). Then, we define

$$Q_{X^n}(x^n) \triangleq \int_\Theta P_{X^n|\theta}(x^n|\theta')P_\theta(\theta')d\theta'\,. \tag{9.10}$$

Before proceeding further, we recall the following method of approximating exponential integrals (called Laplace method). Suppose that $f(\theta)$ has a unique minimum at the interior point $\hat{\theta}$ of Θ

and that Hessian Hess$f$ is uniformly lower-bounded by a multiple of identity (in particular, $f(\theta)$ is strongly convex). Then taking Taylor expansion of π and f we get

$$\int_\Theta \pi(\theta)e^{-nf(\theta)}d\theta = \int (\pi(\hat\theta) + O(\|t\|))e^{-n(f(\hat\theta)-\frac{1}{2}t^T\mathrm{Hess}f(\hat\theta)t+o(\|t\|^2))}dt \tag{9.11}$$

$$= \pi(\hat\theta)e^{-nf(\hat\theta)}\int_{\mathbb{R}^d} e^{-x^T\mathrm{Hess}f(\hat\theta)x}\frac{dx}{\sqrt{n^d}}(1+O(n^{-1/2})) \tag{9.12}$$

$$= \pi(\hat\theta)e^{-nf(\hat\theta)}\left(\frac{2\pi}{n}\right)^{\frac{d}{2}}\frac{1}{\sqrt{\det \mathrm{Hess}f(\hat\theta)}}(1+O(n^{-1/2})) \tag{9.13}$$

where in the last step we computed Gaussian integral.

Next, we notice that

$$P_{X^n|\theta}(x^n|\theta') = e^{-n(D(\hat P_{x^n}\|P_{X|\theta=\theta'})+H(\hat P_{x^n}))\log e},$$

and therefore, denoting

$$\hat\theta(x^n) \triangleq \hat P_{x^n}$$

we get from applying (9.13) to (9.10)

$$\log Q_{X^n}(x^n) = -nH(\hat\theta) + \frac{d}{2}\log\frac{2\pi}{n\log e} + \log\frac{P_\theta(\hat\theta)}{\sqrt{\det J_F(\hat\theta)}} + O(n^{-\frac{1}{2}}),$$

where we used the fact that $\mathrm{Hess}_{\theta'}D(\hat P\|P_{X|\theta=\theta'}) = \frac{1}{\log e}J_F(\theta')$ with J_F – Fisher information matrix, see (4.13). From here, using the fact that under $X^n \sim P_{X^n|\theta=\theta'}$ the random variable $\hat\theta = \theta' + O(n^{-1/2})$ we get by linearizing $J_F(\cdot)$ and $P_\theta(\cdot)$

$$D(P_{X^n|\theta=\theta'}\|Q_{X^n}) = n(\mathbb{E}[H(\hat\theta)] - H(X|\theta=\theta')) + \frac{d}{2}\log n - \log\frac{P_\theta(\theta')}{\sqrt{\det J_F(\theta')}} + \mathrm{const} + O(n^{-\frac{1}{2}}), \tag{9.14}$$

where const is some constant (independent of prior P_θ or θ'). The first term is handled by the next Lemma.

Lemma 9.1. *Let $X^n \overset{i.i.d.}{\sim} P$ on finite alphabet $\mathcal{X}$ and let $\hat P$ be the empirical type of X^n then*

$$\mathbb{E}[D(\hat P\|P)] = \frac{|\mathcal{X}|-1}{2n}\log e + o(\frac{1}{n}).$$

Proof. Notice that $\sqrt{n}(\hat P - P)$ converges in distribution to $\mathcal{N}(0,\Sigma)$, where $\Sigma = \mathrm{diag}(P) - PP^T$, where P is an $|\mathcal{X}|$-by-1 column vector. Thus, computing second-order Taylor expansion of $D(\cdot\|P)$, cf. (4.15), we get the result. □

Continuing (9.14) we get in the end

$$D(P_{X^n|\theta=\theta'}\|Q_{X^n}) = \frac{d}{2}\log n - \log\frac{P_\theta(\theta')}{\sqrt{\det J_F(\theta')}} + \mathrm{const} + O(n^{-\frac{1}{2}}) \tag{9.15}$$

under the assumption of smoothness of prior P_θ and that θ' is not too close to the boundary. Consequently, we can see that in order for the prior P_θ be the saddle point solution, we should have

$$P_\theta(\theta') \sim \sqrt{\det J_F(\theta')},$$

provided that such density is normalizable. Prior proportional to square-root of the determinant of Fisher information matrix is known as *Jeffreys prior*. In our case, using the explicit expression for Fisher information (4.16) we get

$$P_\theta^* = \mathrm{Beta}(1/2, 1/2, \cdots, 1/2) = c_d \frac{1}{\sqrt{\prod_{j=0}^d \theta_j}}\,, \tag{9.16}$$

where c_d is the normalization constant. The corresponding redundancy is then

$$R_n^* = \frac{d}{2} \log \frac{n}{2\pi e} - \log c_d + o(1)\,. \tag{9.17}$$

Remark 9.3. In statistics Jeffreys prior is justified as being invariant to smooth reparametrization, as evidenced by (4.14). For example, in answering "will the sun rise tomorrow", Laplace proposed to estimate the probability by modeling sunrise as i.i.d. Bernoulli process with a uniform prior on $\theta \in [0,1]$. However, this is clearly not very logical, as one may equally well postulate uniformity of $\alpha = \theta^{10}$ or $\beta = \sqrt{\theta}$. Jeffreys prior $\theta \sim \frac{1}{\sqrt{\theta(1-\theta)}}$ is invariant to reparametrization in the sense that if one computed $\sqrt{\det J_F(\alpha)}$ under α-parametrization the result would be exactly the pushforward of the $\frac{1}{\sqrt{\theta(1-\theta)}}$ along the map $\theta \mapsto \theta^{10}$.

Making the arguments in this subsection rigorous is far from trivial, see [CB90, CB94] for details.

9.5 Sequential probability assignment: Krichevsky-Trofimov

From (9.16) it is not hard to derive the (asymptotically) optimal universal probability assignment Q_{X^n}. For simplicity we consider Bernoulli case, i.e. $d = 1$ and $\theta \in [0,1]$ is the 1-dimensional parameter. Then,[1]

$$P_\theta^* = \frac{1}{\pi\sqrt{\theta(1-\theta)}} \tag{9.18}$$

$$Q_{X^n}^*(x^n) = \frac{(2t_0 - 1)!! \cdot (2t_1 - 1)!!}{2^n n!}\,, \qquad t_a = \#\{j \le n : x_j = a\} \tag{9.19}$$

This assignment can now be used to create a universal compressor via one of the methods outlined in the beginning of this lecture. However, what is remarkable is that it has a very nice sequential interpretation (as does any assignment obtained via $Q_{X^n} = \int P_\theta P_{X^n|\theta}$ with P_θ not depending on n).

$$Q_{X_n|X^{n-1}}(1|x^{n-1}) = \frac{t_1 + \frac{1}{2}}{n}, \qquad t_1 = \#\{j \le n-1 : x_j = 1\} \tag{9.20}$$

$$Q_{X_n|X^{n-1}}(0|x^{n-1}) = \frac{t_0 + \frac{1}{2}}{n}, \qquad t_0 = \#\{j \le n-1 : x_j = 0\} \tag{9.21}$$

This is the famous "add 1/2" rule of Krichevsky and Trofimov. Note that this sequential assignment is very convenient for use in prediction as well as in implementing an arithmetic coder.

[1] This is obtained from identity $\int_0^1 \frac{\theta^a(1-\theta)^b}{\sqrt{\theta(1-\theta)}} d\theta = \pi \frac{1\cdot3\cdots(2a-1)\cdot1\cdot3\cdots(2b-1)}{2^{a+b}(a+b)!}$ for integer $a, b \ge 0$. This identity can be derived by change of variable $z = \frac{\theta}{1-\theta}$ and using the standard keyhole contour on the complex plain.

Remark 9.4. Notice that attaining the first order term $\frac{d}{2}\log n$ in (9.17) is easy. For example, taking Q_{X^n} to be the result of uniform P_θ does achieve this redundancy. In the Bernoulli ($d = 1$) case, the corresponding successive probability is given by

$$Q_{X_n|X^{n-1}}(1|x^{n-1}) = \frac{t_1+1}{n+1}\,, \quad t_1 = \#\{j \le n-1 : x_j = 1\}\,.$$

This is known as Laplace's "add 1" rule.

9.6 Lempel-Ziv compressor

So given a class of sources $\{P_{X^n|\theta}, \theta \in \Theta\}$ we have shown how to produce an asymptotically optimal compressors by using Jeffreys' prior. Although we have done so only for i.i.d. class, it can be extended to handle a class of all r-th order Markov chains with minimal modifications. However, the resulting sequential probability becomes rather complex. Can we do something easier at the expense of losing optimal redundancy?

In principle, the problem is rather straightforward: as we observe a stationary process, we may estimate with better and better precision the conditional probability $\hat{P}_{X_n|X_{n-r}^{n-1}}$ and then use it as the basis for arithmetic coding. As long as $\hat{P}$ converges to the actual conditional probability, we will get to the entropy rate of $H(X_n|X_{n-r}^{n-1})$. Note that Krichevsky-Trofimov assignment (9.21) is clearly learning the distribution too: as n grows, the estimator $Q_{X_n|X^{n-1}}$ converges to the true P_X (provided sequence is i.i.d.). So in some sense the converse is also true: *any good universal compression scheme is inherently learning the true distribution.*

The main drawback of the learn-then-compress approach is the following. Once we extend the class of sources to include those with memory, we invariably are lead to the problem of learning the joint distribution $P_{X_0^{r-1}}$ of r-blocks. However, the number of samples required to obtain a good estimate of $P_{X_0^{r-1}}$ is exponential in r. Thus learning may proceed rather slowly. Lempel-Ziv family of algorithms works around this in an ingeniously elegant way:

- First, estimating probabilities of rare substrings takes longest, but it is also the least useful, as these substrings almost never appear at the input.

- Second, *and most crucial*, observation is that a great estimate of the $P_{X^r}(x^r)$ is given by the reciprocal of the distance to the last observation of x^r in the incoming stream.

- Third, there is a prefix code[2] mapping any integer n to binary string of length roughly $\log_2 n$:

$$f_{int} : \mathbb{Z}_+ \to \{0,1\}^+, \qquad \ell(f_{int}(n)) = \log_2 n + O(\log\log n)\,. \tag{9.22}$$

  Thus, by encoding the pointer to the last observation of x^r via such a code we get a string of length roughly $\log P_{X^r}(x^r)$ automatically.

There are a number of variations of these basic ideas, so we will only attempt to give a rough explanation of why it works, without analyzing any particular algorithm.

We proceed to formal details. First, we need to establish a Kac's lemma.

[2] For this just notice that $\sum_{k\ge1} 2^{-\log_2 k - 2\log_2\log(k+1)} < \infty$ and use Kraft's inequality.

Lemma 9.2 (Kac)**.** *Consider a finite-alphabet stationary ergodic process* $\ldots, X_{-1}, X_0, X_1 \ldots$. *Let* $L = \inf\{t > 0 : X_{-t} = X_0\}$ *be the last appearance of symbol* X_0 *in the sequence* $X_{-\infty}^{-1}$. *Then for any* u *such that* $\mathbb{P}[X_0 = u] > 0$ *we have*

$$\mathbb{E}[L|X_0 = u] = \frac{1}{\mathbb{P}[X_0 = u]}\,.$$

In particular, mean recurrence time $\mathbb{E}[L] = |\mathrm{supp} P_X|$.

Proof. Note that from stationarity the following probability

$$\mathbb{P}[\exists t \geq k : X_t = u]$$

does not depend on $k \in \mathbb{Z}$. Thus by continuity of probability we can take $k = -\infty$ to get

$$\mathbb{P}[\exists t \geq 0 : X_t = u] = \mathbb{P}[\exists t \in \mathbb{Z} : X_t = u]\,.$$

However, the last event is shift-invariant and thus must have probability zero or one by ergodic assumption. But since $\mathbb{P}[X_0 = u] > 0$ it cannot be zero. So we conclude

$$\mathbb{P}[\exists t \geq 0 : X_t = u] = 1\,. \tag{9.23}$$

Next, we have

$$\begin{aligned}
\mathbb{E}[L|X_0 = u] &= \sum_{t\geq 1} \mathbb{P}[L \geq t | X_0 = u] && (9.24)\\
&= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t\geq 1} \mathbb{P}[L \geq t, X_0 = u] && (9.25)\\
&= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t\geq 1} \mathbb{P}[X_{-t+1} \neq u, \ldots, X_{-1} \neq u, X_0 = u] && (9.26)\\
&= \frac{1}{\mathbb{P}[X_0 = u]} \sum_{t\geq 1} \mathbb{P}[X_0 \neq u, \ldots, Xt - 2 \neq u, X_{t-1} = u] && (9.27)\\
&= \frac{1}{\mathbb{P}[X_0 = u]} \mathbb{P}[\exists t \geq 0 : X_t = u] && (9.28)\\
&= \frac{1}{\mathbb{P}[X_0 = u]}\,, && (9.29)
\end{aligned}$$

where (9.24) is the standard expression for the expectation of a $\mathbb{Z}_+$-valued random variable, (9.27) is from stationarity, (9.28) is because the events corresponding to different t are disjoint, and (9.29) is from (9.23). □

The following proposition serves to explain the basic principle behind operation of Lempel-Ziv:

Theorem 9.1. *Consider a finite-alphabet stationary ergodic process* $\ldots, X_{-1}, X_0, X_1 \ldots$ *with entropy rate* H. *Suppose that* $X_{-\infty}^{-1}$ *is known to the decoder. Then there exists a sequence of prefix-codes* $f_n(x_0^{n-1}, x_{-\infty}^{-1})$ *with expected length*

$$\frac{1}{n}\mathbb{E}[\ell(f_n(X_0^{n-1}, X_{\infty}^{-1}))] \to H\,,$$

Proof. Let L_n be the last occurence of the block x_0^{n-1} in the string $x_{-\infty}^{-1}$ (recall that the latter is known to decoder), namely

$$L_n = \inf\{t > 0 : x_{-t}^{-t+n-1} = x_0^{n-1}\}\,.$$

Then, by Kac's lemma applied to the process $Y_t^{(n)} = X_t^{t+n-1}$ we have

$$\mathbb{E}[L_n | X_0^{n-1} = x_0^{n-1}] = \frac{1}{\mathbb{P}[X_0^{n-1} = x_0^{n-1}]}\,.$$

We know encode L_n using the code (9.22). Note that there is crucial subtlety: even if $L_n < n$ and thus $[-t, -t+n-1]$ and $[0, n-1]$ overlap, the substring x_0^{n-1} can be decoded from the knowledge of L_n.

We have, by applying Jensen's inequality twice and noticing that $\frac{1}{n}H(X_0^{n-1}) \searrow H$ and $\frac{1}{n}\log H(X_0^{n-1}) \to 0$ that

$$\frac{1}{n}\mathbb{E}[\ell(f_{int}(L_n))] \le \frac{1}{n}\mathbb{E}[\log\frac{1}{P_{X_0^{n-1}}(X_0^{n-1})}] + o(1) \to H\,.$$

From Kraft's inequality we know that for any prefix code we must have

$$\frac{1}{n}\mathbb{E}[\ell(f_{int}(L_n))] \ge \frac{1}{n}H(X_0^{n-1} | X_{-\infty}^{-1}) = H\,.$$

□

Part III

Binary hypothesis testing

10.1 Binary Hypothesis Testing

Two possible distributions on a space $\mathcal{X}$

$$\begin{aligned} H_0 &: X \sim P \\ H_1 &: X \sim Q \end{aligned}$$

Where under hypothesis H_0 (the null hypothesis) X is distributed according to P, and under H_1 (the alternative hypothesis) X is distributed according to Q. A *test* between two distributions chooses either H_0 or H_1 based on an observation of X

- Deterministic test: $f : \mathcal{X} \to \{0, 1\}$
- Randomized test: $P_{Z|X} : \mathcal{X} \to \{0, 1\}$, so that $P_{Z|X}(0|x) \in [0, 1]$.

Let $Z = 0$ denote that the test chooses P, and $Z = 1$ when the test chooses Q.

Remark: This setting is called "testing simple hypothesis against simple hypothesis". Simple here refers to the fact that under each hypothesis there is only one distribution that could generate the data. Composite hypothesis is when $X \sim P$ and P is only known to belong to some class of distributions.

10.1.1 Performance Metrics

In order to determine the "effectiveness" of a test, we look at two metrics. Let $\pi_{i|j}$ denote the probability of the test choosing i when the correct hypothesis is j. With this

$$\begin{aligned} \alpha = \pi_{0|0} = P[Z = 0] & \quad \text{(Probability of success given } H_0 \text{ true)} \\ \beta = \pi_{0|1} = Q[Z = 0] & \quad \text{(Probability of error given } H_1 \text{ true)} \end{aligned}$$

Remark: $P[Z = 0]$ is a slight abuse of notation, more accurately $P[Z = 0] = \sum_{x \in \mathcal{X}} P(x) P_{Z|X}(0|x) = \mathbb{E}_{X \sim P_X}[1 - f(x)]$. Also, the choice of these two metrics to judge the test is not unique, we can use many other pairs from $\{\pi_{0|0}, \pi_{0|1}, \pi_{1|0}, \pi_{1|1}\}$.

So for any test $P_{Z|X}$ there is an associated (α, β). There are a few ways to determine the "best test"

- Bayesian: Assume prior distributions $\mathbb{P}[H_0] = \pi_0$ and $\mathbb{P}[H_1] = \pi_1$, minimize the expected error

$$P_b^* = \min_{\text{tests}} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}$$

- Minimax: Assume there is a prior distribution but it is unknown, so choose the test that preforms the best for the worst case priors

$$P_m^* = \min_{\text{tests}} \max_{\pi_0} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}$$

- Neyman-Pearson: Minimize error β subject to success probability at least α.

In this course, the Neyman-Pearson formulation will play a vital role.

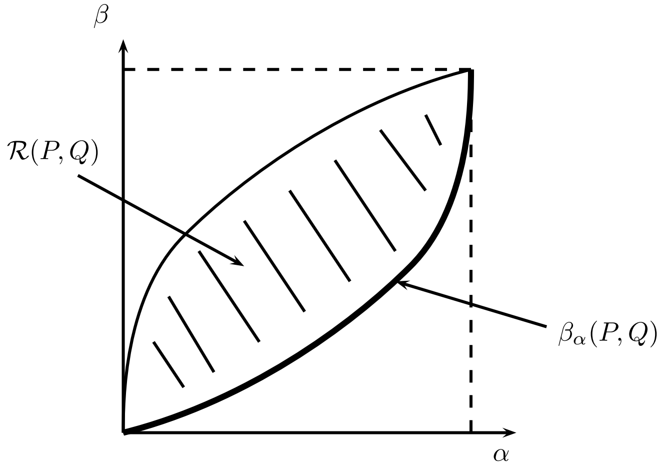
10.2 Neyman-Pearson formulation

Definition 10.1. Given that we require $P[Z = 0] \ge \alpha$,

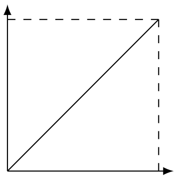
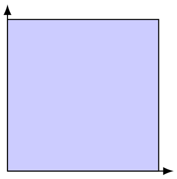
$$\beta_\alpha(P, Q) \triangleq \inf_{P[Z=0] \ge \alpha} Q[Z = 0]$$

Definition 10.2. Given (P, Q), the region of achievable points for all randomized tests is

$$\mathcal{R}(P, Q) = \bigcup_{P_{Z|X}} \{(P[Z = 0], Q[Z = 0])\} \subset [0, 1]^2 \tag{10.1}$$



Remark 10.1. This region encodes a lot of useful information about the relationship between P and Q. For example,[1]

$P = Q \Leftrightarrow \mathcal{R}(P, Q) =$  $P \perp Q \Leftrightarrow \mathcal{R}(P, Q) =$ 

Moreover, $\text{TV}(P, Q)$ = maximal length of vertical line intersecting the lower half of $\mathcal{R}(P, Q)$ (HW).

Theorem 10.1 (Properties of $\mathcal{R}(P, Q)$)**.**

1. *$\mathcal{R}(P, Q)$ is a closed, convex subset of $[0, 1]^2$.*
2. *$\mathcal{R}(P, Q)$ contains the diagonal.*

[1] Recall that P is mutually singular w.r.t. Q, denoted by $P \perp Q$, if $P[E] = 0$ and $Q[E] = 1$ for some E.

3. *Symmetry:* $(\alpha, \beta) \in \mathcal{R}(P, Q) \Leftrightarrow (1-\alpha, 1-\beta) \in \mathcal{R}(P, Q)$.

Proof. 1. For convexity, suppose $(\alpha_0, \beta_0), (\alpha_1, \beta_1) \in \mathcal{R}(P, Q)$, then each specifies a test $P_{Z_0|X}, P_{Z_1|X}$ respectively. Randomize between these two test to get the test $\lambda P_{Z_0|X} + \bar{\lambda} P_{Z_1|X}$ for $\lambda \in [0,1]$, which achieves the point $(\lambda\alpha_0 + \bar{\lambda}\alpha_1, \lambda\beta_0 + \bar{\lambda}\beta_1) \in \mathcal{R}(P, Q)$.

Closedness will follow from the explicit determination of all boundary points via Neyman-Pearson Lemma – see Remark 10.2. In more complicated situations (e.g. in testing against composite hypothesis) simple explicit solutions similar to Neyman-Pearson Lemma are not available but closedness of the region can frequently be argued still. The basic reason is that the collection of functions $\{g : \mathcal{X} \to [0,1]\}$ forms a weakly-compact set and hence its image under a linear functional $g \mapsto (\int g dP, \int g dQ)$ is closed.

2. Test by blindly flipping a coin, i.e., let $Z \sim \text{Bern}(1-\alpha) \perp\!\!\!\perp X$. This achieves the point (α, α).

3. If $(\alpha, \beta) \in \mathcal{R}(P, Q)$, then form the test that chooses P whenever $P_{Z|X}$ choses Q, and chooses Q whenever $P_{Z|X}$ choses P, which gives $(1-\alpha, 1-\beta) \in \mathcal{R}(P, Q)$.

□

The region $\mathcal{R}(P, Q)$ consists of the operating points of all randomized tests, which include deterministic tests as special cases. The achievable region of deterministic tests are denoted by

$$\mathcal{R}_{\text{det}}(P, Q) = \bigcup_E \{(P(E), Q(E)\}. \tag{10.2}$$

One might wonder the relationship between these two regions. It turns out that $\mathcal{R}(P, Q)$ is given by the closed convex hull of $\mathcal{R}_{\text{det}}(P, Q)$.

We first recall a couple of notations:

- Closure: $\mathbf{cl}(E) \triangleq$ the smallest closed set containing E.
- Convex hull: $\mathbf{co}(E) \triangleq$ the smallest convex set containing $E = \{\sum_{i=1}^n \alpha_i x_i : \alpha_i \geq 0, \sum_{i=1}^n \alpha_i = 1, x_i \in E, n \in \mathbb{N}\}$. A useful example: if $(f(x), g(x)) \in E, \forall x$, then $(\mathbb{E}[f(X)], \mathbb{E}[g(X)]) \in \mathbf{cl}(\mathbf{co}(E))$.

Theorem 10.2 (Randomized test v.s. deterministic tests)**.**

$$\mathcal{R}(P, Q) = \mathbf{cl}(\mathbf{co}(\mathcal{R}_{\text{det}}(P, Q))).$$

Consequently, if P and Q are on a finite alphabet $\mathcal{X}$, then $\mathcal{R}(P, Q)$ is a polygon of at most $2^{|\mathcal{X}|}$ vertices.

Proof. "$\supset$": Comparing (10.1) and (10.2), by definition, $\mathcal{R}(P, Q) \supset \mathcal{R}_{\text{det}}(P, Q))$. By Theorem 10.1, $\mathcal{R}(P, Q)$ is closed convex, and we are done with the $\supset$ direction.

"$\subset$": Given any randomized test $P_{Z|X}$, put $g(x) = P_{Z=0|X=x}$. Then g is a measurable function. Moreover,

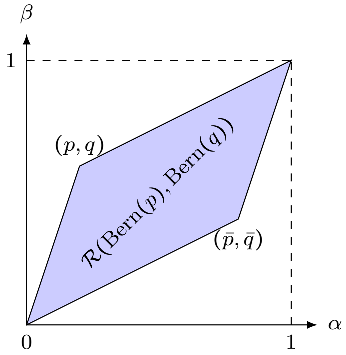
$$P[Z=0] = \sum_x g(x)P(x) = \mathbb{E}_P[g(X)] = \int_0^1 P[g(X) \geq t] \mathrm{d}t$$

$$Q[Z=0] = \sum_x g(x)Q(x) = \mathbb{E}_Q[g(X)] = \int_0^1 Q[g(X) \geq t] \mathrm{d}t$$

where we applied the formula $E[U] = \int \mathbb{P}[U \geq t]\,\mathrm{d}t$ for $U \geq 0$. Therefore the point $(P[Z=0], Q[Z=0]) \in \mathcal{R}$ is a mixture of points $(P[g(X) \geq t], Q[g(X) \geq t]) \in \mathcal{R}_{\mathrm{det}}$, averaged according to t uniformly distributed on the unit interval. Hence $\mathcal{R} \subset \mathbf{cl}(\mathbf{co}(\mathcal{R}_{\mathrm{det}}))$.

The last claim follows because there are at most $2^{|\mathcal{X}|}$ subsets in (10.2). □

Example: Testing Bern(p) versus Bern(q), $p < \frac{1}{2} < q$. Using Theorem 10.2, note that there are $2^2 = 4$ events $E = \varnothing, \{0\}, \{1\}, \{0,1\}$. Then



10.3 Likelihood ratio tests

Definition 10.3. The log likelihood ratio (LLR) is $F = \log\frac{\mathrm{d}P}{\mathrm{d}Q} : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\}$. The likelihood ratio test (LRT) with threshold $\tau \in \mathbb{R}$ is $\mathbf{1}\{\log\frac{\mathrm{d}P}{\mathrm{d}Q} \leq \tau\}$. Formally, we assume that $dP = p(x)d\mu$ and $dQ = q(x)d\mu$ (one can take $\mu = P + Q$, for example) and set

$$F(x) \triangleq \begin{cases} \log\frac{p(x)}{q(x)}, & p(x) > 0, q(x) > 0 \\ +\infty, & p(x) > 0, q(x) = 0 \\ -\infty, & p(x) = 0, q(x) > 0 \\ n/a, & p(x) = 0, q(x) = 0 \end{cases}$$

Notes:

- LRT is a deterministic test. The intuition is that upon observing x, if $\frac{Q(x)}{P(x)}$ exceeds a certain threshold, suggesting Q is more likely, one should reject the null hypothesis and declare Q.
- The rationale for defining extended values $\pm\infty$ of $F(x)$ are the following observations:

$$\forall x, \forall \tau \in \mathbb{R}: \quad \begin{aligned} &(p(x) - \exp\{\tau\}q(x))1\{F(x) > \tau\} \geq 0 \\ &(p(x) - \exp\{\tau\}q(x))1\{F(x) \geq \tau\} \geq 0 \\ &(q(x) - \exp\{-\tau\}p(x))1\{F(x) < \tau\} \geq 0 \\ &(q(x) - \exp\{-\tau\}p(x))1\{F(x) \leq \tau\} \geq 0 \end{aligned}$$

  This leads to the following useful consequence: For any $g \geq 0$ and any $\tau \in \mathbb{R}$ (note: $\tau = \pm\infty$ is excluded) we have

$$\mathbb{E}_P[g(X)1\{F \geq \tau\}] \geq \exp\{\tau\} \cdot \mathbb{E}_Q[g(X)1\{F \geq \tau\}] \tag{10.3}$$
$$\mathbb{E}_Q[g(X)1\{F \leq \tau\}] \geq \exp\{-\tau\} \cdot \mathbb{E}_P[g(X)1\{F \leq \tau\}] \tag{10.4}$$

Below, these and similar inequalities are only checked for the cases of F not taking extended values, but from this remark it should be clear how to treat the general case.

- Another useful observation:

$$Q[F = +\infty] = P[F = -\infty] = 0\,. \tag{10.5}$$

Theorem 10.3.

1. *F is a sufficient statistic for testing H_0 vs H_1.*
2. *For discrete alphabet $\mathcal{X}$ and when $Q \ll P$ we have*

$$Q[F = f] = \exp(-f)P[F = f] \qquad \forall f \in \mathbb{R} \cup \{+\infty\}$$

More generally, we have for any $g : \mathbb{R} \cup \{\pm\infty\} \to \mathbb{R}$

$$\mathbb{E}_Q[g(F)] = g(-\infty)Q[F = -\infty] + \mathbb{E}_P[\exp\{-F\}g(F)] \tag{10.6}$$

$$\mathbb{E}_P[g(F)] = g(+\infty)P[F = +\infty] + \mathbb{E}_Q[\exp\{F\}g(F)] \tag{10.7}$$

Proof. **(2)**

$$\begin{aligned} Q_F(f) &= \sum_{\mathcal{X}} Q(x)\mathbf{1}\{\log \frac{P(x)}{Q(x)} = f\} = \sum_{\mathcal{X}} Q(x)\mathbf{1}\{e^f Q(x) = P(x)\} \\ &= e^{-f}\sum_{\mathcal{X}} P(x)\mathbf{1}\{\log \frac{P(x)}{Q(x)} = f\} = e^{-f}P_F(f) \end{aligned}$$

To prove the general version (10.6), note that

$$\begin{aligned} \mathbb{E}_Q[g(F)] &= \int_{\{-\infty < F(x) < \infty\}} d\mu\; q(x)g(F(x)) + g(-\infty)Q[F = -\infty] && (10.8) \\ &= \int_{\{-\infty < F(x) < \infty\}} d\mu\; p(x)\exp\{-F(x)\}g(F(x)) + g(-\infty)Q[F = -\infty] && (10.9) \\ &= \mathbb{E}_P[\exp\{-F\}g(F)] + g(-\infty)Q[F = -\infty]\,, && (10.10) \end{aligned}$$

where we used (10.5) to justify restriction to finite values of F.

(1) To show F is a s.s, we need to show $P_{X|F} = Q_{X|F}$. For the discrete case we have:

$$\begin{aligned} P_{X|F}(x|f) &= \frac{P_X(x)P_{F|X}(f|x)}{P_F(f)} = \frac{P(x)\mathbf{1}\{\frac{P(x)}{Q(x)} = e^f\}}{P_F(f)} = \frac{e^f Q(x)\mathbf{1}\{\frac{P(x)}{Q(x)} = e^f\}}{P_F(f)} \\ &= \frac{Q_{XF}(xf)}{e^{-f}P_F(f)} \overset{\textbf{(2)}}{=} \frac{Q_{XF}}{Q_F} = Q_{X|F}(x|f). \end{aligned} \qquad \square$$

The general argument is done similarly to the proof of (10.6).

From Theorem 10.2 we know that to obtain the achievable region $\mathcal{R}(P,Q)$, one can iterate over all subsets and compute the region $\mathcal{R}_{\text{det}}(P,Q)$ first, then take its closed convex hull. But this is a formidable task if the alphabet is huge or infinite. But we know that the LLR $\log\frac{\mathrm{d}P}{\mathrm{d}Q}$ is a sufficient statistic. Next we give bounds to the region $\mathcal{R}(P,Q)$ in terms of the statistics of $\log\frac{\mathrm{d}P}{\mathrm{d}Q}$. As usual, there are two types of statements:

- Converse (outer bounds): any point in $\mathcal{R}(P,Q)$ must satisfy ...
- Achievability (inner bounds): the following point belong to $\mathcal{R}(P,Q)$...

10.4 Converse bounds on $\mathcal{R}(P,Q)$

Theorem 10.4 (Weak Converse). *$\forall(\alpha,\beta)\in\mathcal{R}(P,Q)$,*

$$d(\alpha\|\beta)\le D(P\|Q)$$
$$d(\beta\|\alpha)\le D(Q\|P)$$

where $d(\cdot\|\cdot)$ is the binary divergence.

Proof. Use data processing with $P_{Z|X}$. □

Lemma 10.1 (Deterministic tests). *$\forall E,\ \forall\gamma>0:\ P[E]-\gamma Q[E]\le P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma\Big]$*

Proof. (Discrete version)

$$\begin{aligned}P[E]-\gamma Q[E]&=\sum_{x\in E}p(x)-\gamma q(x)\le\sum_{x\in E}(p(x)-\gamma q(x))\mathbf{1}_{\{p(x)>\gamma q(x)\}}\\&=P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma,X\in E\Big]-\gamma Q\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma,X\in E\Big]\le P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma\Big].\end{aligned}$$

(General version) WLOG, suppose $P,Q\ll\mu$ for some measure μ (since we can always take $\mu=P+Q$). Then $\mathrm{d}P=p(x)\mathrm{d}\mu$, $\mathrm{d}Q=q(x)\mathrm{d}\mu$. Then

$$\begin{aligned}P[E]-\gamma Q[E]&=\int_E\mathrm{d}\mu(p(x)-\gamma q(x))\le\int_E\mathrm{d}\mu(p(x)-\gamma q(x))\mathbf{1}_{\{p(x)>\gamma q(x)\}}\\&=P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma,X\in E\Big]-Q\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma,X\in E\Big]\le P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma\Big].\end{aligned}$$

where the second line follows from $\frac{p}{q}=\frac{\frac{\mathrm{d}P}{\mathrm{d}\mu}}{\frac{\mathrm{d}Q}{\mathrm{d}\mu}}=\frac{\mathrm{d}P}{\mathrm{d}Q}$.

[So we see that the only difference between the discrete and the general case is that the counting measure is replaced by some other measure μ.] □

Note: In this case, we do not need $P\ll Q$, since $\pm\infty$ is a reasonable and meaningful value for the log likelihood ratio.

Lemma 10.2 (Randomized tests). *$P[Z=0]-\gamma Q[Z=0]\le P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma\Big]$.*

Proof. Almost identical to the proof of the previous Lemma 10.1:

$$\begin{aligned}P[Z=0]-\gamma Q[Z=0]&=\sum_x P_{Z|X}(0|x)(p(x)-\gamma q(x))\le\sum_x P_{Z|X}(0|x)(p(x)-\gamma q(x))\mathbf{1}_{\{p(x)>\gamma q(x)\}}\\&=P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma,Z=0\Big]-Q\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma,Z=0\Big]\\&\le P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma\Big].\end{aligned}$$

□

Theorem 10.5 (Strong Converse). *$\forall(\alpha,\beta)\in\mathcal{R}(P,Q),\forall\gamma>0$,*

$$\alpha-\gamma\beta\le P\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}>\log\gamma\Big]\tag{10.11}$$

$$\beta-\frac{1}{\gamma}\alpha\le Q\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q}<\log\gamma\Big]\tag{10.12}$$

Proof. Apply Lemma 10.2 to (P, Q, γ) and $(Q, P, 1/\gamma)$. □

Note: Theorem 10.5 provides an outer bound for the region $\mathcal{R}(P,Q)$ in terms of half-spaces. To see this, suppose one fixes $\gamma > 0$ and looks at the line $\alpha - \gamma\beta = c$ and slowing increases c from zero, there is going to be a maximal c, say c^*, at which point the line touches the lower boundary of the region. Then (10.11) says that c^* cannot exceed $P[\log \frac{dP}{dQ} > \log\gamma]$. Hence $\mathcal{R}$ must lie to the left of the line. Similarly, (10.12) provides bounds for the upper boundary. Altogether Theorem 10.5 states that $\mathcal{R}(P,Q)$ is contained in the intersection of a collection of half-spaces indexed by γ.
Note: To apply the strong converse Theorem 10.5, we need to know the CDF of the LLR, whereas to apply the weak converse Theorem 10.4 we need only to know the expectation of the LLR, i.e., divergence.

10.5 Achievability bounds on $\mathcal{R}(P,Q)$

Since we know that the set $\mathcal{R}(P,Q)$ is convex, it is natural to try to find all of its supporting lines (hyperplanes), as it is well known that closed convex set equals the intersection of the halfspaces correposponding to all supporting hyperplanes. So thus, we are naturally lead to solving the problem

$$\max\{\alpha - t\beta : (\alpha,\beta) \in \mathcal{R}(P,Q)\}\,.$$

This can be done rather simply:

$$\alpha^* - t\beta^* = \max_{(\alpha,\beta)\in\mathcal{R}} (\alpha - t\beta) = \max_{P_{Z|X}} \sum_{x\in\mathcal{X}} (P(x) - tQ(x)) P_{Z|X}(0|x) = \sum_{x\in\mathcal{X}} |P(x) - tQ(x)|^+$$

where the last equality follows from the fact that we are free to choose $P_{Z|X}(0|x)$, and the best choice is obvious:

$$P_{Z|X}(0|x) = 1\left\{\log\frac{P(x)}{Q(x)} \geq \log t\right\}\,.$$

Thus, we have shown that all supporting hyperplanes are parameterized by LLR-tests. This completely recovers the region $\mathcal{R}(P,Q)$ except for the points corresponding to the faces (linear pieces) of the region. To be precise, we state the following result.

Theorem 10.6 (Neyman-Pearson Lemma). *"LRT is optimal": For any α, β_α is attained by the following test:*

$$P_{Z|X}(0|x) = \begin{cases} 1 & \log\frac{dP}{dQ} > \tau \\ \lambda & \log\frac{dP}{dQ} = \tau \\ 0 & \log\frac{dP}{dQ} < \tau \end{cases} \tag{10.13}$$

where $\tau \in \mathbb{R}$ and $\lambda \in [0,1]$ are the unique solutions to $\alpha = P[\log\frac{dP}{dQ} > \tau] + \lambda P[\log\frac{dP}{dQ} = \tau]$.

Proof of Theorem 10.6. Let $t = \exp(\tau)$. Given any test $P_{Z|X}$, let $g(x) = P_{Z|X}(0|x) \in [0,1]$. We want to show that

$$\alpha = P[Z=0] = \mathbb{E}_P[g(X)] = P\Big[\frac{dP}{dQ} > t\Big] + \lambda P\Big[\frac{dP}{dQ} = t\Big] \tag{10.14}$$

$$\Rightarrow \beta = Q[Z=0] = \mathbb{E}_Q[g(X)] \overset{\text{goal}}{\geq} Q\Big[\frac{dP}{dQ} > t\Big] + \lambda Q\Big[\frac{dP}{dQ} = t\Big] \tag{10.15}$$

Using the simple fact that $\mathbb{E}_Q[f(X)\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}\leq t\right\}}] \geq t^{-1}\mathbb{E}_P[f(X)\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}\leq t\right\}}]$ for any $f \geq 0$ twice, we have

$$\begin{aligned}
\beta &= \mathbb{E}_Q[g(X)\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}\leq t\right\}}] + \mathbb{E}_Q[g(X)\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}> t\right\}}] \\
&\geq \frac{1}{t}\underbrace{\mathbb{E}_P[g(X)\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}\leq t\right\}}]} + \mathbb{E}_Q[g(X)\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}> t\right\}}] \\
&\overset{(10.14)}{=} \frac{1}{t}\underbrace{\Big(\mathbb{E}_P[(1-g(X))\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}> t\right\}}] + \lambda P\Big[\frac{\mathrm{d}P}{\mathrm{d}Q} = t\Big]\Big)} + \mathbb{E}_Q[g(X)\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}> t\right\}}] \\
&\geq \mathbb{E}_Q[(1-g(X))\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}> t\right\}}] + \lambda Q\Big[\frac{\mathrm{d}P}{\mathrm{d}Q} = t\Big] + \mathbb{E}_Q[g(X)\mathbf{1}_{\left\{\frac{\mathrm{d}P}{\mathrm{d}Q}> t\right\}}] \\
&= Q\Big[\frac{\mathrm{d}P}{\mathrm{d}Q} > t\Big] + \lambda Q\Big[\frac{\mathrm{d}P}{\mathrm{d}Q} = t\Big].
\end{aligned}$$

□

Remark 10.2. As a consequence of the Neyman-Pearson lemma, all the points on the boundary of the region $\mathcal{R}(P,Q)$ are attainable. Therefore

$$\mathcal{R}(P,Q) = \{(\alpha,\beta) : \beta_\alpha \leq \beta \leq 1 - \beta_{1-\alpha}\}.$$

Since $\alpha \mapsto \beta_\alpha$ is convex on $[0,1]$, hence continuous, the region $\mathcal{R}(P,Q)$ is a closed convex set. Consequently, the infimum in the definition of β_α is in fact a minimum.

Furthermore, the lower half of the region $\mathcal{R}(P,Q)$ is the convex hull of the union of the following two sets:

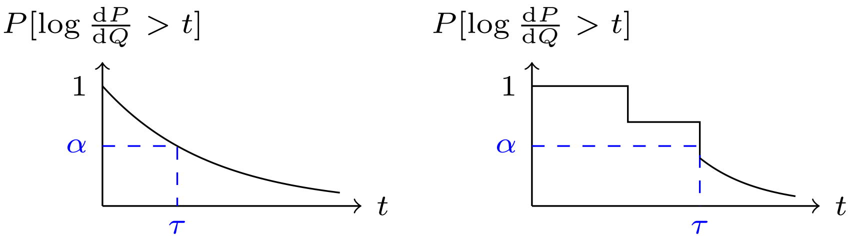
$$\begin{cases} \alpha = P\big[\log\frac{\mathrm{d}P}{\mathrm{d}Q} > \tau\big] \\ \beta = Q\big[\log\frac{\mathrm{d}P}{\mathrm{d}Q} > \tau\big] \end{cases} \qquad \tau \in \mathbb{R} \cup \{\pm\infty\}.$$

and

$$\begin{cases} \alpha = P\big[\log\frac{\mathrm{d}P}{\mathrm{d}Q} \geq \tau\big] \\ \beta = Q\big[\log\frac{\mathrm{d}P}{\mathrm{d}Q} \geq \tau\big] \end{cases} \qquad \tau \in \mathbb{R} \cup \{\pm\infty\}.$$

Therefore it does not lose optimality to restrict our attention on tests of the form $\mathbf{1}\{\log\frac{\mathrm{d}P}{\mathrm{d}Q} \geq \tau\}$ or $\mathbf{1}\{\log\frac{\mathrm{d}P}{\mathrm{d}Q} > \tau\}$.

Remark 10.3. The test (10.13) is related to LRT[2] as follows:



1. Left figure: If $\alpha = P[\log\frac{\mathrm{d}P}{\mathrm{d}Q} > \tau]$ for some τ, then $\lambda = 0$, and (10.13) becomes the LRT $Z = \mathbf{1}_{\left\{\log\frac{\mathrm{d}P}{\mathrm{d}Q}\leq\tau\right\}}$.
2. Right figure: If $\alpha \neq P[\log\frac{\mathrm{d}P}{\mathrm{d}Q} > \tau]$ for any τ, then we have $\lambda \in (0,1)$, and (10.13) is equivalent to randomize over tests: $Z = \mathbf{1}_{\left\{\log\frac{\mathrm{d}P}{\mathrm{d}Q}\leq\tau\right\}}$ with probability $\bar{\lambda}$ or $\mathbf{1}_{\left\{\log\frac{\mathrm{d}P}{\mathrm{d}Q}<\tau\right\}}$ with probability λ.

[2] Note that it so happens that in Definition 10.3 the LRT is defined with an $\leq$ instead of $<$.

Corollary 10.1. *$\forall \tau \in \mathbb{R}$, there exists $(\alpha, \beta) \in \mathcal{R}(P, Q)$ s.t.*

$$\alpha = P\Big[\log \frac{\mathrm{d}P}{\mathrm{d}Q} > \tau\Big]$$
$$\beta \le \exp(-\tau) P\Big[\log \frac{\mathrm{d}P}{\mathrm{d}Q} > \tau\Big] \le \exp(-\tau)$$

Proof.

$$\begin{aligned} Q\Big[\log \frac{\mathrm{d}P}{\mathrm{d}Q} > \tau\Big] &= \sum Q(x) \mathbf{1}\Big\{\frac{P(x)}{Q(x)} > e^{\tau}\Big\} \\ &\le \sum P(x) e^{-\tau} \mathbf{1}\Big\{\frac{P(x)}{Q(x)} > e^{\tau}\Big\} = e^{-\tau} P\Big[\log \frac{\mathrm{d}P}{\mathrm{d}Q} > \tau\Big]. \end{aligned}$$ □

10.6 Asymptotics

Now we have many samples from the underlying distribution

$$H_0: X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$$
$$H_1: X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} Q$$

We're interested in the asymptotics of the error probabilities $\pi_{0|1}$ and $\pi_{1|0}$. There are two main types of tests, both which the convergence rate to zero error is exponential.

1. Stein Regime: What is the best exponential rate of convergence for $\pi_{0|1}$ when $\pi_{1|0}$ has to be $\le \epsilon$?

$$\begin{cases} \pi_{1|0} \le \epsilon \\ \pi_{0|1} \to 0 \end{cases}$$

2. Chernoff Regime: What is the trade off between exponents of the convergence rates of $\pi_{1|0}$ and $\pi_{0|1}$ when we want both errors to go to 0?

$$\begin{cases} \pi_{1|0} \to 0 \\ \pi_{0|1} \to 0 \end{cases}$$

§ 11. Hypothesis testing asymptotics I

Setup:

$$
\begin{aligned}
&H_0 : X^n \sim P_{X^n} \qquad H_1 : X^n \sim Q_{X^n}\\
&\text{test } P_{Z|X^n} : \mathcal{X}^n \to \{0,1\}\\
&\text{specification } 1-\alpha = \pi_{1|0} \qquad \beta = \pi_{0|1}
\end{aligned}
$$

11.1 Stein's regime

$$
\begin{aligned}
&1-\alpha = \pi_{1|0} \le \epsilon\\
&\beta = \pi_{0|1} \to 0 \quad \text{at the rate } 2^{-nV_\epsilon}
\end{aligned}
$$

Note: interpretation of this specification, usually a "miss"(0|1) is much worse than a "false alarm" (1|0).

Definition 11.1 (ϵ-optimal exponent)**.** V_ϵ is called an *ϵ-optimal exponent in Stein's regime* if

$$
\begin{aligned}
&V_\epsilon = \sup\{E : \exists n_0, \forall n \ge n_0, \exists P_{Z|X^n} \text{ s.t. } \alpha > 1-\epsilon, \beta < 2^{-nE}, \}\\
\Leftrightarrow\ &V_\epsilon = \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_{1-\epsilon}(P_{X^n}, Q_{X^n})}
\end{aligned}
$$

where $\beta_\alpha(P,Q) = \min_{P_{Z|X}, P(Z=0)\ge\alpha} Q(Z=0)$.

Exercise: Check the equivalence.

Definition 11.2 (Stein's exponent)**.**

$$V = \lim_{\epsilon\to 0} V_\epsilon.$$

Theorem 11.1 (Stein's lemma)**.** *Let $P_{X^n} = P_X^n$ i.i.d. and $Q_{X^n} = Q_X^n$ i.i.d. Then*

$$V_\epsilon = D(P\|Q), \quad \forall \epsilon \in (0,1).$$

Consequently,

$$V = D(P\|Q).$$

Example: If it is required that $\alpha \ge 1-10^{-3}$, and $\beta \le 10^{-40}$, what's the number of samples needed? Stein's lemma provides a rule of thumb: $n \gtrsim -\frac{\log 10^{-40}}{D(P\|Q)}$.

Proof. Denote $F = \log \frac{dP}{dQ}$, and $F_n = \log \frac{dP_{X^n}}{dQ_{X^n}} = \sum_{i=1}^n \log \frac{dP}{dQ}(X_i)$ – iid sum.

Recall Neyman Pearson's lemma on optimal tests (likelihood ratio test): $\forall \tau$,

$$\alpha = P(F > \tau), \quad \beta = Q(F > \tau) \le e^{-\tau}$$

Also notice that by WLLN, under P, as $n \to \infty$,

$$\frac{1}{n} F_n = \frac{1}{n} \sum_{i=1}^n \log \frac{dP(X_i)}{dQ(X_i)} \xrightarrow{\mathbb{P}} \mathbb{E}_P \left[\log \frac{dP}{dQ} \right] = D(P\|Q). \tag{11.1}$$

Alternatively, under Q, we have

$$\frac{1}{n} F_n \xrightarrow{\mathbb{P}} \mathbb{E}_Q [\log \frac{dP}{dQ}] = -D(Q\|P) \tag{11.2}$$

1. Show $V_\epsilon \ge D(P\|Q) = D$.
   Pick $\tau = n(D - \delta)$, for some small $\delta > 0$. Then the optimal test achieves:

   $$\alpha = P(F_n > n(D - \delta)) \to 1, \text{ by (11.1)}$$
   $$\beta \le e^{-n(D-\delta)}$$

   then pick n large enough (depends on ϵ, δ) such that $\alpha \ge 1 - \epsilon$, we have the exponent $E = D - \delta$ achievable, $V_\epsilon \ge E$. Further let $\delta \to 0$, we have that $V_\epsilon \ge D$.

2. Show $V_\epsilon \le D(P\|Q) = D$.

   a) (weak converse) $\forall (\alpha, \beta) \in \mathcal{R}(P_{X^n}, Q_{X^n})$, we have

   $$-h(\alpha) + \alpha \log \frac{1}{\beta} \le d(\alpha\|\beta) \le D(P_{X^n}\|Q_{X^n}) \tag{11.3}$$

   where the first inequality is due to

   $$d(\alpha\|\beta) = \alpha \log \frac{\alpha}{\beta} + \bar{\alpha} \log \frac{\bar{\alpha}}{\bar{\beta}} = -h(\alpha) + \alpha \log \frac{1}{\beta} + \underbrace{\bar{\alpha} \log \frac{1}{\bar{\beta}}}_{\ge 0 \text{ and } \approx 0 \text{ for small } \beta}$$

   and the second is due to the weak converse Theorem 10.4 proved in the last lecture (data processing inequality for divergence).
   $\forall$ achievable exponent $E < V_\epsilon$, by definition, there exists a sequence of tests $P_{Z|X^n}$ such that $\alpha_n \ge 1 - \epsilon$ and $\beta_n \le 2^{-nE}$. Plugging it in (11.3) and using $h \le \log 2$, we have

   $$-\log 2 + (1 - \epsilon) nE \le nD(P\|Q) \Rightarrow E \le \frac{D(P\|Q)}{1 - \epsilon} + \underbrace{\frac{\log 2}{n(1 - \epsilon)}}_{\to 0, \text{ as } n \to \infty}.$$

   Therefore

   $$V_\epsilon \le \frac{D(P\|Q)}{1 - \epsilon}$$

   Notice that this is weaker than what we hoped to prove, and this weak converse result is tight for $\epsilon \to 0$, i.e., for Stein's exponent we did have the desired result $V = \lim_{\epsilon \to 0} V_\epsilon \ge D(P\|Q)$.

b) (strong converse) In proving the weak converse, we only made use of the *expectation* of F_n in (11.3), we need to make use of the *entire distribution (CDF)* in order to obtain stronger results.

Recall the strong converse result which we showed in the last lecture:

$$\forall(\alpha,\beta)\in\mathcal{R}(P,Q),\forall\gamma,\quad \alpha-\gamma\beta\le P(F>\log\gamma)$$

Here, suppose there exists a sequence of tests $P_{Z|X_n}$ which achieve $\alpha_n\ge 1-\epsilon$ and $\beta_n\le 2^{-nE}$. Then

$$1-\epsilon-\gamma 2^{-nE}\le\alpha_n-\gamma\beta_n\le P_{X^n}[F_n>\log\gamma].$$

Pick $\log\gamma=n(D+\delta)$, by (11.1) the RHS goes to 0, and we have

$$\begin{aligned}&1-\epsilon-2^{n(D+\delta)}2^{-nE}\le o(1)\\ \Rightarrow&D+\delta-E\ge\frac{1}{n}\log(1-\epsilon+o(1))\to 0\\ \Rightarrow&E\le D\text{ as }\delta\to 0\\ \Rightarrow&V_\epsilon\le D\end{aligned}$$

□

Note: [Ergodic] Just like in last section of data compression. Ergodic assumptions on P_{X^n} and Q_{X^n} allow one to show that

$$V_\epsilon=\lim_{n\to\infty}\frac{1}{n}D(P_{X^n}\|Q_{X^n})$$

the counterpart of (11.3), which is the key for picking the appropriate τ, for ergodic sequence X^n is the Birkhoff-Khintchine convergence theorem.

Note: The theoretical importance of knowing the Stein's exponents is that:

$$\forall E\subset\mathcal{X}^n,\quad P_{X^n}[E]\ge 1-\epsilon\quad\Rightarrow Q_{X^n}[E]\ge 2^{-nV_\epsilon+o(n)}$$

Thus knowledge of Stein's exponent V_ϵ allows one to prove exponential bounds on probabilities of arbitrary sets, the technique is known as "change of measure".

11.2 Chernoff regime

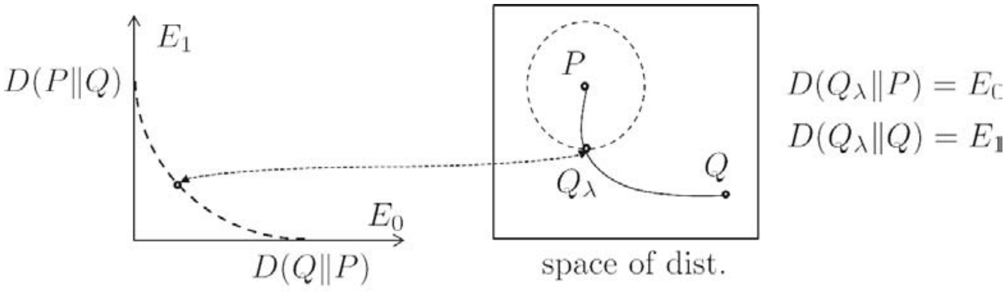
We are still considering i.i.d. sequence X^n, and binary hypothesis

$$H_0: X^n\sim P_X^n\qquad H_1: X^n\sim Q_X^n$$

But our objective in this section is to have both types of error probability to vanish exponentially fast simultaneously. We shall look at the following specification:

$$\begin{aligned}1-\alpha=\pi_{1|0}\to 0&\quad\text{at the rate }2^{-nE_0}\\ \beta=\pi_{0|1}\to 0&\quad\text{at the rate }2^{-nE_1}\end{aligned}$$

Apparently, E_0 (resp. E_1) can be made arbitrarily big at the price of making E_1 (resp. E_0) arbitrarily small. So the problem boils down to the optimal tradeoff, i.e., what's the achievable region of (E_0,E_1)? This problem is solved by [Hoeffding '65], [Blahut '74].



characterize the boundary of the achievable region of (E_0, E_1)

The optimal tests give the explict error probability:

$$\alpha_n = P\left[\frac{1}{n}F_n > \tau\right], \quad \beta_n = Q\left[\frac{1}{n}F_n > \tau\right]$$

and we are interested in the asymptotics when $n \to \infty$, in which scenario we know (11.1) and (11.2) occur.

Stein's regime corresponds to the corner points. Indeed, Theorem 11.1 tells us that when fixing $\alpha_n = 1 - \epsilon$, namely $E_0 = 0$, picking $\tau = D(P\|Q) - \delta$ ($\delta \to 0$) gives the exponential convergence rate of β_n as $E_1 = D(P\|Q)$. Similarly, exchanging the role of P and Q, we can achieves the point $(E_0, E_1) = (D(Q\|P), 0)$. More generally, to achieve the optimal tradeoff between the two corner points, we need to introduce a powerful tool – Large Deviation Theory.

Note: Here is a roadmap of the upcoming 2 lectures:

1. basics of large deviation (ψ_X, ψ_X^*, tilted distribution P_λ)

2. information projection problem

$$\min_{Q:\mathbb{E}_Q[X]\geq\gamma} D(Q\|P) = \psi^*(\gamma)$$

3. use information projection to prove tight Chernoff bound

$$\mathbb{P}\left[\frac{1}{n}\sum_{k=1}^n X_k \geq \gamma\right] = 2^{-n\psi^*(\gamma)+o(n)}$$

4. apply the above large deviation theorem to (E_0, E_1) to get

$$(E_0(\theta) = \psi_P^*(\theta), \quad E_1(\theta) = \psi_P^*(\theta) - \theta) \quad \text{characterize the achievable boundary.}$$

11.3 Basics of Large deviation theory

Let X^n be an i.i.d. sequence and $X_i \sim P$. Large deviation focuses on the following inequality:

$$P\left[\sum_{i=1}^n X_i \geq n\gamma\right] = 2^{-nE(\gamma)+o(n)}$$

what is the rate function $E(\gamma) = -\lim_{n\to\infty}\frac{1}{n}\log P\left[\frac{\sum_{i=1}^n X_i}{n} \geq \gamma\right]$? (Chernoff's ineq.)

To motivate, let us recall the usual Chernoff bound: For iid X^n, for any $\lambda \geq 0$,

$$\begin{aligned}
\mathbb{P}\left[\sum_{i=1}^n X_i \geq n\gamma\right] &= \mathbb{P}\left[\exp\left(\lambda\sum_{i=1}^n X_i\right) \geq \exp(n\lambda\gamma)\right] \\
&\overset{\text{Markov}}{\leq} \exp(-n\lambda\gamma)\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n X_i\right)\right] \\
&= \exp\left\{-n\lambda\gamma + n\log\mathbb{E}\left[\exp(\lambda X)\right]\right\}.
\end{aligned}$$

Optimizing over $\lambda \geq 0$ gives the *non-asymptotic* upper bound (concentration inequality) which holds for any n:

$$\mathbb{P}\left[\sum_{i=1}^n X_i \geq n\gamma\right] \leq \exp\Big\{-n\sup_{\lambda\geq 0}(\lambda\gamma - \underbrace{\log\mathbb{E}\left[\exp(\lambda X)\right]}_{\text{log MGF}})\Big\}.$$

Of course we still need to show the lower bound.

Let's first introduce the two key quantities: *log MGF* (also known as the *cumulant generating function*) $\psi_X(\lambda)$ and *tilted distribution* P_λ.

11.3.1 log MGF

Definition 11.3 (log MGF)**.**

$$\psi_X(\lambda) = \log(\mathbb{E}[\exp(\lambda X)]), \quad \lambda \in \mathbb{R}.$$

Per the usual convention, we will also denote $\psi_P(\lambda) = \psi_X(\lambda)$ if $X \sim P$.

Assumptions: In this section, we shall restrict to the distribution P_X such that

1. MGF exists, i.e., $\forall\lambda \in \mathbb{R}, \psi_X(\lambda) < \infty$,
2. $X \neq$const.

Example:

- Gaussian: $X \sim \mathcal{N}(0,1) \Rightarrow \psi_X(\lambda) = \frac{\lambda^2}{2}$.
- Example of R.V. such that $\psi_X(\lambda)$ does not exist: $X = Z^3$ with $Z \sim$ Gaussian. Then $\psi_X(\lambda) = \infty, \forall\lambda] \neq 0$.

Theorem 11.2 (Properties of ψ_X)**.**

1. *ψ_X is convex;*
2. *ψ_X is continuous;*
3. *ψ_X is infinitely differentiable and*

$$\psi'_X(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} = e^{-\psi_X(\lambda)}\mathbb{E}[Xe^{\lambda X}].$$

In particular, $\psi_X(0) = 0, \psi'_X(0) = \mathbb{E}[X]$.

4. *If $a \leq X \leq b$ a.s., then $a \leq \psi'_X \leq b$;*

5. *Conversely, if*

$$A = \inf_{\lambda\in\mathbb{R}} \psi'_X(\lambda), \quad B = \sup_{\lambda\in\mathbb{R}} \psi'_X(\lambda),$$

then $A \le X \le B$ *a.s.;*

6. ψ_X *is strictly convex, and consequently,* ψ'_X *is strictly increasing.*

7. *Chernoff bound:*

$$P(X \ge \gamma) \le \exp(-\lambda\gamma + \psi_X(\lambda)), \quad \lambda \ge 0.$$

Remark 11.1. The slope of log MGF encodes the range of X. Indeed, 4) and 5) of Theorem 11.2 together show that the smallest closed interval containing the support of P_X equals (closure of) the range of ψ'_X. In other words, A and B coincide with the essential infimum and supremum (min and max of RV in the probabilistic sense) of X respectively,

$$\begin{aligned} A &= \operatorname{essinf} X \triangleq \sup\{a : X \ge a \text{ a.s.}\} \\ B &= \operatorname{esssup} X \triangleq \inf\{b : X \le b \text{ a.s.}\} \end{aligned}$$

Proof. Note: 1–4 can be proved right now. 7 is the usual Chernoff bound. The proof of 5–6 relies on Theorem 11.4, which can be skipped for now.

1. Fix $\theta \in (0,1)$. Recall Holder's inequality:

$$\mathbb{E}[|UV|] \le \|U\|_p \|V\|_q, \quad \text{for } p,q \ge 1, \frac{1}{p} + \frac{1}{q} = 1$$

where the L_p-norm of RV is defined by $\|U\|_p = (\mathbb{E}|U|^p)^{1/p}$. Applying to $\mathbb{E}[e^{(\theta\lambda_1+\bar{\theta}\lambda_2)X}]$ with $p = 1/\theta, q = 1/\bar{\theta}$, we get

$$\mathbb{E}[\exp((\lambda_1/p + \lambda_2/q)X)] \le \|\exp(\lambda_1 X/p)\|_p \|\exp(\lambda_2 X/q)\|_q = \mathbb{E}[\exp(\lambda_1 X)]^\theta \mathbb{E}[\exp(\lambda_2 X)]^{\bar{\theta}},$$

i.e., $e^{\psi_X(\theta\lambda_1+\bar{\theta}\lambda_2)} \le e^{\psi_X(\lambda_1)\theta} e^{\psi_X(\lambda_2)\bar{\theta}}$.

2. By our assumptions on X, domain of ψ_X is $\mathbb{R}$, and by the fact that convex function must be continuous on the interior of its domain, we have that ψ_X is continuous on $\mathbb{R}$.

3. Be careful when exchanging the order of differentiation and expectation.

   Assume $\lambda > 0$ (similar for $\lambda \le 0$).
   First, we show that $\mathbb{E}[|Xe^{\lambda X}|]$ exists. Since

$$\begin{aligned} &e^{|X|} \le e^X + e^{-X} \\ &|Xe^{\lambda X}| \le e^{|(\lambda+1)X|} \le e^{(\lambda+1)X} + e^{-(\lambda+1)X} \end{aligned}$$

   by assumption on X, both of the summands are absolutely integrable in X. Therefore by dominated convergence theorem (DCT), $\mathbb{E}[|Xe^{\lambda X}|]$ exists and is continuous in λ.

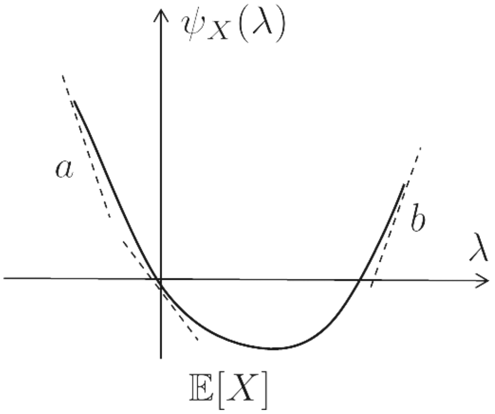
   Second, by the existence and continuity of $\mathbb{E}[|Xe^{\lambda X}|]$, $u \mapsto \mathbb{E}[|Xe^{uX}|]$ is integrable on $[0,\lambda]$, we can switch order of integration and differentiation as follows:

$$\begin{aligned} &e^{\psi_X(\lambda)} = \mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[1 + \int_0^\lambda Xe^{uX} du\right] \overset{\text{Fubini}}{=} 1 + \int_0^\lambda \mathbb{E}\left[Xe^{uX}\right] du \\ \Rightarrow\ &\psi'_X(\lambda) e^{\psi_X(\lambda)} = \mathbb{E}[Xe^{\lambda X}] \end{aligned}$$

thus $\psi'_X(\lambda) = e^{-\psi_X(\lambda)}\mathbb{E}[Xe^{\lambda X}]$ exists and is continuous in λ on $\mathbb{R}$.

Furthermore, using similar application of DCT we can extend to $\lambda \in \mathbb{C}$ and show that $\lambda \mapsto \mathbb{E}[e^{\lambda X}]$ is a holomorphic function. Thus it is infinitely differentiable.

4. $$a \le X \le b \Rightarrow \psi'_X(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \in [a, b].$$



5. Suppose $P_X[X > B] > 0$ (for contradiction), then $P_X[X > B + 2\epsilon] > 0$ for some small $\epsilon > 0$. But then $P_\lambda[X \le B + \epsilon] \to 0$ for $\lambda \to \infty$ (see Theorem 11.4.3 below). On the other hand, we know from Theorem 11.4.2 that $\mathbb{E}_{P_\lambda}[X] = \psi'_X(\lambda) \le B$. This is not yet a contradiction, since P_λ might still have some very small mass at a very negative value. To show that this cannot happen, we first assume that $B - \epsilon > 0$ (otherwise just replace X with $X - 2B$). Next note that

$$\begin{aligned} B \ge \mathbb{E}_{P_\lambda}[X] &= \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{X<B-\epsilon\}}] + \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{B-\epsilon\le X\le B+\epsilon\}}] + \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{X>B+\epsilon\}}] \\ &\ge \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{X<B-\epsilon\}}] + \mathbb{E}_{P_\lambda}[X\mathbf{1}_{\{X>B+\epsilon\}}] \\ &\ge -\mathbb{E}_{P_\lambda}[|X|\mathbf{1}_{\{X<B-\epsilon\}}] + (B+\epsilon)\underbrace{P_\lambda[X > B+\epsilon]}_{\to 1} \end{aligned} \tag{11.4}$$

therefore we will obtain a contradiction if we can show that $\mathbb{E}_{P_\lambda}[|X|\mathbf{1}_{\{X<B-\epsilon\}}] \to 0$ as $\lambda \to \infty$. To that end, notice that convexity of ψ_X implies that $\psi'_X \nearrow B$. Thus, for all $\lambda \ge \lambda_0$ we have $\psi'_X(\lambda) \ge B - \frac{\epsilon}{2}$. Thus, we have for all $\lambda \ge \lambda_0$

$$\psi_X(\lambda) \ge \psi_X(\lambda_0) + (\lambda - \lambda_0)(B - \frac{\epsilon}{2}) = c + \lambda(B - \frac{\epsilon}{2}), \tag{11.5}$$

for some constant c. Then,

$$\begin{aligned} \mathbb{E}_{P_\lambda}[|X|1\{X < B - \epsilon\}] &= \mathbb{E}[|X|e^{\lambda X - \psi_X(\lambda)}1\{X < B - \epsilon\}] && (11.6) \\ &\le \mathbb{E}[|X|e^{\lambda X - c - \lambda(B-\frac{\epsilon}{2})}1\{X < B - \epsilon\}] && (11.7) \\ &\le \mathbb{E}[|X|e^{\lambda(B-\epsilon) - c - \lambda(B-\frac{\epsilon}{2})}] && (11.8) \\ &= \mathbb{E}[|X|]e^{-\lambda\frac{\epsilon}{2} - c} \to 0 \quad \lambda \to \infty && (11.9) \end{aligned}$$

where the first inequality is from (11.5) and the second from $X < B - \epsilon$. Thus, the first term in (11.4) goes to 0 implying the desired contradiction.

6. Suppose ψ_X is not strictly convex. Since we know that ψ_X is convex, then ψ_X must be "flat" (affine) near some point, i.e., there exists a small neighborhood of some λ_0 such that $\psi_X(\lambda_0 + u) = \psi_X(\lambda_0) + ur$ for some $r \in \mathbb{R}$. Then $\psi_{P_\lambda}(u) = ur$ for all u in small neighborhood of zero, or equivalently $\mathbb{E}_{P_\lambda}[e^{u(X-r)}] = 1$ for u small. The following Lemma 11.1 implies $P_\lambda[X = r] = 1$, but then $P[X = r] = 1$, contradicting the assumption $X \neq \text{const}$.

□

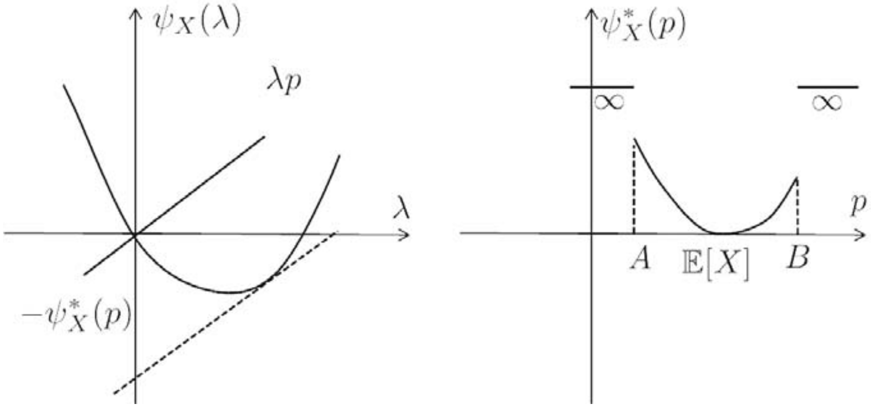
Lemma 11.1. *$\mathbb{E}[e^{uS}] = 1$ for all $u \in (-\epsilon, \epsilon)$ then $S = 0$.*

Proof. Expand in Taylor series around $u = 0$ to obtain $E[S] = 0$, $E[S^2] = 0$. Alternatively, we can extend the argument we gave for differentiating $\psi_X(\lambda)$ to show that the function $z \mapsto \mathbb{E}[e^{zS}]$ is holomorphic on the entire complex plane[1]. Thus by uniqueness, $\mathbb{E}[e^{uS}] = 1$ for all u. □

Definition 11.4 (Rate function)**.** The rate function $\psi_X^* : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is given by the *Legendre-Fenchel transform* of the log MGF:

$$\psi_X^*(\gamma) = \sup_{\lambda \in \mathbb{R}} \lambda\gamma - \psi_X(\lambda) \tag{11.10}$$

Note: The maximization (11.10) is a nice convex optimization problem since ψ_X is strictly convex, so we are maximizing a strictly concave function. So we can find the maximum by taking the derivative and finding the stationary point. In fact, ψ_X^* is the *dual* of ψ_X in the sense of convex analysis.



Theorem 11.3 (Properties of ψ_X^*)**.**

1. *Let $A = \text{essinf}\, X$ and $B = \text{esssup}\, X$. Then*

$$\psi_X^*(\gamma) = \begin{cases} \lambda\gamma - \psi_X(\lambda) \text{ for some } \lambda \text{ s.t. } \gamma = \psi_X'(\lambda), & A < \gamma < B \\ \log \frac{1}{P(X=\gamma)} & \gamma = A \text{ or } B \\ +\infty, & \gamma < A \text{ or } \gamma > B \end{cases}$$

2. *ψ_X^* is strictly convex and strictly positive except $\psi_X^*(\mathbb{E}[X]) = 0$.*

3. *ψ_X^* is decreasing when $\gamma \in (A, \mathbb{E}[X])$, and increasing when $\gamma \in [\mathbb{E}[X], B)$*

[1] More precisely, if we only know that $\mathbb{E}[e^{\lambda S}]$ is finite for $|\lambda| \leq 1$ then the function $z \mapsto \mathbb{E}[e^{zS}]$ is holomorphic in the vertical strip $\{z : |\text{Re}z| < 1\}$.

Proof. By Theorem 11.2.4, since $A \le X \le B$ a.s., we have $A \le \psi'_X \le B$. When $\gamma \in (A, B)$, the strictly concave function $\lambda \mapsto \lambda\gamma - \psi_X(\lambda)$ has a single stationary point which achieves the unique maximum. When $\gamma > B$ (resp. $< A$), $\lambda \mapsto \lambda\gamma - \psi_X(\lambda)$ increases (resp. decreases) without bounds. When $\gamma = B$, since $X \le B$ a.s., we have

$$\begin{aligned}\psi^*_X(B) &= \sup_{\lambda\in\mathbb{R}} \lambda B - \log(\mathbb{E}[\exp(\lambda X)]) = -\log \inf_{\lambda\in\mathbb{R}} \mathbb{E}[\exp(\lambda(X-B))] \\ &= -\log \lim_{\lambda\to\infty} \mathbb{E}[\exp(\lambda(X-B))] = -\log P(X=B),\end{aligned}$$

by monotone convergence theorem.

By Theorem 11.2.6, since ψ_X is strictly convex, the derivative of ψ_X and ψ^*_X are inverse to each other. Hence ψ^*_X is strictly convex. Since $\psi_X(0) = 0$, we have $\psi^*_X(\gamma) \ge 0$. Moreover, $\psi^*_X(\mathbb{E}[X]) = 0$ follows from $\mathbb{E}[X] = \psi'_X(0)$. □

11.3.2 Tilted distribution

As early as in Lecture 3, we have already introduced *tilting* in the proof of Donsker-Varadhan's variational characterization of divergence (Theorem 3.6). Let us formally define it now.

Definition 11.5 (Tilting)**.** Given $X \sim P$, the tilted measure P_λ is defined by

$$P_\lambda(dx) = \frac{e^{\lambda x}}{\mathbb{E}[e^{\lambda X}]} P(dx) = e^{\lambda x - \psi_X(\lambda)} P(dx) \tag{11.11}$$

In other words, if P has a pdf p, then the pdf of P_λ is given by $p_\lambda(x) = e^{\lambda x - \psi_X(\lambda)} p(x)$.

Note: The set of distributions $\{P_\lambda : \lambda \in \mathbb{R}\}$ parametrized by λ is called a *standard exponential family*, a very useful model in statistics. See [Bro86, p. 13].
Example:

- *Gaussian*: $P = \mathcal{N}(0,1)$ with density $p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$. Then P_λ has density $\frac{\exp(\lambda x)}{\exp(\lambda^2/2)} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) = \frac{1}{\sqrt{2\pi}} \exp(-(x-\lambda)^2/2)$. Hence $P_\lambda = \mathcal{N}(\lambda, 1)$.
- *Binary*: P is uniform on $\{\pm 1\}$. Then $P_\lambda(1) = \frac{e^\lambda}{e^\lambda + e^{-\lambda}}$ which puts more (resp. less) mass on 1 if $\lambda > 0$ (resp. < 0). Moreover, $P_\lambda \xrightarrow{\mathrm{D}} \delta_1$ if $\lambda \to \infty$ or δ_{-1} if $\lambda \to -\infty$.
- *Uniform*: P is uniform on $[0,1]$. Then P_λ is also supported on $[0,1]$ with pdf $p_\lambda(x) = \frac{\lambda \exp(\lambda x)}{e^\lambda - 1}$. Therefore as λ increases, P_λ becomes increasingly concentrated near 1, and $P_\lambda \to \delta_1$ as $\lambda \to \infty$. Similarly, $P_\lambda \to \delta_0$ as $\lambda \to -\infty$.

So we see that P_λ shifts the mean of P to the right (resp. left) when $\lambda > 0$ (resp. < 0). Indeed, this is a general property of tilting.

Theorem 11.4 (Properties of P_λ)**.**

1. *Log MGF:*

$$\psi_{P_\lambda}(u) = \psi_X(\lambda + u) - \psi_X(\lambda)$$

2. *Tilting trades mean for divergence:*

$$\mathbb{E}_{P_\lambda}[X] = \psi'_X(\lambda) \gtrless \mathbb{E}_P[X] \text{ if } \lambda \gtrless 0. \tag{11.12}$$

$$D(P_\lambda \| P) = \psi^*_X(\psi'_X(\lambda)) = \psi^*_X(\mathbb{E}_{P_\lambda}[X]). \tag{11.13}$$

3.

$$P(X > b) > 0 \Rightarrow \forall \epsilon > 0, P_\lambda(X \le b - \epsilon) \to 0 \ \textit{as}\ \lambda \to \infty$$
$$P(X < a) > 0 \Rightarrow \forall \epsilon > 0, P_\lambda(X \ge a + \epsilon) \to 0 \ \textit{as}\ \lambda \to -\infty$$

Therefore if $X_\lambda \sim P_\lambda$, then $X_\lambda \xrightarrow{\text{D}} \operatorname{essinf} X = A$ as $\lambda \to -\infty$ and $X_\lambda \xrightarrow{\text{D}} \operatorname{esssup} X = B$ as $\lambda \to \infty$.

Proof. 1. By definition. (DIY)

2. $\mathbb{E}_{P_\lambda}[X] = \frac{\mathbb{E}[X \exp(\lambda X)]}{\mathbb{E}[\exp(\lambda X)]} = \psi_X'(\lambda)$, which is strictly increasing in λ, with $\psi_X'(0) = \mathbb{E}_P[X]$.

   $D(P_\lambda \| P) = \mathbb{E}_{P_\lambda} \log \frac{dP_\lambda}{dP} = \mathbb{E}_{P_\lambda} \log \frac{\exp(\lambda X)}{\mathbb{E}[\exp(\lambda X)]} = \lambda \mathbb{E}_{P_\lambda}[X] - \psi_X(\lambda) = \lambda \psi_X'(\lambda) - \psi_X(\lambda) = \psi_X^*(\psi_X'(\lambda))$, where the last equality follows from Theorem 11.3.1.

3.

$$\begin{aligned} P_\lambda(X \le b - \epsilon) &= \mathbb{E}_P[e^{\lambda X - \psi_X(\lambda)} \mathbf{1}[X \le b - \epsilon]] \\ &\le \mathbb{E}_P[e^{\lambda(b-\epsilon) - \psi_X(\lambda)} \mathbf{1}[X \le b - \epsilon]] \\ &\le e^{-\lambda\epsilon} e^{\lambda b - \psi_X(\lambda)} \\ &\le \frac{e^{-\lambda\epsilon}}{P[X > b]} \to 0 \text{ as } \lambda \to \infty \end{aligned}$$

where the last inequality is due to the usual Chernoff bound (Theorem 11.2.7): $P[X > b] \le \exp(-\lambda b + \psi_X(\lambda))$.

□

12.1 Large-deviation exponents

Large deviations problems make statements about the tail probabilities of a sequence of distributions. We're interested in the speed of decay for probabilities such as $P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right]$ for iid X_k.

In the last lecture we used Chernoff bound to obtain an upper bound on the exponent via the log-MGF and tilting. Next we use a different method to give a formula for the exponent as a convex optimization problem involving the KL divergence (information projection). Later in Section 12.3 we shall revisit the Chernoff bound after we have computed the value of the information projection.

Theorem 12.1. *Let $X^n \overset{i.i.d.}{\sim} P$. Then for any $\gamma \in \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P\left[\frac{1}{n}\sum_{k=1}^{n} X_k > \gamma\right]} = \inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P) \tag{12.1}$$

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right]} = \inf_{Q:\mathbb{E}_Q[X]\geq\gamma} D(Q\|P) \tag{12.2}$$

Proof. We first prove (12.1). Set $P[E_n] = P\left[\frac{1}{n}\sum_{k=1}^{n} X_k > \gamma\right]$.

Lower Bound on $P[E_n]$: Fix a Q such that $\mathbb{E}_Q[X] > \gamma$. Let X^n be iid. Then by WLLN,

$$Q[E_n] = Q\left[\sum_{k=1}^{n} X_k > n\gamma\right] = 1 - o(1).$$

Now the data processing inequality gives

$$d(Q[E_n]\|P[E_n]) \leq D(Q_{X^n}\|P_{X^n}) = nD(Q\|P)$$

And a lower bound for the binary divergence is

$$d(Q[E_n]\|P[E_n]) \geq -h(Q[E_n]) + Q[E_n]\log\frac{1}{P[E_n]}$$

Combining the two bounds on $d(Q[E_n]\|P[E_n])$ gives

$$P[E_n] \geq \exp\left(\frac{-nD(Q\|P) - \log 2}{Q[E_n]}\right) \tag{12.3}$$

Optimizing over Q to give the best bound:

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P[E_n]} \leq \inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P).$$

Upper Bound on $P[E_n]$: The key observation is that given any X and any event E, $P_X(E) > 0$ can be expressed via the divergence between the conditional and unconditional distribution as: $\log \frac{1}{P_X(E)} = D(P_{X|X\in E} \| P_X)$. Define $\tilde{P}_{X^n} = P_{X^n|\sum X_i > n\gamma}$, under which $\sum X_i > n\gamma$ holds a.s. Then

$$\log \frac{1}{P[E_n]} = D(\tilde{P}_{X^n} \| P_{X^n}) \geq \inf_{Q_{X^n}: \mathbb{E}_Q[\sum X_i] > n\gamma} D(Q_{X^n} \| P_{X^n}) \tag{12.4}$$

We know show that the last problem "single-letterizes", i.e. need to be solved only for $n = 1$. Consider the following two steps:

$$D(Q_{X^n} \| P_{X^n}) \geq \sum_{j=1}^{n} D(Q_{X_j} \| P) \tag{12.5}$$

$$\geq n D(\bar{Q} \| P), \qquad \bar{Q} \triangleq \frac{1}{n} \sum_{j=1}^{n} Q_{X_j}, \tag{12.6}$$

where the first step follows from Corollary 2.1 after noticing that $P_{X^n} = P^n$, and the second step is by convexity of divergence Theorem 4.1. From this argument we conclude that

$$\inf_{Q_{X^n}: \mathbb{E}_Q[\sum X_i] > n\gamma} D(Q_{X^n} \| P_{X^n}) = n \cdot \inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q \| P) \tag{12.7}$$

$$\inf_{Q_{X^n}: \mathbb{E}_Q[\sum X_i] \geq n\gamma} D(Q_{X^n} \| P_{X^n}) = n \cdot \inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q \| P) \tag{12.8}$$

In particular, (12.4) and (12.7) imply the required lower bound in (12.1).

Next we prove (12.2). First, notice that the lower bound argument (12.4) applies equally well, so that for each n we have

$$\frac{1}{n} \log \frac{1}{P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right]} \geq \inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q \| P).$$

To get a matching upper bound we consider two cases:

- Case I: $P[X > \gamma] = 0$. If $P[X \geq \gamma] = 0$, then both sides of (12.2) are $+\infty$. If $P[X = \gamma] > 0$, then $P[\sum X_k \geq n\gamma] = P[X_1 = \ldots = X_n = \gamma] = P[X = \gamma]^n$. For the right-hand side, since $D(Q \| P) < \infty \implies Q \ll P \implies Q(X \leq \gamma) = 1$, the only possibility for $\mathbb{E}_Q[X] \geq \gamma$ is that $Q(X = \gamma) = 1$, i.e., $Q = \delta_\gamma$. Then $\inf_{\mathbb{E}_Q[X] \geq \gamma} D(Q \| P) = \log \frac{1}{P(X=\gamma)}$.

- Case II: $P[X > \gamma] > 0$. Since $\mathbb{P}[\sum X_k \geq \gamma] \geq \mathbb{P}[\sum X_k > \gamma]$ from (12.1) we know that

  $$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{P\left[\frac{1}{n}\sum_{k=1}^{n} X_k \geq \gamma\right]} \leq \inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q \| P).$$

  We next show that in this case

  $$\inf_{Q: \mathbb{E}_Q[X] > \gamma} D(Q \| P) = \inf_{Q: \mathbb{E}_Q[X] \geq \gamma} D(Q \| P) \tag{12.9}$$

  Indeed, let $\tilde{P} = P_{X|X>\gamma}$ which is well defined since $P[X > \gamma] > 0$. For any Q such that $\mathbb{E}_Q[X] \geq \gamma$, set $\tilde{Q} = \bar{\epsilon} Q + \epsilon \tilde{P}$ satisfies $\mathbb{E}_{\tilde{Q}}[X] > \gamma$. Then by convexity, $D(Q \| P) \leq \bar{\epsilon} D(Q \| P) + \epsilon D(\tilde{P} \| P) = \bar{\epsilon} D(Q \| P) + \epsilon \log \frac{1}{P[X>\gamma]}$. Sending $\epsilon \to 0$, we conclude the proof of (12.9).

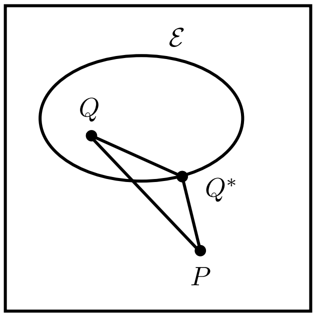
□

12.2 Information Projection

The results of Theorem 12.1 motivate us to study the following general **information projection problem:** Let $\mathcal{E}$ be a convex set of distributions on some abstract space Ω, then for the distribution P on Ω, we want

$$\inf_{Q\in\mathcal{E}} D(Q\|P)$$

Denote the minimizing distribution Q by Q^*. The next result shows that intuitively the "line" between P and optimal Q^* is "orthogonal" to $\mathcal{E}$.



Distributions on $\mathcal{X}$

Theorem 12.2. *Suppose $\exists Q^* \in \mathcal{E}$ such that $D(Q^*\|P) = \min_{Q\in\mathcal{E}} D(Q\|P)$, then $\forall Q \in \mathcal{E}$*

$$D(Q\|P) \geq D(Q\|Q^*) + D(Q^*\|P)$$

Proof. If $D(Q\|P) = \infty$, then we're done, so we can assume that $D(Q\|P) < \infty$, which also implies that $D(Q^*\|P) < \infty$. For $\theta \in [0,1]$, form the convex combination $Q^{(\theta)} = \bar{\theta}Q^* + \theta Q \in \mathcal{E}$. Since Q^* is the minimizer of $D(Q\|P)$, then[1]

$$0 \leq \left.\frac{\partial}{\partial\theta}\right|_{\theta=0} D(Q^{(\theta)}\|P) = D(Q\|P) - D(Q\|Q^*) - D(Q^*\|P)$$

and we're done. □

Remark: If we view the picture above in the Euclidean setting, the "triangle" formed by P, Q^* and Q (for Q^*, Q in a convex set, P outside the set) is always obtuse, and is a right triangle only when the convex set has a "flat face". In this sense, the divergence is similar to the squared Euclidean distance, and the above theorem is sometimes known as a "Pythagorean" theorem.

The interesting set of Q's that we will particularize to is the "half-space" of distributions $\mathcal{E} = \{Q : \mathbb{E}_Q[X] \geq \gamma\}$, where $X : \Omega \to \mathbb{R}$ is some fixed function. This is justified by relation (to be established) with the large deviation exponent in Theorem 12.1. First, we solve this I-projection problem explicitly.

Theorem 12.3. *Given distribution P on Ω and $X : \Omega \to \mathbb{R}$ let*

$$A = \inf \psi'_X = \text{essinf}\, X = \sup\{a : X \geq a\ P\text{-}a.s.\} \tag{12.10}$$
$$B = \sup \psi'_X = \text{esssup}\, X = \inf\{b : X \leq b\ P\text{-}a.s.\} \tag{12.11}$$

[1] This can be found by taking the derivative and matching terms (Exercise). Be careful with exchanging derivatives and integrals. Need to use dominated convergence theorem similar as in the "local behavior of divergence" in Proposition 4.1.

1. *The information projection problem over $\mathcal{E} = \{Q : \mathbb{E}_Q[X] \ge \gamma\}$ has solution*

$$\min_{Q\,:\,\mathbb{E}_Q[X]\ge\gamma} D(Q\|P) = \begin{cases} 0 & \gamma < \mathbb{E}_P[X] \\ \psi^*(\gamma) & \mathbb{E}_P[X] \le \gamma < B \\ \log\frac{1}{P(X=B)} & \gamma = B \\ +\infty & \gamma > B \end{cases} \tag{12.12}$$

$$= \psi^*(\gamma)1\{\gamma \ge \mathbb{E}_P[X]\} \tag{12.13}$$

2. *Whenever the minimum is finite, minimizing distribution is unique and equal to tilting of P along X, namely*[2]

$$dP_\lambda = \exp\{\lambda X - \psi(\lambda)\} \cdot dP \tag{12.14}$$

3. *For all $\gamma \in [\mathbb{E}_P[X], B)$ we have*

$$\min_{\mathbb{E}_Q[X]\ge\gamma} D(Q\|P) = \inf_{\mathbb{E}_Q[X]>\gamma} D(Q\|P) = \min_{\mathbb{E}_Q[X]=\gamma} D(Q\|P)\,.$$

Note: An alternative expression is

$$\min_{Q:\mathbb{E}_Q[X]\ge\gamma} = \sup_{\lambda\ge 0} \lambda\gamma - \psi_X(\lambda)\,.$$

Proof. First case: Take $Q = P$.

Fourth case: If $\mathbb{E}_Q[X] > B$, then $Q[X \ge B + \epsilon] > 0$ for some $\epsilon > 0$, but $P[X \ge B + \epsilon] = 0$, since $P(X \le B) = 1$, by Theorem 11.2.5. Hence $Q \not\ll P \implies D(Q\|P) = \infty$.

Third case: If $P(X = B) = 0$, then $X < B$ a.s. under P, and $Q \not\ll P$ for any Q s.t. $\mathbb{E}_Q[X] \ge B$. Then the minimum is ∞. Now assume $P(X = B) > 0$. Since $D(Q\|P) < \infty \implies Q \ll P \implies Q(X \le B) = 1$. Therefore the only possibility for $\mathbb{E}_Q[X] \ge B$ is that $Q(X = B) = 1$, i.e., $Q = \delta_B$. Then $D(Q\|P) = \log\frac{1}{P(X=B)}$.

Second case: Fix $\mathbb{E}_P[X] \le \gamma < B$, and find the unique λ such that $\psi'_X(\lambda) = \gamma = \mathbb{E}_{P_\lambda}[X]$ where $dP_\lambda = \exp(\lambda X - \psi_X(\lambda))dP$. This corresponds to tilting P far enough to the right to increase its mean from $\mathbb{E}_P X$ to γ, in particular $\lambda \ge 0$. Moreover, $\psi^*_X(\gamma) = \lambda\gamma - \psi_X(\lambda)$. Take any Q such that $\mathbb{E}_Q[X] \ge \gamma$, then

$$D(Q\|P) = \mathbb{E}_Q\left[\log\frac{dQdP_\lambda}{dPdP_\lambda}\right] \tag{12.15}$$

$$= D(Q\|P_\lambda) + \mathbb{E}_Q[\log\frac{dP_\lambda}{dP}] \tag{12.16}$$

$$= D(Q\|P_\lambda) + \mathbb{E}_Q[\lambda X - \psi_X(\lambda)] \tag{12.17}$$

$$\ge D(Q\|P_\lambda) + \lambda\gamma - \psi_X(\lambda) \tag{12.18}$$

$$= D(Q\|P_\lambda) + \psi^*_X(\gamma) \tag{12.19}$$

$$\ge \psi^*_X(\gamma), \tag{12.20}$$

where the last inequality holds with equality if and only if $Q = P_\lambda$. In addition, this shows the minimizer is unique, proving the second claim. Note that even in the corner case of $\gamma = B$ (assuming $P(X = B) > 0$) the minimizer is a point mass $Q = \delta_B$, which is also a tilted measure (P_∞), since $P_\lambda \to \delta_B$ as $\lambda \to \infty$, cf. Theorem 11.4.3.

[2] Note that unlike previous Lecture, here P and P_λ are measures on an abstract space Ω, not on a real line.

Another version of the solution, given by expression (12.13), follows from Theorem 11.3.

For the third claim, notice that there is nothing to prove for $\gamma < \mathbb{E}_P[X]$, while for $\gamma \ge \mathbb{E}_P[X]$ we have just shown

$$\psi_X^*(\gamma) = \min_{Q:\mathbb{E}_Q[X]\ge\gamma} D(Q\|P)$$

while from the next corollary we have

$$\inf_{Q:\mathbb{E}_Q[X]>\gamma} D(Q\|P) = \inf_{\gamma'>\gamma} \psi_X^*(\gamma') .$$

The final step is to notice that ψ_X^* is increasing and continuous by Theorem 11.3, and hence the right-hand side infimum equalis $\psi_X^*(\gamma)$. The case of $\min_{Q:\mathbb{E}_Q[X]=\gamma}$ is handled similarly. □

Corollary 12.1. *$\forall Q$ with $\mathbb{E}_Q[X] \in (A,B)$, there exists a unique $\lambda \in \mathbb{R}$ such that the tilted distribution P_λ satisfies*

$$\mathbb{E}_{P_\lambda}[X] = \mathbb{E}_Q[X]$$
$$D(P_\lambda\|P) \le D(Q\|P)$$

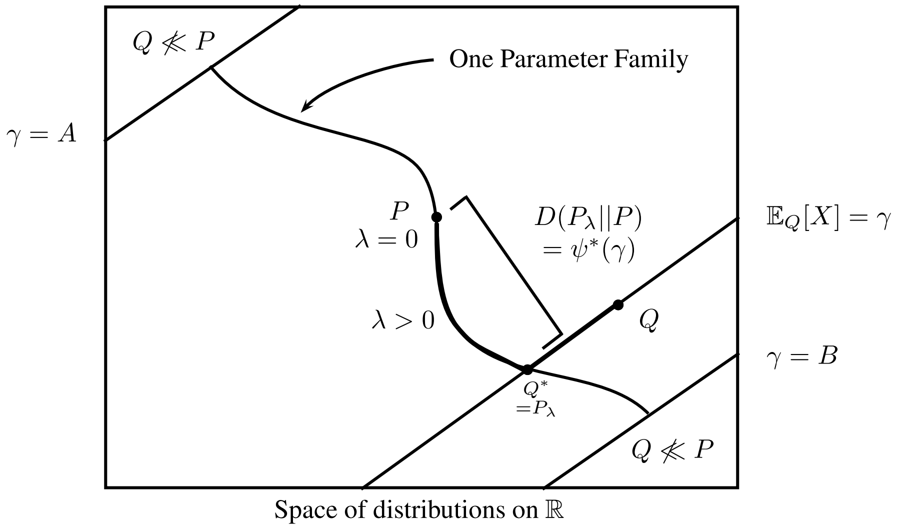
and furthermore the gap in the last inequality equals $D(Q\|P_\lambda) = D(Q\|P) - D(P_\lambda\|P)$.

Proof. Same as in the proof of Theorem 12.3, find the unique λ s.t. $\mathbb{E}_{P_\lambda}[X] = \psi_X'(\lambda) = \mathbb{E}_Q[X]$. Then $D(P_\lambda\|P) = \psi_X^*(\mathbb{E}_Q[X]) = \lambda\mathbb{E}_Q[X] - \psi_X(\lambda)$. Repeat the steps (12.15)-(12.20) obtaining $D(Q\|P) = D(Q\|P_\lambda) + D(P_\lambda\|P)$. □

Remark: For any Q, this allows us to find a tilted measure P_λ that has the same mean yet smaller (or equal) divergence.

12.3 Interpretation of Information Projection

The following picture describes many properties of information projections.



- Each set $\{Q : \mathbb{E}_Q[X] = \gamma\}$ corresponds to a slice. As γ varies from A to B, the curves fill the entire space minus the corner regions.

- When $\gamma < A$ or $\gamma > B$, $Q \not\ll P$.
- As γ varies, the P_λ's trace out a curve via $\psi^*(\gamma) = D(P_\lambda \| P)$. This set of distributions is called a *one parameter family*, or *exponential family.*

Key Point: The one parameter family curve intersects each γ-slice $\mathcal{E} = \{Q : \mathbb{E}_Q[X] = \gamma\}$ "orthogonally" at the minimizing $Q^* \in \mathcal{E}$, and the distance from P to Q^* is given by $\psi^*(\lambda)$. To see this, note that applying Theorem 12.2 to the convex set $\mathcal{E}$ gives us $D(Q\|P) \geq D(Q\|Q^*) + D(Q^*\|P)$. Now thanks to Corollary 12.1, we in fact have *equality* $D(Q\|P) = D(Q\|Q^*) + D(Q^*\|P)$ and $Q^* = P_\lambda$ for some tilted measure.

Chernoff bound revisited: The proof of upper bound in Theorem 12.1 is via the definition of information projection. Theorem 12.3 shows that the value of the information projection coincides with the rate function (conjugate of log-MGF). This shows the optimality of the Chernoff bound (recall Theorem 11.2.7). Indeed, we directly verify this for completeness: For all $\lambda \geq 0$,

$$P\left[\sum_{k=1}^{n} X_k \geq n\gamma\right] \leq e^{-n\gamma\lambda}(\mathbb{E}_P[e^{\lambda X}])^n = e^{-n(\lambda\gamma - \psi_X(\lambda))}$$

where we used iid X_k's in the expectation. Optimizing over $\lambda \geq 0$ to get the best upper bound:

$$\sup_{\lambda\geq 0} \lambda\gamma - \psi_X(\lambda) = \sup_{\lambda\in\mathbb{R}} \lambda\gamma - \psi_X(\lambda) = \psi_X^*(\gamma)$$

where the first equality follows since $\gamma \geq \mathbb{E}_P[X]$, therefore $\lambda \mapsto \lambda\gamma - \psi_X(\lambda)$ is increasing when $\lambda \leq 0$.

Remark: The Chernoff bound is tight precisely because, from information projection, the lower bound showed that the best change of measure is to change to the tilted measure P_λ.

12.4 Generalization: Sanov's theorem

Theorem 12.4 (Sanov's Theorem)**.** *Consider observing n samples $X_1, \ldots, X_n \sim$ iid P. Let $\hat{P}$ be the empirical distribution, i.e., $\hat{P} = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$. Let $\mathcal{E}$ be a convex set of distributions. Then under regularity conditions on $\mathcal{E}$ and P we have*

$$\mathbb{P}[\hat{P} \in \mathcal{E}] = e^{-n \min_{Q\in\mathcal{E}} D(Q\|P) + o(n)}$$

Note: Examples of regularity conditions: space $\mathcal{X}$ is finite and $\mathcal{E}$ is closed with non-empty interior; space $\mathcal{X}$ is Polish and the set $\mathcal{E}$ is weakly closed and has non-empty interior.

Proof sketch. The lower bound is proved as in Theorem 12.1: Just take an arbitrary $Q \in \mathcal{E}$ and apply a suitable version of WLLN to conclude $Q^n[\hat{P} \in \mathcal{E}] = 1 + o(1)$.

For the upper bound we can again adapt the proof from Theorem 12.1. Alternatively, we can write the convex set $\mathcal{E}$ as an intersection of half spaces. Then we've already solved the problem for half-spaces $\{Q : \mathbb{E}_Q[X] \geq \gamma\}$. The general case follows by the following consequence of Theorem 12.2: if Q^* is projection of P onto $\mathcal{E}_1$ and Q^{**} is projection of Q^* on $\mathcal{E}_2$, then Q^{**} is also projection of P onto $\mathcal{E}_1 \cap \mathcal{E}_2$:

$$D(Q^{**}\|P) = \min_{Q\in\mathcal{E}_1\cap\mathcal{E}_2} D(Q\|P) \Leftarrow \begin{cases} D(Q^*\|P) = \min_{Q\in\mathcal{E}_1} D(Q\|P) \\ D(Q^{**}\|Q^*) = \min_{Q\in\mathcal{E}_2} D(Q\|Q^*) \end{cases}$$

(Repeated projection property)

Indeed, by first tilting from P to Q^* we find

$$P[\hat{P} \in \mathcal{E}_1 \cap \mathcal{E}_2] \le 2^{-nD(Q^* \| P)} Q^*[\hat{P} \in \mathcal{E}_1 \cap \mathcal{E}_2] \tag{12.21}$$

$$\le 2^{-nD(Q^* \| P)} Q^*[\hat{P} \in \mathcal{E}_2] \tag{12.22}$$

and from here proceed by tilting from Q^* to Q^{**} and note that $D(Q^* \| P) + D(Q^{**} \| Q^*) = D(Q^{**} \| P)$. □

Remark: Sanov's theorem tells us the probability that, after observing n iid samples of a distribution, our empirical distribution is still far away from the true distribution, is exponentially small.

Setup:

$$\begin{aligned}&H_0 : X^n \sim P_{X^n} \qquad H_1 : X^n \sim Q_{X^n} \quad \text{(i.i.d.)}\\&\text{test } P_{Z|X^n} : \mathcal{X}^n \to \{0,1\}\\&\text{specification: } 1-\alpha = \pi^{(n)}_{1|0} \le 2^{-nE_0} \qquad \beta = \pi^{(n)}_{0|1} \le 2^{-nE_1}\end{aligned}$$

Bounds:

- achievability (Neyman Pearson)

$$\alpha = 1-\pi_{1|0} = P_{X^n}[F_n > \tau], \qquad \beta = \pi_{0|1} = Q_{X^n}[F_n > \tau]$$

- converse (strong)

$$\forall(\alpha,\beta) \text{ achievable, } \alpha - \gamma\beta \le P_{X^n}[F > \log\gamma]$$

where

$$F = \log \frac{dP_{X^n}}{dQ_{X^n}}(X^n),$$

13.1 (E_0, E_1)-Tradeoff

Goal:

$$1-\alpha \le 2^{-nE_0}, \quad \beta \le 2^{-nE_1}.$$

Our goal in the Chernoff regime is to find the best tradeoff, which we formally define as follows (compare to Stein's exponent in Lecture 11)

$$\begin{aligned}E_1^*(E_0) &\triangleq \sup\{E_1 : \exists n_0, \forall n \ge n_0, \exists P_{Z|X^n} \text{ s.t. } \alpha > 1-2^{-nE_0}, \beta < 2^{-nE_1}, \}\\&= \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_{1-2^{-nE_0}}(P^n, Q^n)}\end{aligned}$$

Define

$$T = \log\frac{dQ}{dP}(X), \quad T_k = \log\frac{dQ}{dP}(X_k), \quad \text{thus } \log\frac{dQ^n}{dP^n}(X^n) = \sum_{k=1}^n T_k$$

Log MGF of T under P (again assumed to be finite and also $T \ne \text{const}$ since $P \ne Q$):

$$\begin{aligned}\psi_P(\lambda) &= \log \mathbb{E}_P[e^{\lambda T}]\\&= \log\sum_x P(x)^{1-\lambda}Q(x)^\lambda = \log\int (dP)^{1-\lambda}(dQ)^\lambda\\ \psi_P^*(\theta) &= \sup_{\lambda\in\mathbb{R}} \theta\lambda - \psi_P(\lambda)\end{aligned}$$

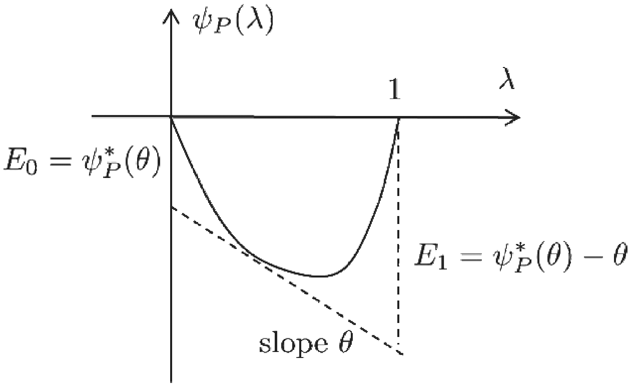
Note that since $\psi_P(0) = \psi_P(1) = 0$ from convexity $\psi_P(\lambda)$ is finite on $0 \le \lambda \le 1$. Furthermore, assuming $P \ll Q$ and $Q \ll P$ we also have that $\lambda \mapsto \psi_P(\lambda)$ continuous everywhere on $[0,1]$ (on $(0,1)$ it follows from convexity, but for boundary points we need more detailed arguments). Consequently, all the results in this section apply under just the conditions of $P \ll Q$ and $Q \ll P$. However, since in previous lecture we were assuming that log-MGF exists for all λ, we will only present proofs under this extra assumption.

Theorem 13.1. *Let $P \ll Q$, $Q \ll P$, then*

$$E_0(\theta) = \psi_P^*(\theta), \qquad E_1(\theta) = \psi_P^*(\theta) - \theta \tag{13.1}$$

parametrized by $-D(P\|Q) \le \theta \le D(Q\|P)$ characterizes the best exponents on the boundary of achievable (E_0, E_1).

Note: The geometric interpretation of the above theorem is shown in the following picture, which rely on the properties of $\psi_P(\lambda)$ and $\psi_P^*(\theta)$. Note that $\psi_P(0) = \psi_P(1) = 0$. Moreover, by Theorem 11.3 (Properties of ψ_X^*), $\theta \mapsto E_0(\theta)$ is increasing, $\theta \mapsto E_1(\theta)$ is decreasing.



Remark 13.1 (Rényi divergence)**.** Rényi defined a family of divergence indexed by $\lambda \neq 1$

$$D_\lambda(P\|Q) \triangleq \frac{1}{\lambda - 1} \log \mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^\lambda\right] \ge 0.$$

which generalizes Kullback-Leibler divergence since $D_\lambda(P\|Q) \xrightarrow{\lambda \to 1} D(P\|Q)$. Note that $\psi_P(\lambda) = (\lambda - 1)D_\lambda(Q\|P) = -\lambda D_{1-\lambda}(P\|Q)$. This provides another explanation that ψ_P is negative between 0 and 1, and the slope at endpoints is: $\psi_P'(0) = -D(P\|Q)$ and $\psi_P'(1) = D(Q\|P)$.

Corollary 13.1 (Bayesian criterion)**.** *Fix a prior (π_0, π_1) such that $\pi_0 + \pi_1 = 1$ and $0 < \pi_0 < 1$. Denote the optimal Bayesian (average) error probability by*

$$P_e^*(n) \triangleq \inf_{P_{Z|X^n}} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}$$

with exponent

$$E \triangleq \lim_{n \to \infty} \frac{1}{n} \log \frac{1}{P_e^*(n)}.$$

Then

$$E = \max_\theta \min(E_0(\theta), E_1(\theta)) = \psi_P^*(0) = -\inf_{\lambda \in \mathbb{R}} \psi_P(\lambda),$$

regardless of the prior, and $\psi_P^*(0) \triangleq C(P,Q)$ is called the Chernoff exponent*.*

Proof of Theorem 13.1. The idea is to apply the large deviation theory to iid sum $\sum_{k=1}^n T_k$. Specifically, let's rewrite the bounds in terms of T:

- Achievability (Neyman Pearson)

$$\text{let } \tau = -n\theta, \quad \pi_{1|0}^{(n)} = P\left[\sum_{k=1}^n T_k \geq n\theta\right] \quad \pi_{0|1}^{(n)} = Q\left[\sum_{k=1}^n T_k < n\theta\right]$$

- Converse (strong)

$$\text{let } \gamma = 2^{-n\theta}, \quad \pi_{1|0} + 2^{-n\theta}\pi_{0|1} \geq P\left[\sum_{k=1}^n T_k \geq n\theta\right]$$

Achievability: Using Neyman Pearson test, for fixed $\tau = -n\theta$, apply the large deviation theorem:

$$1-\alpha = \pi_{1|0}^{(n)} = P\left[\sum_{k=1}^n T_k \geq n\theta\right] = 2^{-n\psi_P^*(\theta)+o(n)}, \qquad \text{for } \theta \geq \mathbb{E}_P T = -D(P\|Q)$$
$$\beta = \pi_{0|1}^{(n)} = Q\left[\sum_{k=1}^n T_k < n\theta\right] = 2^{-n\psi_Q^*(\theta)+o(n)}, \qquad \text{for } \theta \leq \mathbb{E}_Q T = D(Q\|P)$$

Notice that by the definition of T we have

$$\psi_Q(\lambda) = \log \mathbb{E}_Q[e^{\lambda \log(Q/P)}] = \log \mathbb{E}_P[e^{(\lambda+1)\log(Q/P)}] = \psi_P(\lambda+1)$$
$$\Rightarrow \quad \psi_Q^*(\theta) = \sup_{\lambda\in\mathbb{R}} \theta\lambda - \psi_P(\lambda+1) = \psi_P^*(\theta) - \theta$$

thus (E_0, E_1) in (13.1) is achievable.

Converse: We want to show that any achievable (E_0, E_1) pair must be below the curve $(E_0(\theta), E_1(\theta))$ in the above Neyman-Pearson test with parameter θ. Apply the strong converse bound we have:

$$2^{-nE_0} + 2^{-n\theta}2^{-nE_1} \geq 2^{-n\psi_P^*(\theta)+o(n)}$$
$$\Rightarrow \min(E_0, E_1+\theta) \leq \psi_P^*(\theta), \ \forall n, \theta, -D(P\|Q) \leq \theta \leq D(Q\|P)$$
$$\Rightarrow \text{ either } E_0 \leq \psi_P^*(\theta) \text{ or } E_1 \leq \psi_P^*(\theta) - \theta$$

□

13.2 Equivalent forms of Theorem 13.1

Alternatively, the optimal (E_0, E_1)-tradeoff can be stated in the following equivalent forms:

Theorem 13.2. *1. The optimal exponents are given (parametrically) in terms of $\lambda \in [0,1]$ as*

$$E_0 = D(P_\lambda\|P), \qquad E_1 = D(P_\lambda\|Q) \tag{13.2}$$

where the distribution P_λ is tilting of P along T, cf. (12.14), which moves from $P_0 = P$ to $P_1 = Q$ as λ ranges from 0 to 1:

$$dP_\lambda = (dP)^{1-\lambda}(dQ)^\lambda \exp\{-\psi_P(\lambda)\}$$

2. *Yet another characterization of the boundary is*

$$E_1^*(E_0) = \min_{Q':D(Q'\|P)\le E_0} D(Q'\|Q), \qquad 0 \le E_0 \le D(Q\|P) \tag{13.3}$$

Proof. The first part is verified trivially. Indeed, if we fix λ and let $\theta(\lambda) \triangleq \mathbb{E}_{P_\lambda}[T]$, then from (11.13) we have

$$D(P_\lambda\|P) = \psi_P^*(\theta),$$

whereas

$$D(P_\lambda\|Q) = \mathbb{E}_{P_\lambda}[\log\frac{dP_\lambda}{dQ}] = \mathbb{E}_{P_\lambda}[\log\frac{dP_\lambda}{dP}\frac{dP}{dQ}] = D(P_\lambda\|P) - \mathbb{E}_{P_\lambda}[T] = \psi_P^*(\theta) - \theta.$$

Also from (11.12) we know that as λ ranges in $[0,1]$ the mean $\theta = \mathbb{E}_{P_\lambda}[T]$ ranges from $-D(P\|Q)$ to $D(Q\|P)$.

To prove the second claim (13.3), the key observation is the following: Since Q is itself a tilting of P along T (with $\lambda = 1$), the following two families of distributions

$$dP_\lambda = \exp\{\lambda T - \psi_P(\lambda)\}\cdot dP \tag{13.4}$$

$$dQ_{\lambda'} = \exp\{\lambda' T - \psi_Q(\lambda')\}\cdot dQ \tag{13.5}$$

are in fact the same family with $Q_{\lambda'} = P_{\lambda'+1}$.

Now, suppose that Q^* achieves the minimum in (13.3) and that $Q^* \ne Q$, $Q^* \ne P$ (these cases should be verified separately). Note that we have not shown that this minimum is achieved, but it will be clear that our argument can be extended to the case of when Q'_n is a sequence achieving the infimum. Then, on one hand, obviously

$$D(Q^*\|Q) = \min_{Q':D(Q'\|P)\le E_0} D(Q'\|Q) \le D(P\|Q)$$

On the other hand, since $E_0 \le D(Q\|P)$ we also have

$$D(Q^*\|P) \le D(Q\|P).$$

Therefore,

$$\mathbb{E}_{Q^*}[T] = \mathbb{E}_{Q^*}[\log\frac{dQ^*}{dP}\frac{dQ}{dQ^*}] = D(Q^*\|P) - D(Q^*\|Q) \in [-D(P\|Q), D(Q\|P)]. \tag{13.6}$$

Next, we have from Corollary 12.1 that there exists a unique P_λ with the following three properties:[1]

$$\mathbb{E}_{P_\lambda}[T] = \mathbb{E}_{Q^*}[T] \tag{13.7}$$

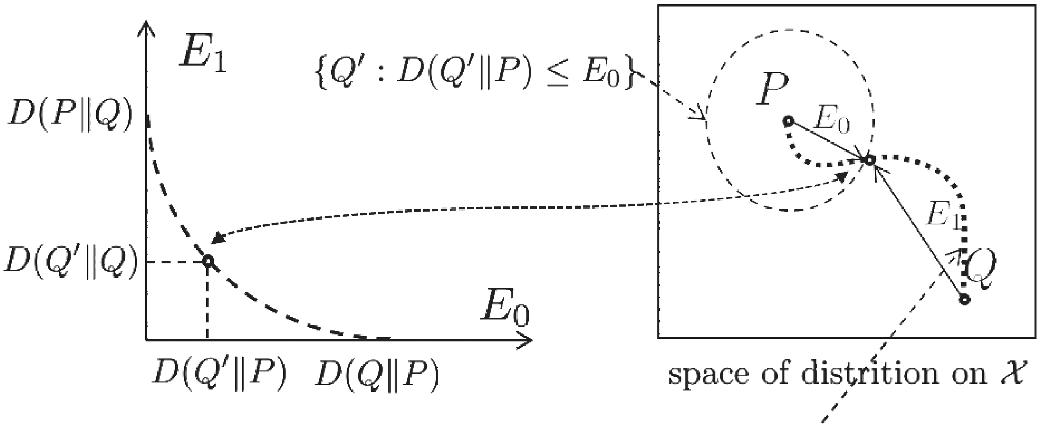
$$D(P_\lambda\|P) \le D(Q^*\|P) \tag{13.8}$$

$$D(P_\lambda\|Q) \le D(Q^*\|Q) \tag{13.9}$$

Thus, we immediately conclude that minimization in (13.3) can be restricted to Q^* belonging to the family of tilted distributions $\{P_\lambda, \lambda \in \mathbb{R}\}$. Furthermore, from (13.6) we also conclude that $\lambda \in [0,1]$. Hence, characterization of $E_1^*(E_0)$ given by (13.2) coincides with the one given by (13.3). □

[1] Small subtlety: In Corollary 12.1 we ask $\mathbb{E}_{Q^*}[T] \in (A,B)$. But A,B – the essential range of T – depend on the distribution under which the essential range is computed, cf. (12.10). Fortunately, we have $Q \ll P$ and $P \ll Q$, so essential range is the same under both P and Q. And furthermore (13.6) implies that $\mathbb{E}_{Q^*}[T] \in (A,B)$.

Note: Geometric interpretation of (13.3) is as follows: As λ increases from 0 to 1, or equivalently, θ increases from $-D(P\|Q)$ to $D(Q\|P)$, the optimal distribution traverses down the curve. This curve is in essense a geodesic connecting P to Q and exponents E_0,E_1 measure distances to P and Q. It may initially sound strange that the sum of distances to endpoints actually varies along the geodesic, but it is a natural phenomenon: just consider the unit circle with metric induced by the ambient Euclidean metric. Than if p and q are two antipodal points, the distance from intermediate point to endpoints do not sum up to $d(p,q) = 2$.



Non-linearity of the boundary corresponds to the scenario when the triangle inequality is not "="

$\forall$ distribution Q' in the tilted family, it minimizes E_0, E_1 simultaneously.
$\exists$ a unique optimal path from P to Q

13.3* Sequential Hypothesis Testing

Review: Filtrations, stopping times

- A sequence of nested σ-algebras $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_n \cdots \subset \mathcal{F}$ is called a filtration of $\mathcal{F}$.
- A random variable τ is called a stopping time of a filtration $\mathcal{F}_n$ if a) τ is valued in $\mathbb{Z}_+$ and b) for every $n \ge 0$ the event $\{\tau \le n\} \in \mathcal{F}_n$.
- The σ-algebra $\mathcal{F}_\tau$ consists of all events E such that $E \cap \{\tau \le n\} \in \mathcal{F}_n$ for all $n \ge 0$.
- When $\mathcal{F}_n = \sigma\{X_1, \dots, X_n\}$ the interpretation is that τ is a time that can be determined by causally observing the sequence X_j, and random variables measurable with respect to $\mathcal{F}_\tau$ are precisely those whose value can be determined on the basis of knowing $(X_1, \dots, X_\tau)$.
- Let M_n be a martingale adapted to $\mathcal{F}_n$, i.e. M_n is $\mathcal{F}_n$-measurable and $\mathbb{E}[M_n|\mathcal{F}_k] = M_{\min(n,k)}$. Then $\tilde{M}_n = M_{\min(n,\tau)}$ is also a martingale. If collection $\{M_n\}$ is uniformly integrable then
$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0]\,.$$
- For more details, see [Ç11, Chapter V].

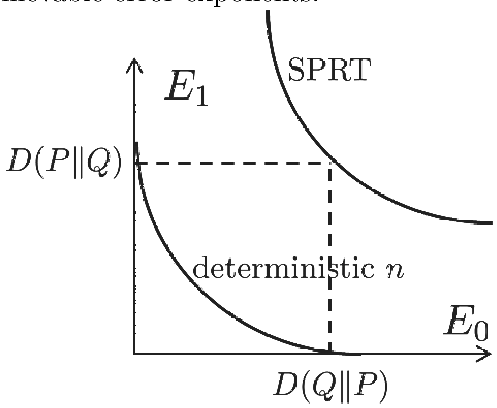
Different realizations of X_k are informative to different levels, the total "information" we receive follows a random process. Therefore, instead of fixing the sample size n, we can make n a stopping time τ, which gives a "better" (E_0, E_1) tradeoff. Solution is the concept of **sequential test**:

- Informally: Sequential test Z at each step declares either "H_0", "H_1" or "give me one more sample".
- Rigorous definition is as follows: A sequential hypothesis test is a stopping time τ of the filtration $\mathcal{F}_n \triangleq \sigma\{X_1, \dots, X_n\}$ and a random variable $Z \in \{0,1\}$ measurable with respect to $\mathcal{F}_\tau$.
- Each sequential test has the following performance metrics:

$$\alpha = \mathbb{P}[Z=0], \qquad \beta = \mathbb{Q}[Z=0] \tag{13.10}$$

$$l_0 = \mathbb{E}_{\mathbb{P}}[\tau], \qquad l_1 = \mathbb{E}_{\mathbb{Q}}[\tau] \tag{13.11}$$

The easiest way to see why sequential tests may be dramatically superior to fixed-sample size tests is the following example: Consider $P = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ and $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{-1}$. Since $P \not\perp Q$, we also have $P^n \not\perp Q^n$. Consequently, no finite-sample-size test can achieve zero error rates under both hypotheses. However, an obvious sequential test (wait for the first appearance of ±1) achieves zero error probability with finite average number of samples (2) under both hypotheses. This advantage is also seem very clearly in achievable error exponents.



Theorem 13.3. *Assume bounded LLR:*[2]

$$\left|\log \frac{P(x)}{Q(x)}\right| \le c_0, \forall x$$

where c_0 is some positive constant. If the error probabilities satisfy:

$$\pi_{1|0} \le 2^{-l_0 E_0}, \qquad \pi_{0|1} \le 2^{-l_1 E_1}$$

for large l_0, l_1, then the following inequality for the exponents holds

$$E_0 E_1 \le D(P\|Q) D(Q\|P).$$

[2] This assumption is satisfied for discrete distributions on finite spaces.

with optimal boundary achieved by the sequential probability ratio test SPRT(A, B) *(A, B are large positive numbers) defined as follows:*

$$\tau = \inf\{n : S_n \geq B \ \textit{or} \ S_n \leq -A\}$$
$$Z = \begin{cases} 0, & \textit{if } S_\tau \geq B \\ 1, & \textit{if } S_\tau < -A \end{cases}$$

where

$$S_n = \sum_{k=1}^{n} \log \frac{P(X_k)}{Q(X_k)}$$

is the log likelihood function of the first k observations.

Note: (Intuition on SPRT) Under the usual hypothesis testing setup, we collect n samples, evaluate the LLR S_n, and compare it to the threshold to give the optimal test. Under the sequential setup with iid data, $\{S_n : n \geq 1\}$ is a *random walk*, which has positive (resp. negative) drift $D(P\|Q)$ (resp. $-D(Q\|P)$) under the null (resp. alternative)! SPRT test simply declare P if the random walk crosses the upper boundary B, or Q if the random walk crosses the upper boundary $-A$.

Proof. As preparation we show two useful identities:

- For any stopping time with $\mathbb{E}_P[\tau] < \infty$ we have

$$\mathbb{E}_P[S_\tau] = \mathbb{E}_P[\tau] D(P\|Q) \tag{13.12}$$

  and similarly, if $\mathbb{E}_Q[\tau] < \infty$ then

$$\mathbb{E}_Q[S_\tau] = -\mathbb{E}_Q[\tau] D(Q\|P)\,.$$

  To prove these, notice that

$$M_n = S_n - nD(P\|Q)$$

  is clearly a martingale w.r.t. $\mathcal{F}_n$. Consequently,

$$\tilde{M}_n \triangleq M_{\min(\tau,n)}$$

  is also a martingale. Thus

$$\mathbb{E}[\tilde{M}_n] = \mathbb{E}[\tilde{M}_0] = 0\,,$$

  or, equivalently,

$$\mathbb{E}[S_{\min(\tau,n)}] = \mathbb{E}[\min(\tau, n)] D(P\|Q)\,. \tag{13.13}$$

  This holds for every $n \geq 0$. From boundedness assumption we have $|S_n| \leq nc$ and thus $|S_{\min(n,\tau)}| \leq n\tau$, implying that collection $\{S_{\min(n,\tau)}, n \geq 0\}$ is uniformly integrable. Thus, we can take $n \to \infty$ in (13.13) and interchange expectation and limit safely to conclude (13.12).

- Let τ be a stopping time. The Radon-Nikodym derivative of $\mathbb{P}$ w.r.t. $\mathbb{Q}$ on σ-algebra $\mathcal{F}_\tau$ is given by

$$\frac{d\mathbb{P}|_{\mathcal{F}_\tau}}{d\mathbb{Q}|_{\mathcal{F}_\tau}} = \exp\{S_\tau\}\,.$$

  Indeed, what we need to verify is that for every event $E \in \mathcal{F}_\tau$ we have

$$\mathbb{E}_P[1_E] = \mathbb{E}_Q[\exp\{S_\tau\} 1_E] \tag{13.14}$$

To that end, consider a decomposition

$$1_E = \sum_{n\ge 0} 1_{E\cap\{\tau=n\}}\,.$$

By monotone convergence theorem applied to (13.14) it is sufficient to verify that for every n

$$\mathbb{E}_P[1_{E\cap\{\tau=n\}}] = \mathbb{E}_Q[\exp\{S_\tau\}1_{E\cap\{\tau=n\}}]\,. \tag{13.15}$$

This, however, follows from the fact that $E\cap\{\tau=n\}\in\mathcal{F}_n$ and $\frac{d\mathbb{P}|_{\mathcal{F}_n}}{d\mathbb{Q}|_{\mathcal{F}_n}} = \exp\{S_n\}$ by the very definition of S_n.

We now proceed to the proof. For **achievability** we apply (13.14) to infer

$$\begin{aligned}\pi_{1|0} &= \mathbb{P}[S_\tau \le -A]\\ &= \mathbb{E}_Q[\exp\{S_\tau\}1\{S_\tau\le -A\}]\\ &\le e^{-A}\end{aligned}$$

Next, we denot $\tau_0 = \inf\{n : S_n \ge B\}$ and observe that $\tau \le \tau_0$, whereas expectation of τ_0 we estimate from (13.12):

$$\mathbb{E}_P[\tau] \le \mathbb{E}_P[\tau_0] = \mathbb{E}_P[S_{\tau_0}] \le B + c_0\,,$$

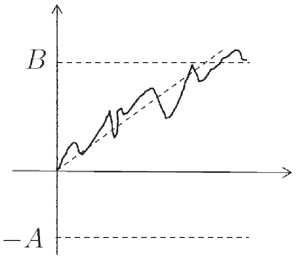
where in the last step we used the boundedness assumption to infer

$$S_{\tau_0} \le B + c_0$$

Thus

$$l_0 = \mathbb{E}_{\mathbb{P}}[\tau] \le \mathbb{E}_{\mathbb{P}}[\tau_0] \le \frac{B+c_0}{D(P\|Q)} \approx \frac{B}{D(P\|Q)} \text{ for large } B$$

Similarly we can show $\pi_{0|1}\le e^{-B}$ and $l_1 \le \frac{A}{D(Q\|P)}$ for large A. Take $B = l_0 D(P\|Q)$, $A = l_1 D(Q\|P)$, this shows the achievability.



under P, $S_n - nD(P\|Q)$ is a martingale

Converse: Assume (E_0, E_1) achievable for large l_0, l_1 and apply data processing inequality of divergence:

$$\begin{aligned}d(\mathbb{P}(Z=1)\|\mathbb{Q}(Z=1)) &\le D(\mathbb{P}\|\mathbb{Q})\Big|_{\mathcal{F}_\tau}\\ &= \mathbb{E}_P[S_\tau] \qquad\qquad = \mathbb{E}_{\mathbb{P}}[\tau]D(P\|Q) \quad \text{from (13.12)}\\ &= l_0 D(P\|Q)\end{aligned}$$

notice that for l_0E_0 and l_1E_1 large, we have $d(\mathbb{P}(Z=1)\|\mathbb{Q}(Z=1)) \approx l_1E_1$, therefore $l_1E_1 \lesssim l_0D(P\|Q)$. Similarly we can show that $l_0E_0 \lesssim l_1D(Q\|P)$, finally we have

$$E_0E_1 \le D(P\|Q)D(Q\|P), \text{ as } l_0, l_1 \to \infty$$

□

Part IV

Channel coding

§ 14. Channel coding

Objects of study so far:

1. P_X - Single distribution, Compression
2. P_X vs Q_X - Comparing two distributions, Hypothesis testing
3. Now: $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ (called a *random transformation*) - A collection of distributions

14.1 Channel Coding

Definition 14.1. An M-code for $P_{Y|X}$ is an encoder/decoder pair (f, g) of (randomized) functions[1]

- encoder $f : [M] \to \mathcal{X}$
- decoder $g : \mathcal{Y} \to [M] \cup \{\mathsf{e}\}$

Notation: $[M] \triangleq \{1, \ldots, M\}$.

In most cases f and g are deterministic functions, in which case we think of them (equivalently) in terms of codewords, codebooks, and decoding regions

- $\forall i \in [M] : c_i = f(i)$ are *codewords*, the collection $\mathcal{C} = \{c_1, \ldots, c_M\}$ is called a *codebook*.
- $\forall i \in [M], D_i = g^{-1}(\{i\})$ is the *decoding region* for i.

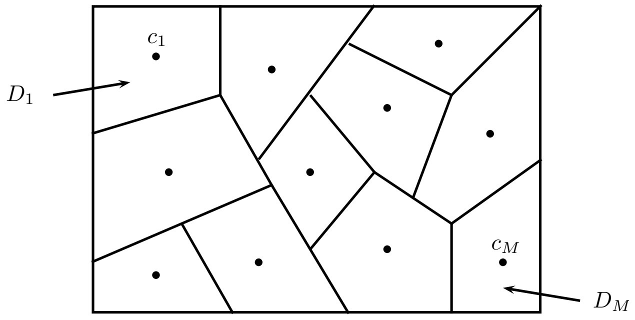


Figure 14.1: When $\mathcal{X} = \mathcal{Y}$, the decoding regions can be pictured as a partition of the space, each containing one codeword.

Note: The underlying probability space for channel coding problems will always be

$$W \xrightarrow{f} X \xrightarrow{P_{Y|X}} Y \xrightarrow{g} \hat{W}$$

[1] For randomized encoder/decoders, we identify f and g as probability transition kernels $P_{X|W}$ and $P_{\hat{W}|Y}$.

When the source alphabet is $[M]$, the joint distribution is given by:

$$\text{(general) } P_{WXY\hat{W}}(m,a,b,\hat{m}) = \frac{1}{M} P_{X|W}(a|m) P_{Y|X}(b|a) P_{\hat{W}|Y}(\hat{m}|b)$$
$$\text{(deterministic } f,g\text{) } P_{WXY\hat{W}}(m,c_m,b,\hat{m}) = \frac{1}{M} P_{Y|X}(b|c_m) \mathbf{1}\{b \in D_{\hat{m}}\}$$

Throughout the notes, these quantities will be called:

- W - Original message
- X - (Induced) Channel input
- Y - Channel output
- $\hat{W}$ - Decoded message

14.1.1 Performance Metrics

Three ways to judge the quality of a code in terms of error probability:

1. $P_e \triangleq \mathbb{P}[W \neq \hat{W}]$ - Average error probability.
2. $P_{e,\max} \triangleq \max_{m \in [M]} \mathbb{P}[\hat{W} \neq m | W = m]$ - Maximum error probability.
3. In the special case when $M = 2^k$, think of $W = S^k \in \mathbb{F}_2^k$ as a length k bit string. Then the bit error rate is $P_b \triangleq \frac{1}{k}\sum_{j=1}^k \mathbb{P}[S_j \neq \hat{S}_j]$, which means the average fraction of errors in a k-bit block. It is also convenient to introduce in this case the Hamming distance

   $$d_H(S^k, \hat{S}^k) \triangleq \#\{i : S_i \neq \hat{S}_j\}\,.$$

   Then, the bit-error rate becomes the normalized expected Hamming distance:

   $$P_b = \frac{1}{k}\mathbb{E}[d_H(S^k, \hat{S}^k)]\,.$$

   To distinguish the bit error rate P_b from the previously defined P_e and $P_{e,\max}$, we will also call the latter the average (resp. max) block error rate.

The most typical metric is average probability of error, but the others will be used occasionally in the course as well. By definition, $P_e \le P_{e,\max}$. Therefore maximum error probability is a more stringent criterion which offers uniform protection for all codewords.

14.1.2 Fundamental Limit of $P_{Y|X}$

Definition 14.2. A code (f,g) is an (M,ϵ)-code for $P_{Y|X}$ if $f : [M] \to \mathcal{X}$, $g : \mathcal{Y} \to [M] \cup \{e\}$, and $P_e \le \epsilon$. Similarly, an $(M,\epsilon)_{\max}$-code must satisfy $P_{e,\max} \le \epsilon$.

Then the fundamental limits of channel codes are defined as

$$M^*(\epsilon) = \max\{M : \exists (M,\epsilon) - code\}$$
$$M^*_{\max}(\epsilon) = \max\{M : \exists (M,\epsilon)_{\max} - code\}$$

Remark: $\log_2 M^*$ gives the maximum number of bits that we can pump through a channel $P_{Y|X}$ while still having the error probability (in the appropriate sense) at most ϵ.

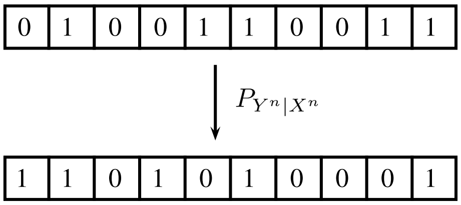
Example: The random transformation BSC(n, δ) (binary symmetric channel) is defined as

$$\mathcal{X} = \{0,1\}^n$$
$$\mathcal{Y} = \{0,1\}^n$$

where the input X^n is contaminated by additive noise $Z^n \perp\!\!\!\perp X^n$ and the channel outputs

$$Y^n = X^n \oplus Z^n$$

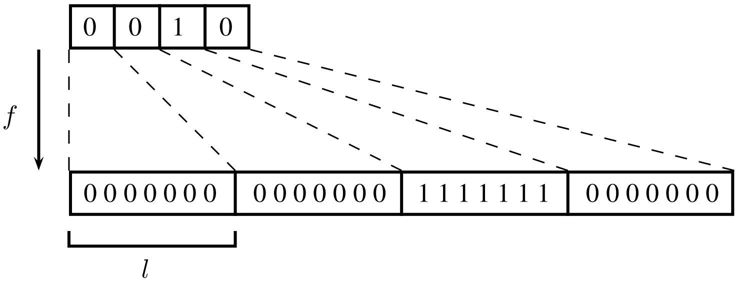
where $Z^n \overset{\text{i.i.d.}}{\sim} \text{Bern}(\delta)$. Pictorially, the BSC$(n, \delta)$ channel takes a binary sequence length n and flips the bits independently with probability δ:



Question: When $\delta = .11$, $n = 1000$, what is the max number of bits you can send with $P_e \leq 10^{-3}$?
Ideas:

0. Can one send 1000 bits with $P_e \leq 10^{-3}$? No and apparently the probability that at least one bit is flipped is $P_e = 1 - (1-\delta)^n \approx 1$. This implies that uncoded transmission does not meet our objective and coding is necessary – tradeoff: reduce number of bits to send, increase probability of success.

1. Take each bit and repeat it l times (l-repetition code).



   With majority decoding, the probability of error of this scheme is $P_e \approx k\mathbb{P}[\text{Binom}(l, \delta) > l/2]$ and $kl \leq n = 1000$, which for $P_e \leq 10^{-3}$ gives $l = 21$, $k = 47$ bits.

2. Reed-Muller Codes $(1, r)$. Interpret a message $a_0, \dots, a_{r-1} \in \mathbb{F}_2^r$ as the polynomial (in this case, a degree-1 and $(r-1)$-variate polynomial) $\sum_{i=1}^{r-1} a_i x_i + a_0$, then codewords are formed by evaluating the polynomial at all possible $x^{r-1} \in \mathbb{F}_2^{r-1}$. This code, which maps r bits to 2^{r-1} bits, has minimum distance 2^{r-2}. For $r = 7$, there is a $[64, 7, 32]$ Reed-Muller code and it can be shown that the MAP decoder of this code passed over the $BSC(n = 64, \delta = 0.11)$ achieves probability of error $\leq 6 \cdot 10^{-6}$. Thus, we can use 16 such blocks (each carrying 7 data bits and occupying 64 bits on the channel) over the BSC$(1024, \delta)$, and still have (union bound) overall $P_e \lesssim 10^{-4}$. This allows us to send $7 \cdot 16 = 112$ bits in 1024 channel uses, more than double that of the repetition code.

3. Shannon's theorem (to be shown soon) tells us that over memoryless channel of blocklength n the fundamental limit satisfies
$$\log M^* = nC + o(n) \tag{14.1}$$
as $n \to \infty$ and for arbitrary $\epsilon \in (0,1)$. Here $C = \max_X I(X_1;Y_1)$ is the capacity of the single-letter channel. In our case we have
$$I(X;Y) = \max_{P_X} I(X;X+Z) = \log 2 - h(\delta) \approx \frac{1}{2} \text{ bit}$$
So Shannon's expansion (14.1) can be used to predict (non-rigorously, of course) that it should be possible to send around 500 bits reliably. As it turns out, for this blocklength this is not quite possible.

4. Even though calculating $\log M^*$ is not computationally feasible (searching over all codebooks is doubly exponential in block length n), we can find bounds on $\log M^*$ that are easy to compute. We will show later in the course that in fact, for BSC(1000, .11)
$$414 \le \log M^* \le 416$$

5. The first codes to approach the bounds on $\log M^*$ are called *Turbo codes* (after the turbocharger engine – where exhaust is fed back in to power the engine). This class of codes is known as *sparse graph codes*, of which LDPC codes are particularly well studied. As a rule of thumb, these codes typically approach $80\ldots90\%$ of $\log M^*$ when $n \approx 10^3 \ldots 10^4$.

14.2 Basic Results

Recall that the object of our study is $M^*(\epsilon) = \max\{M : \exists (M,\epsilon) - \mathit{code}\}$.

14.2.1 Determinism

1. Given any encoder $f : [M] \to \mathcal{X}$, the decoder that minimizes P_e is the *Maximum A Posteriori (MAP)* decoder, or equivalently, the *Maximal Likelihood (ML)* decoder, since the codewords are equiprobable:
$$\begin{aligned} g^*(y) &= \operatorname*{argmax}_{m\in[M]} \mathbb{P}[W = m|Y = y] \\ &= \operatorname*{argmax}_{m\in[M]} \mathbb{P}[Y = y|W = m] \end{aligned}$$
Furthermore, for a fixed f, the MAP decoder g is deterministic

2. For given M, $P_{Y|X}$, the P_e-minimizing encoder is deterministic.

Proof. Let $f : [M] \to \mathcal{X}$ be a random transformation. We can always represent randomized encoder as deterministic encoder with auxiliary randomness. So instead of $f(a|m)$, consider the deterministic encoder $\tilde{f}(m,u)$, that receives external randomness u. Then looking at all possible values of the randomness,
$$P_e = P[W \neq \hat{W}] = \mathbb{E}_U[\mathbb{P}[W \neq \hat{W}|U] = \mathbb{E}_U[P_e(U)]$$
Each u in the expectation gives a deterministic encoder, hence there is a deterministic encoder that is at least as good as the average of the collection, i.e., $\exists u_0$ s.t. $P_e(u_0) \le \mathbb{P}[W \neq \hat{W}]$ □

Remark: If instead we use maximal probability of error as our performance criterion, then these results don't hold; randomized encoders and decoders may improve performance. Example: consider $M = 2$ and we are back to the binary hypotheses testing setup. The optimal decoder (test) that minimizes the maximal Type-I and II error probability, i.e., $\max\{1-\alpha,\beta\}$, is not deterministic, if $\max\{1-\alpha,\beta\}$ is not achieved at a vertex of the region $\mathcal{R}(P,Q)$.

14.2.2 Bit Error Rate vs Block Error Rate

Now we give a bound on the average probability of error in terms of the bit error probability.

Theorem 14.1. *For all* (f,g), $M = 2^k \implies P_b \le P_e \le kP_b$

Remark: The most often used direction $P_b \ge \frac{1}{k}P_e$ is rather loose for large k.

Proof. Recall that $M = 2^k$ gives us the interpretation of $W = S^k$ sequence of bits.

$$\frac{1}{k}\sum_{i=1}^{k}\mathbf{1}\{S_i \ne \hat{S}_i\} \le \mathbf{1}\{S^k \ne \hat{S}^k\} \le \sum_{i=1}^{k}\mathbf{1}\{S_i \ne \hat{S}_i\}$$

Where the first inequality is obvious and the second follow from the union bound. Taking expectation of the above expression gives the theorem. □

Theorem 14.2 (Assouad). *If* $M = 2^k$ *then*

$$P_b \ge \min\{\mathbb{P}[\hat{W} = c'|W = c] : c, c' \in \mathbb{F}_2^k, d_H(c,c') = 1\}\,.$$

Proof. Let e_i be a length k vector that is 1 in the i-th position, and zero everywhere else. Then

$$\sum_{i=1}^{k}\mathbf{1}\{S_i \ne \hat{S}_i\} \ge \sum_{i=1}^{k}\mathbf{1}\{S^k = \hat{S}^k + e_i\}$$

Dividing by k and taking expectation gives

$$\begin{aligned} P_b &\ge \frac{1}{k}\sum_{i=1}^{k}\mathbb{P}[S^k = \hat{S}^k + e_i] \\ &\ge \min\{\mathbb{P}[\hat{W} = c'|W = c] : c, c' \in \mathbb{F}_2^k, d_H(c,c') = 1\}\,. \end{aligned}$$

□

Similarly, we can prove the following generalization:

Theorem 14.3. *If* $A, B \in \mathbb{F}_2^k$ *(with arbitrary marginals!) then for every* $r \ge 1$ *we have*

$$P_b = \frac{1}{k}\,\mathbb{E}[d_H(A,B)] \ge \binom{k-1}{r-1}P_{r,\min} \tag{14.2}$$

$$P_{r,\min} \triangleq \min\{\mathbb{P}[B = c'|A = c] : c, c' \in \mathbb{F}_2^k, d_H(c,c') = r\} \tag{14.3}$$

Proof. First, observe that

$$\mathbb{P}[d_H(A,B) = r|A = a] = \sum_{b: d_H(a,b)=r} P_{B|A}(b|a) \ge \binom{k}{r}P_{r,\min}\,.$$

Next, notice

$$d_H(x,y) \ge r1\{d_H(x,y) = r\}$$

and take the expectation with $x \sim A$, $y \sim B$. □

Remark: In statistics, Assouad's Lemma is a useful tool for obtaining lower bounds on the minimax risk of an estimator. Say the data X is distributed according to P_θ parameterized by $\theta \in \mathbb{R}^k$ and let $\hat{\theta} = \hat{\theta}(X)$ be an estimator for θ. The goal is to minimize the maximal risk $\sup_{\theta\in\Theta} \mathbb{E}_\theta[\|\theta - \hat{\theta}\|_1]$. A lower bound (Bayesian) to this worst-case risk is the average risk $\mathbb{E}[\|\theta - \hat{\theta}\|_1]$, where θ is distributed to any prior. Consider θ uniformly distributed on the hypercube $\{0, \epsilon\}^k$ with side length ϵ embedded in the space of parameters. Then

$$\inf_{\hat{\theta}} \sup_{\theta\in\{0,\epsilon\}^k} \mathbb{E}[\|\theta - \hat{\theta}\|_1] \geq \frac{k\epsilon}{4} \min_{d_H(\theta,\theta')=1} (1 - \mathrm{TV}(P_\theta, P_{\theta'}))\,. \tag{14.4}$$

This can be proved using similar ideas to Theorem 14.2. WLOG, assume that $\epsilon = 1$.

$$\begin{aligned}
\mathbb{E}[\|\theta - \hat{\theta}\|_1] &\overset{(a)}{\geq} \frac{1}{2}\mathbb{E}[\|\theta - \hat{\theta}_{dis}\|_1] = \frac{1}{2}\mathbb{E}[d_H(\theta, \hat{\theta}_{dis})] \\
&\geq \frac{1}{2}\sum_{i=1}^{k} \min_{\hat{\theta}_i = \hat{\theta}_i(X)} \mathbb{P}[\theta_i \neq \hat{\theta}_i] \overset{(b)}{=} \frac{1}{4}\sum_{i=1}^{k}(1 - \mathrm{TV}(P_{X|\theta_i=0}, P_{X|\theta_i=1})) \\
&\overset{(c)}{\geq} \frac{k}{4} \min_{d_H(\theta,\theta')=1} (1 - \mathrm{TV}(P_\theta, P_{\theta'}))\,.
\end{aligned}$$

Here $\hat{\theta}_{dis}$ is the discretized version of $\hat{\theta}$, i.e. the closest point on the hypercube to $\hat{\theta}$ and so (a) follows from $|\theta_i - \hat{\theta}_i| \geq \frac{1}{2}\mathbf{1}_{\{|\theta_i - \hat{\theta}_i| > 1/2\}} = \frac{1}{2}\mathbf{1}_{\{\theta_i \neq \hat{\theta}_{dis,i}\}}$, (b) follows from the optimal binary hypothesis testing for θ_i given X, (c) follows from the convexity of TV: $\mathrm{TV}(P_{X|\theta_i=0}, P_{X|\theta_i=1}) = \mathrm{TV}(\frac{1}{2^{k-1}}\sum_{\theta:\theta_i=0} P_{X|\theta}, \frac{1}{2^{k-1}}\sum_{\theta:\theta_i=1} P_{X|\theta}) \leq \frac{1}{2^{k-1}}\sum_{\theta:\theta_i=0}\mathrm{TV}(P_{X|\theta}, P_{X|\theta\oplus e_i}) \leq \max_{d_H(\theta,\theta')=1}\mathrm{TV}(P_\theta, P_{\theta'})$. Alternatively, (c) also follows from by providing the extra information $\theta^{\backslash i}$ and allowing $\hat{\theta}_i = \hat{\theta}_i(X, \theta^{\backslash i})$ in the second line.

14.3 General (Weak) Converse Bounds

Theorem 14.4 (Weak converse)**.**

1. *Any M-code for $P_{Y|X}$ satisfies*

$$\log M \leq \frac{\sup_X I(X;Y) + h(P_e)}{1 - P_e}$$

2. *When $M = 2^k$*

$$\log M \leq \frac{\sup_X I(X;Y)}{\log 2 - h(P_b)}$$

Proof. **(1)** Since $W \to X \to Y \to \hat{W}$, we have the following chain of inequalities, cf. Fano's inequality Theorem 5.4:

$$\begin{aligned}
\sup_X I(X;Y) \geq I(X;Y) \geq I(W;\hat{W}) \\
&\geq d(\mathbb{P}[W = \hat{W}] \| \frac{1}{M}) \\
&\geq -h(\mathbb{P}[W \neq \hat{W}]) + \mathbb{P}[W = \hat{W}]\log M
\end{aligned}$$

Plugging in $P_e = \mathbb{P}[W \neq \hat{W}]$ finishes the first proof.

(2) Now $S^k \to X \to Y \to \hat{S}^k$. Recall from Theorem 5.1 that for iid S^n, $\sum I(S_i;\hat{S}_i) \le I(S^k;\hat{S}^k)$. This gives us

$$\begin{aligned}
\sup_X I(X;Y) \ge I(X;Y) &\ge \sum_{i=1}^k I(S_i,\hat{S}_i) \\
&\ge k\frac{1}{k}\sum d\left(\mathbb{P}[S_i = \hat{S}_i]\middle\|\frac{1}{2}\right) \\
&\ge kd\left(\frac{1}{k}\sum_{i=1}^k \mathbb{P}[S_i = \hat{S}_i]\middle\|\frac{1}{2}\right) \\
&= kd\left(1 - P_b\middle\|\frac{1}{2}\right) = k(\log 2 - h(P_b))
\end{aligned}$$

where the second line used Fano's inequality (Theorem 5.4) for binary random variable (or divergence data processing), and the third line used the convexity of divergence. □

14.4 General achievability bounds: Preview

Remark: Regarding differences between information theory and statistics: in statistics, there is a parametrized set of distributions on a space (determined by the model) from which we try to estimate the underlying distribution from samples. In data transmission, the challenge is to *choose* the structure on the parameter space (channel coding) such that, upon observing a sample, we can estimate the correct parameter with high probability. With this in mind, it is natural to view

$$\log \frac{P_{Y|X=x}}{P_Y}$$

as an LLR of a binary hypothesis test, where we compare the hypothesis $X = x$ to the distribution induced by our codebook: $P_Y = P_{Y|X} \circ P_X$ (so compare c_i to "everything else"). To decode, we ask M different questions of this form. This motivates importance of the random variable (called *information density*):

$$i(X;Y) = \log \frac{P_{Y|X}(Y|X)}{P_Y(Y)}$$

, where $P_Y = P_{Y|X} \circ P_X$. (Note: $I(X;Y) = \mathbb{E}[i(X;Y)]$).

Shortly, we will show a result (**Shannon's Random Coding Theorem**), that states: $\forall P_X$, $\forall \tau$, $\exists (M,\epsilon) -$ *code* with

$$\epsilon \le \mathbb{P}[i(X;Y) \le \log M + \tau] + e^{-\tau}$$

Details in the next lecture.

Notation: in the following proofs, we shall make use of the *independent pairs* $(X,Y) \perp\!\!\!\perp (\overline{X},\overline{Y})$

$$\begin{aligned} X &\to Y \quad (X: \text{ sent codeword}) \\ \overline{X} &\to \overline{Y} \quad (\overline{X}: \text{ unsent codeword}) \end{aligned}$$

The joint distribution is given by:

$$P_{XY\overline{XY}}(a,b,\overline{a},\overline{b}) = P_X(a)P_{Y|X}(b|a)P_X(\overline{a})P_{Y|X}(\overline{b}|\overline{a}).$$

15.1 Information density

Definition 15.1 (Information density)**.** Given joint distribution $P_{X,Y}$ we define

$$i_{P_{XY}}(x;y) = \log\frac{P_{Y|X}(y|x)}{P_Y(y)} = \log\frac{dP_{Y|X=x}(y)}{dP_Y(y)} \tag{15.1}$$

and we define $i_{P_{XY}}(x;y) = +\infty$ for all y in the singular set where $P_{Y|X=x}$ is not absolutely continuous w.r.t. P_Y. We also define $i_{P_{XY}}(x;y) = -\infty$ for all y such that $dP_{Y|X=x}/dP_Y$ equals zero. We will almost always abuse notation and write $i(x;y)$ dropping the subscript $P_{X,Y}$, assuming that the joint distribution defining $i(\cdot;\cdot)$ is clear from the context.
Notice that $i(x;y)$ depends on the underlying P_X and $P_{Y|X}$, which should be understood from the context.

Remark 15.1 (Intuition)**.** Information density is a useful concept in understanding decoding. In discriminating between two codewords, one concerns with (as we learned in binary hypothesis testing) the LLR, $\log\frac{P_{Y|X=c_1}}{P_{Y|X=c_2}}$. In M-ary hypothesis testing, a similar role is played by information density $i(c_1;y)$, which, loosely speaking, evaluates the likelihood of c_1 against the average likelihood, or "everything else", which we model by P_Y.

Remark 15.2 (Joint measurability)**.** There is a measure-theoretic subtlety in (15.1): The so-defined function $i(\cdot;\cdot)$ may not be a measurable function on the product space $\mathcal{X}\times\mathcal{Y}$. For resolution, see Section 2.6* and Remark 2.4 in particular.

Remark 15.3 (Alternative definition)**.** Observe that for discrete $\mathcal{X},\mathcal{Y}$, (15.1) is equivalently written as

$$i(x;y) = \log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} = \log\frac{P_{X|Y}(x|y)}{P_X(x)}$$

For the continuous case, people often use the alternative definition, which is symmetric in X and Y and is measurable w.r.t. $\mathcal{X}\times\mathcal{Y}$:

$$i(x;y) = \log\frac{dP_{X,Y}}{dP_X\times P_Y}(x,y) \tag{15.2}$$

Notice a subtle difference between (15.1) and (15.2) for the continuous case: In (15.2) the Radon-Nikodym derivative is only defined up to sets of measure zero, therefore whenever $P_X(x) = 0$ the value of $P_Y(i(x,Y) > t)$ is undefined. This problem does not occur with definition (15.1), and that is why we prefer it. In any case, for discrete $\mathcal{X}$, $\mathcal{Y}$, or under other regularity conditions, all the definitions are equivalent.

Proposition 15.1 (Properties of information density)**.**

1. *$\mathbb{E}[i(X;Y)] = I(X;Y)$. This justifies the name "(mutual) information density".*

2. *If there is a bijective transformation $(X,Y) \to (A,B)$, then almost surely $i_{P_{XY}}(X;Y) = i_{P_{AB}}(A;B)$ and in particular, distributions of $i(X;Y)$ and $i(A;B)$ coincide.*

3. *(Conditioning and unconditioning trick) Suppose that $f(y) = 0$ and $g(x) = 0$ whenever $i(x;y) = -\infty$, then*[1]

$$\mathbb{E}[f(Y)] = \mathbb{E}[\exp\{-i(x;Y)\}f(Y)|X = x], \forall x \tag{15.3}$$
$$\mathbb{E}[g(X)] = \mathbb{E}[\exp\{-i(X;y)\}g(X)|Y = y], \forall y \tag{15.4}$$

4. *Suppose that $f(x,y) = 0$ whenever $i(x;y) = -\infty$, then*

$$\mathbb{E}[f(\overline{X},Y)] = \mathbb{E}[\exp\{-i(X;Y)\}f(X,Y)] \tag{15.5}$$
$$\mathbb{E}[f(X,\overline{Y})] = \mathbb{E}[\exp\{-i(X;Y)\}f(X,Y)] \tag{15.6}$$

Proof. The proof is simply change of measure. For example, to see (15.3), note

$$\mathbb{E}f(Y) = \sum_{y\in\mathcal{Y}} P_Y(y)f(y) = \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\frac{P_Y(y)}{P_{Y|X}(y|x)}f(y)$$

notice that by the assumption on $f(\cdot)$, the summation is valid even if for some y we have that $P_{Y|X}(y|x) = 0$. Similarly, $\mathbb{E}[f(x,Y)] = \mathbb{E}[\exp\{-i(x;Y)\}f(x,Y)|X = x]$. Integrating over $x \sim P_X$ gives (15.5). The rest are by interchanging X and Y. □

Corollary 15.1.

$$\mathbb{P}[i(x;Y) > t] \le \exp(-t) \tag{15.7}$$
$$\mathbb{P}[i(\overline{X};Y) > t] \le \exp(-t) \tag{15.8}$$

Proof. Pick $f(Y) = \mathbf{1}\{i(x;Y) > t\}$ in (15.3). □

Remark 15.4. We have used this trick before: For any probability measure P and any measure Q,

$$Q\Big[\log\frac{\mathrm{d}P}{\mathrm{d}Q} \ge t\Big] \le \exp(-t). \tag{15.9}$$

for example, in hypothesis testing (Corollary 10.1). In data compression, we frequently used the fact that $|\{x : \log P_X(x) \ge t\}| \le \exp(-t)$, which is also of the form (15.9) with Q = counting measure.

[1] Note that (15.3) holds when $i(x;y)$ is defined as $i = \log\frac{dP_{Y|X}}{P_Y}$, and (15.4) holds when $i(x;y)$ is defined as $i = \log\frac{dP_{X|Y}}{P_X}$. (15.5) and (15.6) hold under either of the definitions. Since in the following we shall only make use of (15.3) and (15.5), this is another reason we adopted definition (15.1).

15.2 Shannon's achievability bound

Theorem 15.1 (Shannon's achievability bound)**.** *For a given $P_{Y|X}$, $\forall P_X$, $\forall \tau > 0$, $\exists (M, \epsilon)$-code with*

$$\epsilon \leq \mathbb{P}[i(X;Y) \leq \log M + \tau] + \exp(-\tau). \tag{15.10}$$

Proof. Recall that for a given codebook $\{c_1, \ldots, c_M\}$, the optimal decoder is MAP, or equivalently, ML, since the codewords are equiprobable:

$$\begin{aligned} g^*(y) &= \operatorname*{argmax}_{m \in [M]} P_{X|Y}(c_m|y) \\ &= \operatorname*{argmax}_{m \in [M]} P_{Y|X}(y|c_m) \\ &= \operatorname*{argmax}_{m \in [M]} i(c_m; y). \end{aligned}$$

The step of selecting the maximum likelihood can make analyzing the error probability difficult. Similar to what we did in almost loss compression (e.g., Theorem 7.4), the magic in showing the following two achievability bounds is to consider a suboptimal decoder. In Shannon's bound, we consider a threshold-based suboptimal decoder $g(y)$ as follows:

$$g(y) = \begin{cases} m, & \exists! c_m \text{ s.t. } i(c_m; y) \geq \log M + \tau \\ \mathsf{e}, & \text{o.w.} \end{cases}$$

Interpretation: $i(c_m; y) \geq \log M + \tau \Leftrightarrow P_{X|Y}(c_m|y) \geq M \exp(\tau) P_X(c_m)$, i.e., the likelihood of c_m being the transmitted codeword conditioned on receiving y exceeds some threshold.

For a given codebook $(c_1, \ldots, c_M)$, the error probability is:

$$P_e(c_1, \ldots, c_M) = \mathbb{P}[\{i(c_W; Y) \leq \log M + \tau\} \cup \{\exists \overline{m} \neq W, i(c_{\overline{m}}; Y) > \log M + \tau\}]$$

where W is uniform on $[M]$.
We generate the codebook $(c_1, \ldots, c_M)$ randomly with $c_m \sim P_X$ i.i.d. for $m \in [M]$. By symmetry, the error probability averaging over all possible codebooks is given by:

$$\begin{aligned} & \mathbb{E}[P_e(c_1, \ldots, c_M)] \\ &= \mathbb{E}[P_e(c_1, \ldots, c_M)|W = 1] \\ &= \mathbb{P}[\{i(c_1; Y) \leq \log M + \tau\} \cup \{\exists \overline{m} \neq 1, i(c_{\overline{m}}, Y) > \log M + \tau\}|W = 1] \\ &\leq \mathbb{P}[i(c_1; Y) \leq \log M + \tau | W = 1] + \sum_{m=2}^{M} \mathbb{P}[i(c_m; Y) > \log M + \tau | W = 1] && \text{(union bound)} \\ &= \mathbb{P}\left[i(X;Y) \leq \log M + \tau\right] + (M-1)\mathbb{P}\left[i(\overline{X}; Y) > \log M + \tau\right] && \text{(random codebook)} \\ &\leq \mathbb{P}\left[i(X;Y) \leq \log M + \tau\right] + (M-1)\exp(-(\log M + \tau)) && \text{(by Corollary 15.1)} \\ &\leq \mathbb{P}\left[i(X;Y) \leq \log M + \tau\right] + \tau) + \exp(-\tau) \end{aligned}$$

Finally, since the error probability averaged over the random codebook satisfies the upper bound, there must exist some code allocation whose error probability is no larger than the bound. □

Remark 15.5 (Typicality)**.**

- The property of a pair (x,y) satisfying the condition $\{i(x;y) \geq \gamma\}$ can be interpreted as "**joint typicality**". Such version of joint typicality is useful when random coding is done in product spaces with $c_j \sim P_X^n$ (i.e. coordinates of the codeword are iid).
- A popular alternative to the definition of typicality is to require that the empirical joint distribution is close to the true joint distribution, i.e., $\hat{P}_{x^n,y^n} \approx P_{XY}$, where

$$\hat{P}_{x^n,y^n}(a,b) = \frac{1}{n} \cdot \#\{j : x_j = a, y_j = b\}\,.$$

This definition is natural for cases when random coding is done with $c_j \sim$ uniform on the set $\{x^n : \hat{P}_{x^n} \approx P_X\}$ (type class).

15.3 Dependence-testing bound

Theorem 15.2 (DT bound). *$\forall P_X$, $\exists(M,\epsilon)$-code with*

$$\epsilon \leq \mathbb{E}\left[\exp\left\{-\left(i(X;Y) - \log\frac{M-1}{2}\right)^+\right\}\right] \tag{15.11}$$

where $x^+ \triangleq \max(x,0)$.

Proof. For a fixed γ, consider the following suboptimal decoder:

$$g(y) = \begin{cases} m, & \text{for the smallest } m \text{ s.t. } i(c_m;y) \geq \gamma \\ \mathsf{e}, & \text{o/w} \end{cases}$$

Note that given a codebook $\{c_1,\ldots,c_M\}$, we have by union bound

$$\begin{aligned}\mathbb{P}[\hat{W} \neq j|W=j] &= \mathbb{P}[i(c_j;Y) \leq \gamma|W=j] + \mathbb{P}[i(c_j;Y) > \gamma, \exists k \in [j-1], \text{ s.t. } i(c_k;Y) > \gamma] \\ &\leq \mathbb{P}[i(c_j;Y) \leq \gamma|W=j] + \sum_{k=1}^{j-1}\mathbb{P}[i(c_k;Y) > \gamma|W=j].\end{aligned}$$

Averaging over the randomly generated codebook, the expected error probability is upper bounded by:

$$\begin{aligned}\mathbb{E}[P_e(c_1,\ldots,c_M)] &= \frac{1}{M}\sum_{j=1}^{M}\mathbb{P}[\hat{W} \neq j|W=j] \\ &\leq \frac{1}{M}\sum_{j=1}^{M}\Big(\mathbb{P}[i(X;Y) \leq \gamma] + \sum_{k=1}^{j-1}\mathbb{P}[i(\overline{X};Y) > \gamma]\Big) \\ &= \mathbb{P}[i(X;Y) \leq \gamma] + \frac{M-1}{2}\mathbb{P}[i(\overline{X};Y) > \gamma] \\ &= \mathbb{P}[i(X;Y) \leq \gamma] + \frac{M-1}{2}\mathbb{E}[\exp(-i(X;Y))\mathbf{1}\{i(X;Y) > \gamma\}] \qquad \text{(by (15.3))} \\ &= \mathbb{E}\Big[\mathbf{1}\{i(X;Y) \leq \gamma\} + \frac{M-1}{2}\exp(-i(X;Y))\mathbf{1}\{i(X,Y) > \gamma\}\Big] \\ &= \mathbb{E}\Big[\min\Big(1, \frac{M-1}{2}\exp(-i(X;Y))\Big)\Big] \qquad \Big(\gamma = \log\frac{M-1}{2} \text{ minimizes the upper bound}\Big) \\ &= \mathbb{E}\left[\exp\left\{-\left(i(X;Y) - \log\frac{M-1}{2}\right)^+\right\}\right].\end{aligned}$$

To optimize over γ, note the simple observation that $U\mathbf{1}_E + V\mathbf{1}_{\{E^c\}} \geq \min\{U, V\}$, with equality iff $U \geq V$ on E. Therefore for any x, y, $\mathbf{1}[i(x;y) \leq \gamma] + \frac{M-1}{2}e^{-i(x;y)}\mathbf{1}[i(x;y) > \gamma] \geq \min(1, \frac{M-1}{2}e^{-i(x;y)})$, achieved by $\gamma = \log\frac{M-1}{2}$ regardless of x, y. □

Note: Dependence-testing: The RHS of (15.11) is equivalent to the minimum error probability of the following Bayesian hypothesis testing problem:

$$H_0 : X, Y \sim P_{X,Y} \text{ versus } \quad H_1 : X, Y \sim P_X P_Y$$
$$\text{prior prob.: } \pi_0 = \frac{2}{M+1}, \pi_1 = \frac{M-1}{M+1}.$$

Note that $X, Y \sim P_{X,Y}$ and $\overline{X}, Y \sim P_X P_Y$, where X is the sent codeword and $\overline{X}$ is the unsent codeword. As we know from binary hypothesis testing, the best threshold for the LRT to minimize the weighted probability of error is $\log\frac{\pi_1}{\pi_0}$.

Note: Here we avoid minimizing over τ in Shannon's bound (15.10) to get the minimum upper bound in Theorem 15.1. Moreover, DT bound is stronger than the best Shannon's bound (with optimized τ).

Note: Similar to the random coding achievability bound of almost lossless compression (Theorem 7.4), in Theorem 15.1 and Theorem 15.2 we only need the random codewords to be *pairwise* independent.

15.4 Feinstein's Lemma

The previous achievability results are obtained using *probabilistic* methods (random coding). In contrast, the following achievability due to Feinstein uses a **greedy** construction. Moreover, Feinstein's construction holds for **maximal** probability of error.

Theorem 15.3 (Feinstein's lemma)**.** *$\forall P_X$, $\forall\gamma > 0$, $\forall\epsilon \in (0,1)$, $\exists(M, \epsilon)_{\max}$-code such that*

$$M \geq \gamma(\epsilon - \mathbb{P}[i(X;Y) < \log\gamma]) \tag{15.12}$$

Remark 15.6 (Comparison with Shannon's bound)**.** We can also interpret (15.12) as for fixed M, there exists an $(M, \epsilon)_{\max}$-code that achieves the maximal error probability bounded as follows:

$$\epsilon \leq \mathbb{P}[i(X;Y) < \log\gamma] + \frac{M}{\gamma}$$

Take $\log\gamma = \log M + \tau$, this gives the bound of exactly the same form in (15.10). However, the two are proved in seemingly quite different ways: Shannon's bound is by random coding, while Feinstein's bound is by greedily selecting the codewords. Nevertheless, Feinstein's bound is stronger in the sense that it concerns about the max error probability instead of the average.

Proof. The idea is to construct the codebook of size M in a greedy way.

For every $x \in \mathcal{X}$, associate it with a preliminary decode region defined as follows:

$$E_x \triangleq \{y : i(x;y) \geq \log\gamma\}$$

Notice that the preliminary decoding regions $\{E_x\}$ may be overlapping, and we denote the final decoding region partition regions by $\{D_x\}$.

We can assume that $\mathbb{P}[i(X;Y) < \log\gamma] \leq \epsilon$, for otherwise the R.H.S. of (15.12) is negative and there is nothing to prove. We first claim that there exists some c such that $P_Y[E_c|X = c] \geq 1 - \epsilon$.

Show by contradiction. Assume that $\forall c \in \mathcal{X}$, $\mathbb{P}[i(c;Y) \geq \log\gamma | X = c] < 1 - \epsilon$, then pick $c \sim P_X$, we have $\mathbb{P}[i(X;Y) \geq \log\gamma] < 1 - \epsilon$, which is a contradiction.

Then we construct the codebook in the following greedy way:

1. Pick c_1 to be any codeword such that $P_Y[E_{c_1} | X = c_1] \geq 1 - \epsilon$, and set $D_1 = E_{c_1}$;
2. Pick c_2 to be any codeword such that $P_Y[E_{c_2} \backslash D_1 | X = c_2] \geq 1 - \epsilon$, and set $D_2 = E_{c_2} \backslash D_1$;
   ...
3. Pick c_M to be any codeword such that $P_Y[E_{c_M} \backslash \cup_{j=1}^{M-1} D_j | X = c_M] \geq 1 - \epsilon$, and set $D_M = E_{c_M} \backslash \cup_{j=1}^{M-1} D_j$. We stop if no more codeword can be found, i.e., M is determined by the stopping condition:
$$\forall x_0 \in \mathcal{X}, P_Y[E_{x_0} \backslash \cup_{j=1}^{M} D_j | X = x_0] < 1 - \epsilon$$

Averaging over $x_0 \sim P_X$, the stopping condition gives that

$$\mathbb{P}(\{i(X;Y) \geq \log\gamma\} \backslash \{Y \in \cup_{j=1}^{M} D_j\}) < 1 - \epsilon$$

by union bound $P(A \backslash B) \geq P(A) - P(B)$, we have

$$\begin{aligned} &\mathbb{P}(i(X;Y) \geq \log\gamma) - \sum_{j=1}^{M} P_Y(D_j) < 1 - \epsilon \\ \Rightarrow &\mathbb{P}(i(X;Y) \geq \log\gamma) - \frac{M}{\gamma} < 1 - \epsilon \end{aligned}$$

where the last step makes use of the following key observation:

$$P_Y(D_j) \leq P_Y(E_{c_j}) = P_Y(i(c_j;Y) \geq \log\gamma) < \frac{1}{\gamma}, \quad \text{(by Corollary 15.1).}$$

□

Recall that last time we showed the following achievability bounds:

$$\begin{aligned} \text{Shannon's:} \quad & P_e \le P[i(X;Y) \le \log M + \tau] + \exp\{-\tau\} \\ & \Uparrow \\ \text{DT:} \quad & P_e \le \mathbb{E}\left[\exp\left\{-\left(i(X;Y) - \log\frac{M-1}{2}\right)^+\right\}\right] \\ \text{Feinstein's:} \quad & P_{e,max} \le P[i(X;Y) \le \log M + \tau] + \exp\{-\tau\} \end{aligned}$$

This time we shall use a shortcut to prove the above bounds and in which case $P_e = P_{e,max}$.

16.1 Linear coding

Definition 16.1 (Linear code)**.** Let $\mathcal{X} = \mathcal{Y} = \mathbb{F}_q^n$, $M = q^k$. Denote the codebook by $\mathcal{C} \triangleq \{c_u : u \in \mathbb{F}_q^k\}$. A code $f : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is a **linear code** if $\forall u \in \mathbb{F}_q^k$, $c_u = uG$ (row-vector convention), where $G \in \mathbb{F}_q^{k\times n}$ is a **generator matrix**.

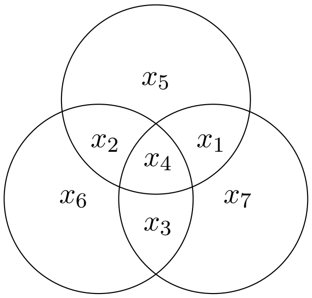
Proposition 16.1.

$$\begin{aligned} & c \in \mathcal{C} \\ \Leftrightarrow\ & c \in \textit{row span of } G \\ \Leftrightarrow\ & c \in \mathrm{Ker}H, \ \textit{for some } H \in \mathbb{F}_q^{(n-k)\times n} \ \textit{s.t.}\ HG^T = 0. \end{aligned}$$

Note: For linear codes, the codebook is a k-dimensional linear subspace of $\mathbb{F}_q^n$ (ImG or Ker$H$). The matrix H is called a **parity check matrix**.

Example: (Hamming code) The $[7,4,3]_2$ Hamming code over $\mathbb{F}_2$ is a linear code with $G = [I;P]$ and $H = [-P^T;I]$ is a parity check matrix.

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \quad H = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$



Parity check: all four bits in the same circle sum up to zero.

Note: Linear codes are almost always examined with channels of additive noise.

Definition 16.2 (Additive noise)**.** $P_{Y|X}$ is additive-noise over $\mathbb{F}_q^n$ if

$$P_{Y|X}(y|x) = P_{Z^n}(y-x) \Leftrightarrow Y = X + Z^n \text{ where } Z^n \perp\!\!\!\perp X$$

Now: Given a linear code and an additive-noise $P_{Y|X}$, what can we say about the decoder?

Theorem 16.1. *Any $[k,n]_{\mathbb{F}_q}$ linear code over an additive-noise $P_{Y|X}$ has a maximum likelihood decoder $g : \mathbb{F}_q^n \to \mathbb{F}_q^n$ such that:*

1. *$g(y) = y - g_{\text{synd}}(Hy^T)$, i.e., the decoder is a function of the "syndrome" Hy^T only*

2. *Decoding regions are translates: $D_u = c_u + D_0, \forall u$*

3. *$P_{e,max} = P_e$,*

where $g_{\text{synd}} : \mathbb{F}_q^{n-k} \to \mathbb{F}_q^n$, defined by $g_{\text{synd}}(s) = \text{argmax}_{z:Hx^T=s} P_Z(z)$, is called the "syndrome decoder", which decodes the most likely realization of the noise.

Proof. 1. The maximum likelihood decoder for linear code is

$$g(y) = \underset{c\in\mathcal{C}}{\text{argmax}}\, P_{Y|X}(y|c) = \underset{c:Hc^T=0}{\text{argmax}}\, P_Z(y-c) = y - \underbrace{\underset{z:Hz^T=Hy^T}{\text{argmax}}\; P_Z(z)}_{\triangleq g_{\text{synd}}(Hy^T)},$$

2. For any u, the decoding region

$$D_u = \{y : g(y) = c_u\} = \{y : y - g_{\text{synd}}(Hy^T) = c_u\} = \{y : y - c_u = g_{\text{synd}}(H(y-c_u)^T)\} = c_u + D_0,$$

where we used $Hc_u^T = 0$ and $c_0 = 0$.

3. For any u,

$$\mathbb{P}[\hat{W} \neq u | W = u] = \mathbb{P}[g(c_u+Z) \neq c_u] = \mathbb{P}[c_u+Z-g_{\text{synd}}(Hc_u^T+HZ^T) \neq c_u] = \mathbb{P}[g_{\text{synd}}(HZ^T) \neq Z].$$

□

Note: The advantages of linear codes include at least

1. Low-complexity encoding

2. Slightly lower complexity ML decoding (syndrome decoding)

3. Under ML decoding, maximum probability of error = average probability of error. This is a consequence of the symmetry of the codes. Note that this holds as long as the decoder is a function of the syndrome only. As shown in Theorem 16.1, syndrome is a **sufficient statistic** for decoding a linear code.

Theorem 16.2 (DT bounds for linear codes)**.** *Let $P_{Y|X}$ be additive noise over $\mathbb{F}_q^n$. $\forall k, \exists$ a linear code $f : \mathbb{F}_q^k \to \mathbb{F}_q^n$ with the error probability:*

$$P_{e,\max} = P_e \leq \mathbb{E}\Big[q^{-\left(n-k-\log_q \frac{1}{P_{Z^n}(Z^n)}\right)^+}\Big] \tag{16.1}$$

Proof. Recall that in proving the Shannon's achievability bounds, we select the code words $c_1, \dots, c_M$ i.i.d $\sim P_X$ and showed that

$$\mathbb{E}[P_e(c_1, \dots, c_M)] \le P[i(X;Y) \le \gamma] + \frac{M-1}{2} P(i(\overline{X};Y) \ge \gamma)$$

As noted after the proof of the DT bound, we only need the random codewords to be **pairwise independent**. Here we will adopt a similar approach. Note that $M = q^k$.

Let's first do a quick check of the capacity achieving input distribution for $P_{Y|X}$ with additive noise over $\mathbb{F}_q^n$:

$$\max_{P_X} I(X;Y) = \max_{P_X} H(Y) - H(Y|X) = \max_{P_X} H(Y) - H(Z^n) = n \log q - H(Z^n) \Rightarrow P_X^* \text{ uniform on } \mathbb{F}_q^n$$

We shall use the uniform distribution P_X in the "random coding" trick.

Moreover, the optimal (MAP) decoder with uniform input is the ML decoder, whose decoding regions are translational invariant by Theorem 16.1, namely $D_u = c_u + D_0, \forall u$, and therefore:

$$P_{e,max} = P_e = P[\hat{W} \ne u | W = u], \forall u.$$

Step 1: Random linear coding with dithering:

$$\forall u \in \mathbb{F}_q^k, c_u = uG + h$$

G and h are drawn from the new ensemble, where the $k \times n$ entries of G and the $1 \times n$ entries of h are i.i.d. uniform over $\mathbb{F}_q$. We add the dithering to eliminate the special role that the all-zero codeword plays (since it is contained in any linear codebook).

Step 2: Claim that the codewords are pairwise independent and uniform: $\forall u \ne u'$, $(c_u, c_{u'}) \sim (X, \overline{X})$, where $P_{X,\overline{X}}(x, \overline{x}) = 1/q^{2n}$. To see this:

$$\begin{aligned} &c_u \sim \text{uniform on } \mathbb{F}_q^n \\ &c_{u'} = u'G + h = uG + h + (u' - u)G = c_u + (u' - u)G \end{aligned}$$

We claim that $c_u \perp\!\!\!\perp G$ because conditioned on the generator matrix $G = G_0$, $c_u \sim$ uniform on $\mathbb{F}_q^n$ due to the dithering h.
We also claim that $c_u \perp\!\!\!\perp c_{u'}$ because conditioned on c_u, $(u' - u)G \sim$ uniform on $\mathbb{F}_q^n$.
Thus random linear coding with dithering indeed gives codewords $c_u, c_{u'}$ pairwise independent and are uniformly distributed.

Step 3: Repeat the same argument in proving DT bound for the symmetric and pairwise independent codewords, we have

$$\begin{aligned} &\mathbb{E}[P_e(c_1, \dots, c_M)] \le P[i(X;Y) \le \gamma] + \frac{M-1}{2} P(i(\overline{X}, Y) \ge \gamma) \\ \Rightarrow &P_e \le \mathbb{E}[\exp\{-\big(i(X;Y) - \log \frac{M-1}{2}\big)^+\}] = \mathbb{E}[q^{-\big(i(X;Y) - \log_q \frac{q^k - 1}{2}\big)^+}] \le \mathbb{E}[q^{-\big(i(X;Y) - k\big)^+}] \end{aligned}$$

where we used $M = q^k$ and picked the base of log to be q.

Step 4: compute $i(X;Y)$:

$$i(a;b) = \log_q \frac{P_{Z^n}(b-a)}{q^{-n}} = n - \log_q \frac{1}{P_{Z^n}(b-a)}$$

therefore

$$P_e \le \mathbb{E}[q^{-\big(n - k - \log_q \frac{1}{P_{Z^n}(Z^n)}\big)^+}] \tag{16.2}$$

Step 5: Kill h. We claim that there exists a linear code without dithering such that (16.2) is satisfied. Indeed shifting a codebook has no impact on its performance. We modify the coding scheme with G, h which achieves the bound in the following way: modify the decoder input $Y' = Y - h$, then when c_u is sent, the additive noise $P_{Y'|X}$ becomes then $Y' = uG + h + Z^n - h = uG + Z^n$, which is equivalent to that the linear code generated by G is used. □

Notes:

- The ensemble $c_u = uG + h$ has the pairwise independence property. The joint entropy $H(c_1, \ldots, c_M) = H(G) + H(h) = (nk + n)\log q$ is significantly smaller than Shannon's "fully random" ensemble we used in the previous lecture. Recall that in that ensemble each c_j was selected independently uniform over $\mathbb{F}_q^n$, implying $H(c_1, \ldots, c_M) = q^k n \log q$. Question:

$$\min H(c_1, \ldots, c_M) = ??$$

  where minimum is over all distributions with $P[c_i = a, c_j = b] = q^{-2n}$ when $i \neq j$ (pairwise independent, uniform codewords). Note that $H(c_1, \ldots, c_M) \geq H(c_1, c_2) = 2n \log q$. Similarly, we may ask for (c_i, c_j) to be uniform over all pairs of *distinct* elements. In this case Wozencraft ensemble for the case of $n = 2k$ achieves $H(c_1, \ldots, c_{q^k}) \approx 2n \log q$.

- There are many different ensembles of random codebooks:

  - Shannon ensemble: $\mathcal{C} = \{c_1, \ldots, c_M\} \overset{\text{i.i.d.}}{\sim} P_X$ – fully random
  - Elias ensemble [Eli55]: $\mathcal{C} = \{uG : u \in \mathbb{F}_q^k\}$, with generator matrix G uniformly drawn at random.
  - Gallager ensemble: $\mathcal{C} = \{c : Hc^T = 0\}$, with parity-check matrix H uniformly drawn at random.

- With some non-zero probability G may fail to be full rank [Exercise: Find $\mathbb{P}[\text{rank}(G) < k]$ as a function of n, k, q!]. In such a case, there are two identical codewords and hence $P_{e,\max} \geq 1/2$. There are two alternative ensembles of codes which do not contain such degenerate codebooks:

  1. $G \sim$ uniform on all full rank matrices
  2. search codeword $c_u \in \text{Ker}H$ where $H \sim$ uniform on all $n \times (n-k)$ full row rank matrices. (random parity check construction)

  Analysis of random coding over such ensemble is similar, except that this time $(X, \bar{X})$ have distribution

$$P_{X,\bar{X}} = \frac{1}{q^{2n} - q^n} \mathbf{1}_{\{X \neq X'\}}$$

  uniform on all pairs of *distinct* codewords and *not* pairwise independent.

16.2 Channels and channel capacity

Basic question of data transmission: How many bits can one transmit reliably if one is allowed to use the channel n times?

- Rate = # of bits per channel use
- Capacity = highest achievable rate

Next we formalize these concepts.

Definition 16.3 (Channel)**.** A channel is specified by:

- input alphabet $\mathcal{A}$
- output alphabet $\mathcal{B}$
- a sequence of random transformation kernels $P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n, n = 1, 2, \ldots$.
- The parameter n is called the *blocklength*.

Note: we do not insist on $P_{Y^n|X^n}$ to have any relation for different n, but it is common that the conditional distribution of the first k letters of the n-th transformation is in fact a function of only the first k letters of the input and this function equals $P_{Y^k|X^k}$ – the k-th transformation. Such channels, in particular, are non-anticipatory: channel outputs are causal functions of channel inputs.

Channel characteristics:

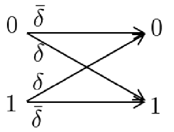
- A channel is *discrete* if $\mathcal{A}$ and $\mathcal{B}$ are finite.
- A channel is *additive-noise* if $\mathcal{A} = \mathcal{B}$ are abelian group, and

$$P_{y^n|x^n} = P_{Z^n}(y^n - x^n) \Leftrightarrow Y^n = X^n + Z^n.$$

- A channel is *memoryless* if there exists a sequence $\{P_{X_k|Y_k}, k = 1, \ldots\}$ of transformations acting $\mathcal{A} \to \mathcal{B}$ such that $P_{Y^n|X^n} = \prod_{k=1}^n P_{Y_k|X_k}$ (in particular, the channels are compatible at different blocklengths).
- A channel is *stationary memoryless* if $P_{Y^n|X^n} = \prod_{k=1}^n P_{Y_1|X_1}$.
- **DMC** (discrete memoryless stationary channel)
  A DMC can be specified in two ways:
  - an $|\mathcal{A}| \times |\mathcal{B}|$-dimensional matrix $P_{Y|X}$ where elements specify the transition probabilities
  - a bipartite graph with edge weight specifying the transition probabilities.

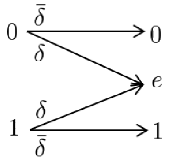
Example:

BSC



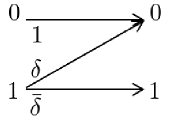
$$\begin{bmatrix} \bar{\delta} & \delta \\ \delta & \bar{\delta} \end{bmatrix}$$

BEC



$$\begin{bmatrix} \bar{\delta} & \delta & 0 \\ 0 & \delta & \bar{\delta} \end{bmatrix}$$

Z channel



$$\begin{bmatrix} 1 & 0 \\ \delta & \bar{\delta} \end{bmatrix}$$

Definition 16.4 (Fundamental Limits)**.** For any channel,

- An (n, M, ϵ)-code is an (M, ϵ)-code for the n-th random transformation $P_{Y^n|X^n}$.

- An $(n, M, \epsilon)_{\max}$-code is analogously defined for maximum probability of error.

The non-asymptotic fundamental limits are

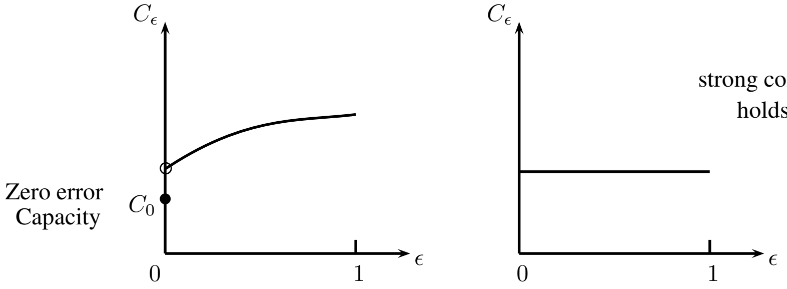
$$M^*(n,\epsilon) = \max\{M : \exists\ (n,M,\epsilon)\text{-code}\} \tag{16.3}$$
$$M^*_{\max}(n,\epsilon) = \max\{M : \exists\ (n,M,\epsilon)_{\max}\text{-code}\} \tag{16.4}$$

Definition 16.5 (Channel capacity)**.** The **ϵ-Capacity** C_ϵ and **Shannon Capacity** C are

$$C_\epsilon \triangleq \liminf_{n\to\infty} \frac{1}{n} \log M^*(n,\epsilon)$$
$$C = \lim_{\epsilon\to 0^+} C_\epsilon$$

Notes:

- This **operational** definition of the capacity represents the maximum achievable rate at which one can communicate through a channel with probability of error less than ϵ. In other words, for any $R < C$, there exists an $(n, \exp(nR), \epsilon_n)$-code, such that $\epsilon_n \to 0$.
- Typically, the ϵ-capacity behaves like the plot below on the left-hand side, where C_0 is called the *zero-error capacity*, which represents the maximal achievable rate with no error. Often times $C_0 = 0$, meaning without tolerating any error zero information can be transmitted. If C_ϵ is constant for all ϵ (see plot on the right-hand side), then we say that the **strong converse** holds (more on this later).



Proposition 16.2 (Equivalent definitions of C_ϵ and C)**.**

$$C_\epsilon = \sup\{R : \forall\delta > 0, \exists n_0(\delta), \forall n \geq n_0(\delta), \exists (n, 2^{n(R-\delta)}, \epsilon)\ \textit{code}\}$$
$$C = \sup\{R : \forall\epsilon > 0, \forall\delta > 0, \exists n_0(\delta,\epsilon), \forall n \geq n_0(\delta,\epsilon), \exists (n, 2^{n(R-\delta)}, \epsilon)\ \textit{code}\}$$

Proof. This trivially follows from applying the definitions of $M^*(n,\epsilon)$ (DIY). □

Question: Why do we define capacity C_ϵ and C with respect to average probability of error, say, $C_\epsilon^{(\max)}$ and $C^{(\max)}$? Why not maximal probability of error? It turns out that these two definitions are equivalent, as the next theorem shows.

Theorem 16.3. $\forall \tau \in (0,1)$,

$$\tau M^*(n, \epsilon(1-\tau)) \leq M^*_{\max}(n,\epsilon) \leq M^*(n,\epsilon)$$

Proof. The second inequality is obvious, since any code that achieves a maximum error probability ϵ also achieves an average error probability of ϵ.

For the first inequality, take an $(n, M, \epsilon(1-\tau))$-code, and define the error probability for the j^{th} codeword as

$$\lambda_j \triangleq \mathbb{P}[\hat{W} \neq j | W = j]$$

Then

$$M(1-\tau)\epsilon \geq \sum \lambda_j = \sum \lambda_j \mathbf{1}_{\{\lambda_j \leq \epsilon\}} + \sum \lambda_j \mathbf{1}_{\{\lambda_j > \epsilon\}} \geq \epsilon |\{j : \lambda_j > \epsilon\}|.$$

Hence $|\{j : \lambda_j > \epsilon\}| \leq (1-\tau)M$. [Note that this is exactly Markov inequality!] Now by removing those codewords[1] whose λ_j exceeds ϵ, we can extract an $(n, \tau M, \epsilon)_{\max}$-code. Finally, take $M = M^*(n, \epsilon(1-\tau))$ to finish the proof. □

Corollary 16.1 (Capacity under maximal probability of error)**.** *$C_\epsilon^{(\max)} = C_\epsilon$ for all $\epsilon > 0$ such that $C_\epsilon = C_{\epsilon-}$. In particular, $C^{(\max)} = C$.*[2]

Proof. Using the definition of M^* and the previous theorem, for any fixed $\tau > 0$

$$C_\epsilon \geq C_\epsilon^{(\max)} \geq \liminf_{n\to\infty} \frac{1}{n} \log \tau M^*(n, \epsilon(1-\tau)) \geq C_{\epsilon(1-\tau)}$$

Sending $\tau \to 0$ yields $C_\epsilon \geq C_\epsilon^{(\max)} \geq C_{\epsilon-}$. □

16.3 Bounds on C_ϵ; Capacity of Stationary Memoryless Channels

Now that we have the basic definitions for C_ϵ, we define another type of capacity, and show that for a *stationary memoryless* channels, the two notions ("operational" and "information" capacity) coincide.

Definition 16.6. The **information capacity** of a channel is

$$C_i = \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n; Y^n)$$

Remark: This quantity is not the same as the Shannon capacity, and has no direct operational interpretation as a quantity related to coding. Rather, it is best to think of this only as taking the n-th random transformation in the channel, maximizing over input distributions, then normalizing and looking at the limit of this sequence.

Next we give **coding theorems** to relate information capacity (information measures) to Shannon capacity (operational quantity).

Theorem 16.4 (Upper Bound for C_ϵ)**.** *For any channel, $\forall \epsilon \in [0,1)$, $C_\epsilon \leq \frac{C_i}{1-\epsilon}$ and $C \leq C_i$.*

Proof. Recall the general weak converse bound, Theorem 14.4:

$$\log M^*(n, \epsilon) \leq \frac{\sup_{P_{X^n}} I(X^n; Y^n) + h(\epsilon)}{1-\epsilon}$$

[1] This operation is usually referred to as *expurgation* which yields a smaller code by killing part of the codewords to reach a desired property.

[2] **Notation:** $f(x-) \triangleq \lim_{y \nearrow x} f(y)$.

Normalizing this by n the taking the $\liminf$ gives

$$C_\epsilon = \liminf_{n\to\infty} \frac{1}{n} \log M^*(n,\epsilon) \le \liminf_{n\to\infty} \frac{1}{n} \frac{\sup_{P_{X^n}} I(X^n;Y^n) + h(\epsilon)}{1-\epsilon} = \frac{C_i}{1-\epsilon}$$

□

Next we give an achievability bound:

Theorem 16.5 (Lower Bound for C_ϵ)**.** *For a stationary memoryless channel, $C_\epsilon \ge C_i$, for any $\epsilon \in (0,1]$.*

The following result follows from pairing the upper and lower bounds on C_ϵ.

Theorem 16.6 (Shannon '1948)**.** *For a stationary memoryless channel,*

$$C = C_i = \sup_{P_X} I(X;Y). \tag{16.5}$$

Remark 16.1. The above result, known as **Shannon's Noisy Channel Theorem**, is perhaps the most significant result in information theory. For communications engineers, the major surprise was that $C > 0$, i.e. communication over a channel is possible with strictly positive rate for any arbitrarily small probability of error. This result influenced the evolution of communication systems to block architectures that used bits as a universal currency for data, along with encoding and decoding procedures.

Before giving the proof of Theorem 16.5, we show the second equality in (16.5). Notice that C_i for stationary memoryless channels is easy to compute: Rather than solving an optimization problem for each n and taking the limit of $n \to \infty$, computing C_i boils down to maximizing mutual information for $n = 1$. This type of result is known as "**single-letterization**" in information theory.

Proposition 16.3 (Memoryless input is optimal for memoryless channels)**.**
For memoryless channels,

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

For stationary memoryless channels,

$$C_i = \sup_{P_X} I(X;Y).$$

Proof. Recall that for product kernels $P_{Y^n|X^n} = \prod P_{Y_i|X_i}$, we have $I(X^n;Y^n) \le \sum_{k=1}^n I(X_k;Y_k)$, with equality when X_i's are independent. Then

$$C_i = \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n;Y^n) = \liminf_{n\to\infty} \sup_{P_X} I(X;Y) = \sup_{P_X} I(X;Y) \square$$

Proof of Theorem 16.5. $\forall P_X$, and let $P_{X^n} = P_X^n$ (iid product). Recall Shannon's (or Feinstein's) achievability bound: For any n, M and any $\gamma > 0$, there exists (n, M, ϵ_n)-code, s.t.

$$\epsilon_n \le \mathbb{P}[i(X^n;Y^n) \le \log M + \gamma] + \exp(-\gamma)$$

Here the information density is defined as

$$i(X^n, Y^n) = \log \frac{dP_{Y^n|X^n}}{dP_{Y^n}}(Y^n|X^n) = \sum_{k=1}^{n} \log \frac{dP_{Y|X}}{dP_Y}(Y_k|X_k) = \sum_{k=1}^{n} i(X_k;Y_k),$$

which is a sum of iid r.v.'s with mean $I(X;Y)$. Set $\log M = n(I(X;Y) - 2\delta)$ for $\delta > 0$, and taking $\gamma = \delta n$ in Shannon's bound, we have

$$\epsilon_n \le \mathbb{P}\Big[\sum_{k=1}^{n} i(X_k;Y_k) \le nI(X;Y) - \delta n\Big] + \exp(-\delta n) \xrightarrow{n\to\infty} 0$$

The second terms goes to zero since $\delta > 0$, and the first terms goes to zero by WLLN.

Therefore, $\forall P_X$, $\forall \delta > 0$, there exists a sequence of (n, M_n, ϵ_n)-codes with $\epsilon_n \to 0$ (where $\log M_n = n(I(X;Y) - 2\delta)$). Hence, for all n such that $\epsilon_n \le \epsilon$

$$\log M^*(n,\epsilon) \ge n(I(X;Y) - 2\delta)$$

And so

$$C_\epsilon = \liminf_{n\to\infty} \frac{1}{n} \log M^*(n,\epsilon) \ge I(X;Y) - 2\delta \quad \forall P_X, \forall \delta$$

Since this holds for all P_X and all δ, we conclude $C_\epsilon \ge \sup_{P_X} I(X;Y) = C_i$. □

Remark 16.2. Shannon's noisy channel theorem (Theorem 16.6) shows that by employing codes of large blocklength, we can approach the channel capacity arbitrarily close. Given the asymptotic nature of this result (or any other asymptotic result), two natural questions are in order dealing with the different aspects of the price to reach capacity:

1. The **complexity** of achieving capacity: Is it possible to find low-complexity encoders and decoders with polynomial number of operations in the blocklength n which achieve the capacity? This question is resolved by Forney in 1966 who showed that this is possible in *linear* time with exponentially small error probability. His main idea is concatenated codes. We will study the complexity question in detail later.

   Note that if we are content with polynomially small probability of error, e.g., $P_e = O(n^{-100})$, then we can construct polynomially decodable codes as follows. First, it can be shown that with rate strictly below capacity, the error probability of optimal codes decays exponentially w.r.t. the blocklenth. Now divide the block of length n into shorter block of length $C \log n$ and apply the optimal code for blocklength $C \log n$ with error probability n^{-101}. The by the union bound, the whole block is has error with probability at most n^{-100}. The encoding and exhaustive-search decoding are obviously polynomial time.

2. The **speed** of achieving capacity: Suppose we want to achieve 90% of the capacity, we want to know how long do we need wait? The blocklength is a good proxy for delay. In other words, we want to know how fast the gap to capacity vanish as blocklength grows. Shannon's theorem shows that the gap $C - \frac{1}{n}\log M^*(n,\epsilon) = o(1)$. Next theorem shows that under proper conditions, the $o(1)$ term is in fact $O(\frac{1}{\sqrt{n}})$.

The main tool in the proof of Theorem 16.5 is the WLLN. The lower bound $C_\epsilon \ge C_i$ in Theorem 16.5 shows that $\log M^*(n,\epsilon) \ge nC + o(n)$ (since normalizing by n and taking the liminf must result in something $\ge C$). If instead we do a more refined analysis using the CLT, we find

Theorem 16.7. *For any stationary memoryless channel with $C = \max_{P_X} I(X;Y)$ (i.e. $\exists P_X^* = \operatorname{argmax}_{P_X} I(X;Y)$) such that $V = \mathrm{Var}[i(X^*;Y^*)] < \infty$, then*

$$\log M^*(n,\epsilon) \ge nC - \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n}),$$

where $Q(\cdot)$ is the complementary Gaussian CDF and $Q^{-1}(\cdot)$ is its functional inverse.

Proof. Writing the little-o notation in terms of lim inf, our goal is

$$\liminf_{n\to\infty} \frac{\log M^*(n,\epsilon) - nC}{\sqrt{nV}} \geq -Q^{-1}(\epsilon) = \Phi^{-1}(\epsilon),$$

where $\Phi(t)$ is the standard normal CDF.

Recall Feinstein's bound

$$\exists (n, M, \epsilon)_{\max}: \quad M \geq \beta\left(\epsilon - \mathbb{P}[i(X^n;Y^n) \leq \log\beta]\right)$$

Take $\log\beta = nC + \sqrt{nV}t$, then applying the CLT gives

$$\begin{aligned} &\log M \geq nC + \sqrt{nV}t + \log\left(\epsilon - \mathbb{P}\left[\sum i(X_k;Y_k) \leq nC + \sqrt{nV}t\right]\right) \\ \implies &\log M \geq nC + \sqrt{nV}t + \log\left(\epsilon - \Phi(t)\right) \quad \forall \Phi(t) < \epsilon \\ \implies &\frac{\log M - nC}{\sqrt{nV}} \geq t + \frac{\log(\epsilon - \Phi(t))}{\sqrt{nV}} \end{aligned}$$

Where $\Phi(t)$ is the standard normal CDF. Taking the liminf of both sides

$$\liminf_{n\to\infty} \frac{\log M^*(n,\epsilon) - nC}{\sqrt{nV}} \geq t \quad \forall t \ s.t. \ \Phi(t) < \epsilon$$

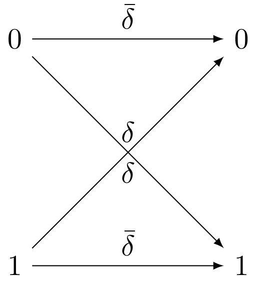
Taking $t \nearrow \Phi^{-1}(\epsilon)$, and writing the liminf in little o form completes the proof

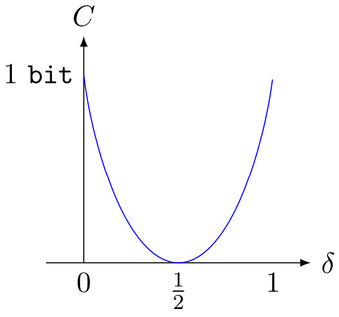
$$\log M^*(n,\epsilon) \geq nC - \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n})$$

□

16.4 Examples of DMC

Binary symmetric channels





$$Y = X + Z, \quad Z \sim \text{Bern}(\delta) \perp\!\!\!\perp X$$

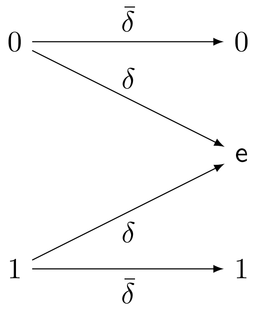
Capacity of BSC:

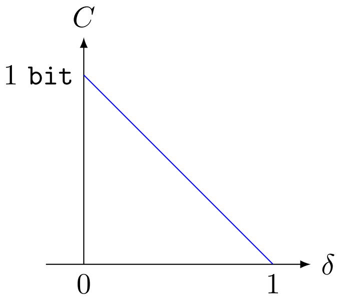
$$C = \sup_{P_X} I(X;Y) = 1 - h(\delta)$$

Proof. $I(X;X+Z) = H(X+Z) - H(X+Z|X) = H(X+Z) - H(Z) \leq 1 - h(\delta)$, with equality iff $X \sim \text{Bern}(1/2)$. □

Note: More generally, for all additive-noise channel over a finite abelian group G, $C = \sup_{P_X} I(X;X+Z) = \log|G| - H(Z)$, achieved by uniform X.

Binary erasure channels





BEC is a **multiplicative** channel: If we think about the input $X \in \{\pm 1\}$, and output $Y \in \{\pm 1, 0\}$. Then equivalently we can write $Y = XZ$ with $Z \sim \text{Bern}(\delta) \perp\!\!\!\perp X$.

Capacity of BEC:

$$C = \sup_{P_X} I(X;Y) = 1 - \delta \quad \texttt{bits}$$

Proof. Note that $P(X=0|Y=\texttt{e}) = \frac{P(X=0)\delta}{\delta} = P(X=0)$. Therefore $I(X;Y) = H(X) - H(X|Y) = H(X) - H(X|Y=\texttt{e}) \le (1-\delta)H(X) \le 1-\delta$, with equality iff $X \sim \text{Bern}(1/2)$. □

16.5* Information Stability

We saw that $C = C_i$ for stationary memoryless channels, but what other channels does this hold for? And what about non-stationary channels? To answer this question, we introduce the notion of *information stability.*

Definition 16.7. A channel is called *information stable* if there exists a sequence of input distribution $\{P_{X^n}, n = 1, 2, \dots\}$ such that

$$\frac{1}{n} i(X^n;Y^n) \longrightarrow C_i \text{ in probability}$$

For example, we can pick $P_{X^n} = (P_X^*)^n$ for stationary memoryless channels. Therefore stationary memoryless channels are information stable.

The purpose for defining information stability is the following theorem.

Theorem 16.8. *For an information stable channel, $C = C_i$.*

Proof. Like the stationary, memoryless case, the upper bound comes from the general converse Theorem 14.4, and the lower bound uses a similar strategy as Theorem 16.5, except utilizing the definition of information stability in place of WLLN. □

The next theorem gives conditions to check for information stability in memoryless channels which are *not* necessarily stationary.

Theorem 16.9. *A memoryless channel is information stable if either of there exists $\{X_k^*, k = 1, \ldots\}$ such that both of the following hold:*

$$\frac{1}{n}\sum_{k=1}^{n} I(X_k^*;Y_k^*) \to C_i \tag{16.6}$$

$$\sum_{n=1}^{\infty} \frac{1}{n^2} Var[i(X_n^*;Y_n^*)] < \infty\,. \tag{16.7}$$

In particular, this is satisfied if

$$|\mathcal{A}| < \infty \;\; or \;\; |\mathcal{B}| < \infty \tag{16.8}$$

Proof. To show the first part, it is sufficient to prove

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{k=1}^{n} i(X_k^*;Y_k^*) - I(X_k^*, Y_k^*)\right| > \delta\right] \to 0$$

So that $\frac{1}{n} i(X^n;Y^n) \to C_i$ in probability. We bound this by Chebyshev's inequality

$$\mathbb{P}\left[\frac{1}{n}\left|\sum_{k=1}^{n} i(X_k^*;Y_k^*) - I(X_k^*, Y_k^*)\right| > \delta\right] \le \frac{\frac{1}{n^2}\sum_{k=1}^{n} \mathrm{Var}[i(X_k^*;Y_k^*)]}{\delta^2} \to 0\,,$$

where convergence to 0 follows from Kronecker lemma (Lemma 16.1 to follow) applied with $b_n = n^2, x_n = \mathrm{Var}[i(X_n^*;Y_n^*)]/n^2$.

The second part follows from the first. Indeed, notice that

$$C_i = \liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \sup_{P_{X_k}} I(X_k;Y_k)\,.$$

Now select $P_{X_k^*}$ such that

$$I(X_k^*;Y_k^*) \ge \sup_{P_{X_k}} I(X_k;Y_k) - 2^{-k}\,.$$

(Note that each $\sup_{P_{X_k}} I(X_k;Y_k) \le \log\min\{|\mathcal{A}|, |\mathcal{B}|\} < \infty$.) Then, we have

$$\sum_{k=1}^{n} I(X_k^*;Y_k^*) \ge \sum_{k=1}^{n} \sup_{P_{X_k}} I(X_k;Y_k) - 1\,,$$

and hence normalizing by n we get (16.6). We next show that for any joint distribution $P_{X,Y}$ we have

$$\mathrm{Var}[i(X;Y)] \le 2\log^2(\min(|\mathcal{A}|, |\mathcal{B}|))\,. \tag{16.9}$$

The argument is symmetric in X and Y, so assume for concreteness that $|\mathcal{B}| < \infty$. Then

$$
\begin{aligned}
&\mathbb{E}[i^2(X;Y)] && (16.10)\\
\triangleq\; &\int_{\mathcal{A}} dP_X(x) \sum_{y\in\mathcal{B}} P_{Y|X}(y|x)\left[\log^2 P_{Y|X}(y|x) + \log^2 P_Y(y) - 2\log P_{Y|X}(y|x)\cdot\log P_Y(y)\right] && (16.11)\\
\le\; &\int_{\mathcal{A}} dP_X(x) \sum_{y\in\mathcal{B}} P_{Y|X}(y|x)\left[\log^2 P_{Y|X}(y|x) + \log^2 P_Y(y)\right] && (16.12)\\
=\; &\int_{\mathcal{A}} dP_X(x)\left[\sum_{y\in\mathcal{B}} P_{Y|X}(y|x)\log^2 P_{Y|X}(y|x)\right] + \left[\sum_{y\in\mathcal{B}} P_Y(y)\log^2 P_Y(y)\right] && (16.13)\\
\le\; &\int_{\mathcal{A}} dP_X(x) g(|\mathcal{B}|) + g(|\mathcal{B}|) && (16.14)\\
=\; &2g(|\mathcal{B}|)\,, && (16.15)
\end{aligned}
$$

where (16.12) is because $2\log P_{Y|X}(y|x)\cdot\log P_Y(y)$ is always non-negative, and (16.14) follows because each term in square-brackets can be upper-bounded using the following optimization problem:

$$g(n) \triangleq \sup_{a_j\ge 0:\sum_{j=1}^n a_j=1} \sum_{j=1}^n a_j \log^2 a_j\,. \tag{16.16}$$

Since the $x\log^2 x$ has unbounded derivative at the origin, the solution of (16.16) is always in the interior of $[0,1]^n$. Then it is straightforward to show that for $n>e$ the solution is actually $a_j=\frac{1}{n}$. For $n=2$ it can be found directly that $g(2)=0.5629\log^2 2<\log^2 2$. In any case,

$$2g(|\mathcal{B}|)\le 2\log^2|\mathcal{B}|\,.$$

Finally, because of the symmetry, a similar argument can be made with $|\mathcal{B}|$ replaced by $|\mathcal{A}|$. □

Lemma 16.1 (Kronecker Lemma)**.** *Let a sequence* $0<b_n\nearrow\infty$ *and a non-negative sequence* $\{x_n\}$ *such that* $\sum_{n=1}^\infty x_n<\infty$*, then*

$$\frac{1}{b_n}\sum_{j=1}^n b_j x_j \longrightarrow 0$$

Proof. Since b_n's are strictly increasing, we can split up the summation and bound them from above

$$\sum_{k=1}^n b_k x_k \le b_m\sum_{k=1}^m x_k + \sum_{k=m+1}^n b_k x_k$$

Now throw in the rest of the x_k's in the summation

$$\Longrightarrow \frac{1}{b_n}\sum_{k=1}^n b_k x_k \le \frac{b_m}{b_n}\sum_{k=1}^\infty x_k + \sum_{k=m+1}^n \frac{b_k}{b_n} x_k \le \frac{b_m}{b_n}\sum_{k=1}^\infty x_k + \sum_{k=m+1}^\infty x_k$$

$$\Longrightarrow \lim_{n\to\infty}\frac{1}{b_n}\sum_{k=1}^n b_k x_k \le \sum_{k=m+1}^\infty x_k \to 0$$

Since this holds for any m, we can make the last term arbitrarily small. □

Important example: For jointly Gaussian (X,Y) we always have bounded variance:

$$\mathrm{Var}[i(X;Y)] = \rho^2(X,Y)\log^2 e \le \log^2 e\,, \qquad \rho(X,Y) = \frac{\mathrm{cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}\,. \tag{16.17}$$

Indeed, first notice that we can always represent $Y=\tilde X+Z$ with $\tilde X=aX \perp\!\!\!\perp Z$. On the other hand, we have

$$i(\tilde x;y) = \frac{\log e}{2}\left[\frac{\tilde x^2+2\tilde x z}{\sigma_Y^2} - \frac{\sigma^2}{\sigma_Y^2\sigma_Z^2}z^2\right], \qquad z\triangleq y-\tilde x\,.$$

From here by using $\mathrm{Var}[\cdot]=\mathrm{Var}[\mathbb{E}[\cdot|\tilde X]]+\mathrm{Var}[\cdot|\tilde X]$ we need to compute two terms separately:

$$\mathbb{E}[i(\tilde X;Y)|\tilde X] = \frac{\log e}{2}\left[\frac{\tilde X^2-\frac{\sigma_{\tilde X}^2}{\sigma_Z^2}}{\sigma_Y^2}\right],$$

and hence

$$\mathrm{Var}[\mathbb{E}[i(\tilde X;Y)|\tilde X]] = \frac{2\log^2 e}{4\sigma_Y^4}\sigma_{\tilde X}^4\,.$$

On the other hand,

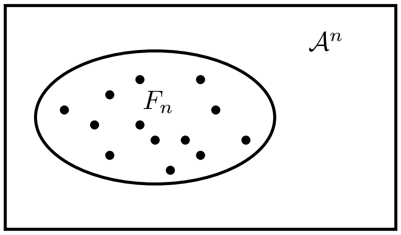
$$\mathrm{Var}[i(\tilde{X};Y)|\tilde{X}] = \frac{2\log^2 e}{4\sigma_Y^4}[4\sigma_{\tilde{X}}^2\sigma_Z^2 + 2\sigma_{\tilde{X}}^4].$$

Putting it all together we get (16.17). Inequality (16.17) justifies information stability of all sorts of Gaussian channels (memoryless and with memory), as we will see shortly.

17.1 Channel coding with input constraints

Motivations: Let us look at the additive Gaussian noise. Then the Shannon capacity is infinite, since $\sup_{P_X} I(X;X+Z) = \infty$ achieved by $X \sim \mathcal{N}(0,P)$ and $P \to \infty$. But this is at the price of infinite second moment. In reality, limitation of transmission power ⇒ constraints on the encoding operations ⇒ constraints on input distribution.

Definition 17.1. An (n, M, ϵ)-code satisfies the input constraint $F_n \subset \mathcal{A}^n$ if the encoder is $f : [M] \to F_n$. (Without constraint, the encoder maps into $\mathcal{A}^n$).



Codewords all land in a subset of $\mathcal{A}^n$

Definition 17.2 (Separable cost constraint)**.** A channel with separable cost constraint is specified as follows:

1. $\mathcal{A}, \mathcal{B}$: input/output spaces
2. $P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n$, $n = 1, 2, \ldots$
3. Cost $\mathsf{c} : \mathcal{A} \to \bar{\mathbb{R}}$

Input constraint: average per-letter cost of a codeword x^n (with slight abuse of notation)

$$\mathsf{c}(x^n) = \frac{1}{n}\sum_{k=1}^{n} \mathsf{c}(x_k) \le P$$

Example: $\mathcal{A} = \mathcal{B} = \mathbb{R}$

- Average power constraint (separable):

$$\frac{1}{n}\sum_{i=1}^{n} |x_i|^2 \le P \quad \Leftrightarrow \quad \|x^n\|_2 \le \sqrt{nP}$$

- Peak power constraint (non-separable):

$$\max_{1\le i\le n} |x_i| \le A \quad \Leftrightarrow \quad \|x^n\|_\infty \le A$$

Definition 17.3. Some basic definitions in parallel with the channel capacity without input constraint.

- A code is an (n, M, ϵ, P)-code if it is an (n, M, ϵ)-code satisfying input constraint $F_n \triangleq \{x^n : \frac{1}{n}\sum \mathsf{c}(x_k) \le P\}$
- Finite-n fundamental limits:

$$\begin{aligned} M^*(n, \epsilon, P) &= \max\{M : \exists (n, M, \epsilon, P)\text{-code}\} \\ M^*_{max}(n, \epsilon, P) &= \max\{M : \exists (n, M, \epsilon, P)_{max}\text{-code}\} \end{aligned}$$

- ϵ-capacity and Shannon capacity

$$\begin{aligned} C_\epsilon(P) &= \liminf_{n\to\infty} \frac{1}{n} \log M^*(n, \epsilon, P) \\ C(P) &= \lim_{\epsilon\downarrow 0} C_\epsilon(P) \end{aligned}$$

- Information capacity

$$C_i(P) = \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}:\mathbb{E}[\sum_{k=1}^n \mathsf{c}(X_k)]\le nP} I(X^n; Y^n)$$

- Information stability: Channel is information stable if for all (admissible) P, there exists a sequence of channel input distributions P_{X^n} such that the following two properties hold:

$$\frac{1}{n} i_{P_{X^n,Y^n}}(X^n; Y^n) \xrightarrow{i.P.} C_i(P) \tag{17.1}$$

$$\mathbb{P}[c(X^n) > P + \delta] \to 0 \qquad \forall \delta > 0\,. \tag{17.2}$$

Note: These are the usual definitions, except that in $C_i(P)$, we are permitted to maximize $I(X^n; Y^n)$ using input distributions from the constraint set $\{P_{X^n} : \mathbb{E}[\sum_{k=1}^n \mathsf{c}(X_k)] \le nP\}$ instead of the distributions supported on F_n.

Definition 17.4 (Admissible constraint)**.** P is an admissible constraint if $\exists x_0 \in \mathcal{A}$ s.t. $\mathsf{c}(x_0) \le P \Leftrightarrow \exists P_X : \mathbb{E}[\mathsf{c}(X)] \le P$. The set of admissible P's is denoted by $\mathcal{D}_\mathsf{c}$, and can be either in the form (P_0, ∞) or $[P_0, \infty)$, where $P_0 \triangleq \inf_{x\in\mathcal{A}} \mathsf{c}(x)$.

Clearly, if $P \notin \mathcal{D}_\mathsf{c}$, then there is no code (even a useless one, with 1 codeword) satisfying the input constraint. So in the remaining we always assume $P \in \mathcal{D}_\mathsf{c}$.

Proposition 17.1. *Define* $f(P) = \sup_{P_X:\mathbb{E}[\mathsf{c}(X)]\le P} I(X; Y)$*. Then*

1. *f is concave and non-decreasing. The domain of f,* $\operatorname{dom} f \triangleq \{x : f(x) > -\infty\} = \mathcal{D}_\mathsf{c}$.
2. *One of the following is true: $f(P)$ is continuous and finite on (P_0, ∞), or $f = \infty$ on (P_0, ∞).*

Furthermore, both properties hold for the function $P \mapsto C_i(P)$.

Proof. In (1) all statements are obvious, except for concavity, which follows from the concavity of $P_X \mapsto I(X;Y)$. For any P_{X_i} such that $\mathbb{E}[\mathsf{c}(X_i)] \le P_i, i = 0,1$, let $X \sim \bar{\lambda} P_{X_0} + \lambda P_{X_1}$. Then $\mathbb{E}[\mathsf{c}(X)] \le \bar{\lambda} P_0 + \lambda P_1$ and $I(X;Y) \ge \bar{\lambda} I(X_0;Y_0) + \lambda I(X_1;Y_1)$. Hence $f(\bar{\lambda} P_0 + \lambda P_1) \ge \bar{\lambda} f(P_0) + \lambda f(P_1)$. The second claim follows from concavity of $f(\cdot)$.

To extend these results to $C_i(P)$ observe that for every n

$$P \mapsto \frac{1}{n} \sup_{P_{X^n}:\mathbb{E}[\mathsf{c}(X^n)]\le P} I(X^n;Y^n)$$

is concave. Then taking $\liminf_{n\to\infty}$ the same holds for $C_i(P)$. □

An immediate consequence is that memoryless input is optimal for memoryless channel with separable cost, which gives us the single-letter formula of the information capacity:

Corollary 17.1 (Single-letterization)**.** *Information capacity of stationary memoryless channel with separable cost:*

$$C_i(P) = f(P) = \sup_{\mathbb{E}[\mathsf{c}(X)]\le P} I(X;Y).$$

Proof. $C_i(P) \ge f(P)$ is obvious by using $P_{X^n} = (P_X)^n$. For "$\le$", use the concavity of $f(\cdot)$, we have that for any P_{X^n},

$$I(X^n;Y^n) \le \sum_{j=1}^n I(X_j;Y_j) \le \sum_{j=1}^n f(\mathbb{E}[\mathsf{c}(X_j)]) \le n f\Big(\frac{1}{n}\sum_{j=1}^n \mathbb{E}[\mathsf{c}(X_j)]\Big) \le n f(P).$$

□

17.2 Capacity under input constraint $C(P) \overset{?}{=} C_i(P)$

Theorem 17.1 (General weak converse)**.**

$$C_\epsilon(P) \le \frac{C_i(P)}{1-\epsilon}$$

Proof. The argument is the same as before: Take any (n, M, ϵ, P)-code, $W \to X^n \to Y^n \to \hat{W}$. Apply Fano's inequality, we have

$$-h(\epsilon) + (1-\epsilon)\log M \le I(W;\hat{W}) \le I(X^n;Y^n) \le \sup_{P_{X^n}:\mathbb{E}[\mathsf{c}(X^n)]\le P} I(X^n;Y^n) \le n f(P)$$

□

Theorem 17.2 (Extended Feinstein's Lemma)**.** *Fix a random transformation $P_{Y|X}$. $\forall P_X, \forall F \subset \mathcal{X}, \forall \gamma > 0, \forall M$, there exists an $(M,\epsilon)_{\max}$-code with:*

- *Encoder satisfies the input constraint: $f : [M] \to F \subset \mathcal{X}$;*
- *Probability of error bound:*

$$\epsilon P_X(F) \le \mathbb{P}[i(X;Y) < \log\gamma] + \frac{M}{\gamma}$$

Note: when $F = \mathcal{X}$, it reduces to the original Feinstein's Lemma.

Proof. Similar to the proof of the original Feinstein's Lemma, define the preliminary decoding regions $E_c = \{y : i(c;y) \ge \log\gamma\}$ for all $c \in \mathcal{X}$. Sequentially pick codewords $\{c_1, \dots, c_M\}$ **from the set** F and the final decoding region $\{D_1, \dots, D_M\}$ where $D_j \triangleq E_{c_j} \backslash \cup_{k=1}^{j-1} D_k$. The stopping criterion is that M is maximal, i.e.,

$$\begin{aligned}
&\forall x_0 \in F, P_Y[E_{x_0} \backslash \cup_{j=1}^{M} D_j \big| X = x_0] < 1 - \epsilon \\
\Leftrightarrow\ &\forall x_0 \in \mathcal{X}, P_Y[E_{x_0} \backslash \cup_{j=1}^{M} D_j \big| X = x_0] < (1-\epsilon)\mathbf{1}[x_0 \in F] + \mathbf{1}[x_0 \in F^c] \\
\Rightarrow\ &\text{average over } x_0 \sim P_X,\ \mathbb{P}[\{i(X;Y) \ge \log\gamma\} \backslash \cup_{j=1}^{M} D_j] \le (1-\epsilon)P_X(F) + P_X(F^c) = 1 - \epsilon P_X(F)
\end{aligned}$$

From here, we can complete the proof by following the same steps as in the proof of Feinstein's lemma (Theorem 15.3). □

Theorem 17.3 (Achievability)**.** *For any information stable channel with input constraints and $P > P_0$ we have*

$$C(P) \ge C_i(P) \tag{17.3}$$

Proof. Let us consider a special case of the stationary memoryless channel (the proof for general information stable channel follows similarly). So we assume $P_{Y^n|X^n} = (P_{Y|X})^n$.

Fix $n \ge 1$. Since the channel is stationary memoryless, we have $P_{Y^n|X^n} = (P_{Y|X})^n$. Choose a P_X such that $\mathbb{E}[\mathsf{c}(X)] < P$, Pick $\log M = n(I(X;Y) - 2\delta)$ and $\log\gamma = n(I(X;Y) - \delta)$.

With the input constraint set $F_n = \{x^n : \frac{1}{n}\sum \mathsf{c}(x_k) \le P\}$, and iid input distribution $P_{X^n} = P_X^n$, we apply the extended Feinstein's Lemma, there exists an $(n, M, \epsilon_n, P)_{\max}$-code with the encoder satisfying input constraint F and the error probability

$$\epsilon_n \underbrace{P_X(F)}_{\to 1} \le \underbrace{P(i(X^n;Y^n) \le n(I(X;Y) - \delta))}_{\to 0 \text{ as } n\to\infty \text{ by WLLN and stationary memoryless assumption}} + \underbrace{\exp(-n\delta)}_{\to 0}$$

Also, since $\mathbb{E}[\mathsf{c}(X)] < P$, by WLLN, we have $P_{X^n}(F_n) = P(\frac{1}{n}\sum \mathsf{c}(x_k) \le P) \to 1$.

$$\begin{aligned}
&\epsilon_n(1 + o(1)) \le o(1) \\
\Rightarrow\ &\epsilon_n \to 0 \text{ as } n \to \infty \\
\Rightarrow\ &\forall\epsilon, \exists n_0, \text{ s.t. } \forall n \ge n_0, \exists(n, M, \epsilon_n, P)_{\max}\text{-code, with } \epsilon_n \le \epsilon
\end{aligned}$$

Therefore

$$\begin{aligned}
&C_\epsilon(P) \ge \frac{1}{n}\log M = I(X;Y) - 2\delta, \quad \forall\delta > 0, \forall P_X \text{ s.t. } \mathbb{E}[\mathsf{c}(X)] < P \\
\Rightarrow\ &C_\epsilon(P) \ge \sup_{P_X : \mathbb{E}[\mathsf{c}(X)] < P} \lim_{\delta\to 0}(I(X;Y) - 2\delta) \\
\Rightarrow\ &C_\epsilon(P) \ge \sup_{P_X : \mathbb{E}[\mathsf{c}(X)] < P} I(X;Y) = C_i(P-) = C_i(P)
\end{aligned}$$

where the last equality is from the continuity of C_i on (P_0, ∞) by Proposition 17.1. Notice that for general information stable channel, we just need to use the definition to show that $P(i(X^n;Y^n) \le n(C_i - \delta)) \to 0$, and all the rest follows. □

Theorem 17.4 (Shannon capacity)**.** *For an information stable channel with cost constraint and for any admissible constraint P we have*

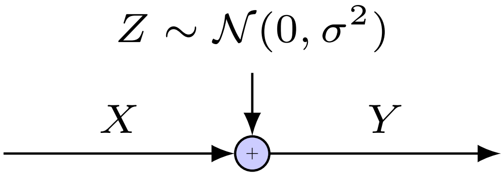
$$C(P) = C_i(P).$$

Proof. The case of $P = P_0$ is treated in the homework. So assume $P > P_0$. Theorem 17.1 shows $C_\epsilon(P) \le \frac{C_i(P)}{1-\epsilon}$, thus $C(P) \le C_i(P)$. On the other hand, from Theorem 17.3 we have $C(P) \ge C_i(P)$. □

Note: In homework, you will show that $C(P_0) = C_i(P_0)$ also holds, even though $C_i(P)$ may be discontinuous at P_0.

17.3 Applications

17.3.1 Stationary AWGN channel



Definition 17.5 (AWGN)**.** The additive Gaussian noise (AWGN) channel is a stationary memoryless additive-noise channel with separable cost constraint: $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{c}(x) = x^2$, $P_{Y|X}$ is given by $Y = X + Z$, where $Z \sim \mathcal{N}(0, \sigma^2) \perp\!\!\!\perp X$, and average power constraint $\mathbb{E}X^2 \le P$.

In other words, $Y^n = X^n + Z^n$, where $Z^n \sim \mathcal{N}(0, I_n)$.

Note: Here "white" = uncorrelated $\overset{\text{Gaussian}}{=}$ independent.
Note: Complex AWGN channel is similarly defined: $\mathcal{A} = \mathcal{B} = \mathbb{C}$, $\mathsf{c}(x) = |x|^2$, and $Z^n \sim \mathbb{C}\mathcal{N}(0, I_n)$

Theorem 17.5. *For stationary ($\mathbb{C}$)-AWGN channel, the channel capacity is equal to information capacity, and is given by:*

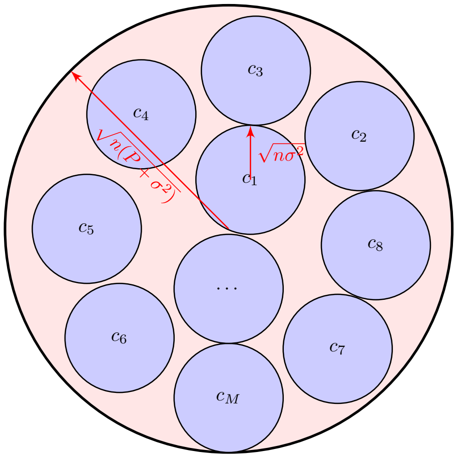
$$C(P) = C_i(P) = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) \quad \textit{for AWGN}$$
$$C(P) = C_i(P) = \log\left(1 + \frac{P}{\sigma^2}\right) \quad \textit{for } \mathbb{C}\textit{-AWGN}$$

Proof. By Corollary 17.1,

$$C_i = \sup_{P_X : \mathbb{E}X^2 \le P} I(X; X + Z)$$

Then use Theorem 4.6 (Gaussian saddlepoint) to conclude $X \sim \mathcal{N}(0, P)$ (or $\mathbb{C}\mathcal{N}(0, P)$) is the unique caid. □

Note: Since $Z^n \sim \mathcal{N}(0,\sigma^2)$, then with high probability, $\|Z^n\|_2$ concentrates around $\sqrt{n\sigma^2}$. Similarly, due the power constraint and the fact that $Z^n \perp\!\!\!\perp X^n$, the received vector Y^n lies in an ℓ_2-ball of radius $\sqrt{n(P+\sigma^2)}$. Since the noise can at most perturb the codeword by $\sqrt{n\sigma^2}$ in Euclidean distance, if we can pack M balls of radius $\sqrt{n\sigma^2}$ into the ℓ_2-ball of radius $\sqrt{n(P+\sigma^2)}$ centered at the origin, then this gives a good codebook and decision regions. The packing number is related to the volume ratio. Note that the volume of an ℓ_2-ball of radius r in $\mathbb{R}^n$ is given by $c_n r^n$ for some constant c_n. Then $\frac{c_n(n(P+\sigma^2))^{n/2}}{c_n(n\sigma^2)^{n/2}} = \left(1+\frac{P}{\sigma^2}\right)^{n/2}$. Take the log and divide by n, we get $\frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right)$.



Theorem 17.5 applies to Gaussian noise. What if the noise is non-Gaussian and how sensitive is the capacity formula $\frac{1}{2}\log(1+\text{SNR})$ to the Gaussian assumption? Recall the Gaussian saddlepoint result we have studied in Lecture 4 where we showed that for the same variance, Gaussian noise is the worst which shows that the capacity of any non-Gaussian noise is at least $\frac{1}{2}\log(1+\text{SNR})$. Conversely, it turns out the increase of the capacity can be controlled by how non-Gaussian the noise is (in terms of KL divergence). The following result is due to Ihara.

Theorem 17.6 (Additive Non-Gaussian noise)**.** *Let Z be a real-valued random variable independent of X and $\mathbb{E}Z^2 < \infty$. Let $\sigma^2 = \text{Var}\, Z$. Then*

$$\frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) \le \sup_{P_X:\mathbb{E}X^2\le P} I(X;X+Z) \le \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) + D(P_Z\|\mathcal{N}(\mathbb{E}Z,\sigma^2)).$$

Proof. Homework. □

Note: The quantity $D(P_Z\|\mathcal{N}(\mathbb{E}Z,\sigma^2))$ is sometimes called the *non-Gaussianness* of Z, where $\mathcal{N}(\mathbb{E}Z,\sigma^2)$ is a Gaussian with the same mean and variance as Z. So if Z has a non-Gaussian density, say, Z is uniform on $[0,1]$, then the capacity can only differ by a constant compared to AWGN, which still scales as $\frac{1}{2}\log\text{SNR}$ in the high-SNR regime. On the other hand, if Z is discrete, then $D(P_Z\|\mathcal{N}(\mathbb{E}Z,\sigma^2)) = \infty$ and indeed in this case one can show that the capacity is infinite because the noise is "too weak".

17.3.2 Parallel AWGN channel

Definition 17.6 (Parallel AWGN)**.** A parallel AWGN channel with L branches is defined as follows: $\mathcal{A} = \mathcal{B} = \mathbb{R}^L$; $\mathsf{c}(x) = \sum_{k=1}^{L}|x_k|^2$; $P_{Y^L|X^L} : Y_k = X_k + Z_k$, for $k = 1,\ldots,L$, and $Z_k \sim \mathcal{N}(0,\sigma_k^2)$ are independent for each branch.

Theorem 17.7 (Waterfilling)**.** *The capacity of L-parallel AWGN channel is given by*

$$C = \frac{1}{2}\sum_{j=1}^{L}\log^+\frac{T}{\sigma_j^2}$$

where $\log^+(x) \triangleq \max(\log x, 0)$, and $T \ge 0$ is determined by

$$P = \sum_{j=1}^{L}|T-\sigma_j^2|^+$$

Proof.

$$\begin{aligned}
C_i(P) &= \sup_{P_{X^L}:\sum \mathbb{E}[X_i^2]\le P} I(X^L;Y^L) \\
&\le \sup_{\sum P_k \le P, P_k\ge 0} \sum_{k=1}^{L} \sup_{\mathbb{E}[X_k^2]\le P_k} I(X_k;Y_k) \\
&= \sup_{\sum P_k \le P, P_k\ge 0} \sum_{k=1}^{L} \frac{1}{2}\log(1+\frac{P_k}{\sigma_k^2})
\end{aligned}$$

with equality if $X_k \sim \mathcal{N}(0,P_k)$ are independent. So the question boils down to the last maximization problem – **power allocation**: Denote the Lagragian multipliers for the constraint $\sum P_k \le P$ by λ and for the constraint $P_k \ge 0$ by μ_k. We want to solve $\max \sum \frac{1}{2}\log(1+\frac{P_k}{\sigma_k^2}) - \mu_k P_k + \lambda(P - \sum P_k)$. First-order condition on P_k gives that

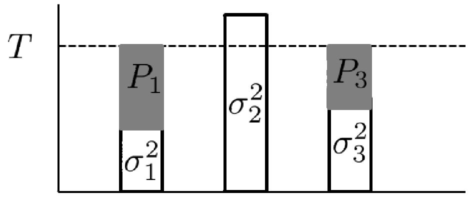
$$\frac{1}{2}\frac{1}{\sigma_k^2 + P_k} = \lambda - \mu_k, \quad \mu_k P_k = 0$$

therefore the optimal solution is

$$P_k = |T - \sigma_k^2|^+, \quad T \text{ is chosen such that } P = \sum_{k=1}^{L} |T-\sigma_k^2|^+$$

□

Note: The figure illustrates the power allocation via water-filling. In this particular case, the second branch is too noisy (σ_2 too big) such that it is better be discarded, i.e., the assigned power is zero.



waterfilling across 3 parallel channels

Note: [Significance of the waterfilling theorem] In the high SNR regime, the capacity for 1 AWGN channel is approximately $\frac{1}{2}\log P$, while the capacity for L parallel AWGN channel is approximately $\frac{L}{2}\log(\frac{P}{L}) \approx \frac{L}{2}\log P$ for large P. This L-fold increase in capacity at high SNR regime leads to the powerful technique of spatial multiplexing in MIMO.

Also notice that this gain does not come from multipath diversity. Consider the scheme that a single stream of data is sent through every parallel channel simultaneously, with multipath diversity, the effective noise level is reduced to $\frac{1}{L}$, and the capacity is approximately $\log(LP)$, which is much smaller than $\frac{L}{2}\log(\frac{P}{L})$ for P large.

17.4* Non-stationary AWGN

Definition 17.7 (Non-stationary AWGN)**.** A non-stationary AWGN channel is defined as follows: $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{c}(x) = x^2$, $P_{Y_j|X_j} : Y_j = X_j + Z_j$, where $Z_j \sim \mathcal{N}(0,\sigma_j^2)$.

Theorem 17.8. *Assume that for every T the following limits exist:*

$$\tilde{C}_i(T) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \frac{1}{2}\log^{+}\frac{T}{\sigma_j^2}$$
$$\tilde{P}(T) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} |T-\sigma_j^2|^{+}$$

then the capacity of the non-stationary AWGN channel is given by the parameterized form: $C(T) = \tilde{C}_i(T)$ with input power constraint $\tilde{P}(T)$.

Proof. Fix $T > 0$. Then it is clear from the waterfilling solution that

$$\sup I(X^n;Y^n) = \sum_{j=1}^{n}\frac{1}{2}\log^{+}\frac{T}{\sigma_j^2}\,, \tag{17.4}$$

where supremum is over all P_{X^n} such that

$$\mathbb{E}[\mathsf{c}(X^n)] \le \frac{1}{n}\sum_{j=1}^{n}|T-\sigma_j^2|^{+}\,. \tag{17.5}$$

Now, by assumption, the LHS of (17.5) converges to $\tilde{P}(T)$. Thus, we have that for every $\delta > 0$

$$C_i(\tilde{P}(T)-\delta) \le \tilde{C}_i(T) \tag{17.6}$$
$$C_i(\tilde{P}(T)+\delta) \ge \tilde{C}_i(T) \tag{17.7}$$

Taking $\delta \to 0$ and invoking continuity of $P \mapsto C_i(P)$, we get that the information capacity satisfies

$$C_i(\tilde{P}(T)) = \tilde{C}_i(T)\,.$$

The channel is information stable. Indeed, from (16.17)

$$\mathrm{Var}(i(X_j;Y_j)) = \frac{\log^2 e}{2}\frac{P_j}{P_j+\sigma_j^2} \le \frac{\log^2 e}{2}$$

and thus

$$\sum_{j=1}^{n}\frac{1}{n^2}\mathrm{Var}(i(X_j;Y_j)) < \infty\,.$$

From here information stability follows via Theorem 16.9. □

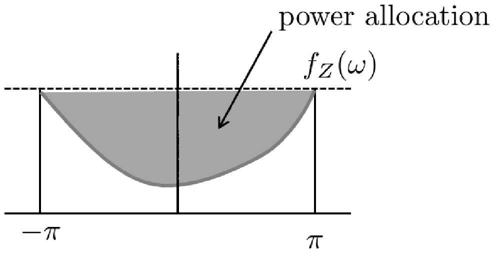
Note: Non-stationary AWGN is primarily interesting due to its relationship to the stationary Additive Colored Gaussian noise channel in the following discussion.

17.5* Stationary Additive Colored Gaussian noise channel

Definition 17.8 (Additive colored Gaussian noise channel). An Additive Colored Gaussian noise channel is defined as follows: $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{c}(x) = x^2$, $P_{Y_j|X_j} : Y_j = X_j + Z_j$, where Z_j is a stationary Gaussian process with spectral density $f_Z(\omega) > 0, \omega \in [-\pi,\pi]$.

Theorem 17.9. *The capacity of stationary ACGN channel is given by the parameterized form:*

$$C(T) = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+ \frac{T}{f_Z(\omega)} d\omega$$
$$P(T) = \frac{1}{2\pi} \int_0^{2\pi} \left| T - f_Z(\omega) \right|^+ d\omega$$



waterfilling across spectrum for stationary ACGN channel

Proof. Take $n \geq 1$, consider the diagonalization of the covariance matrix of Z^n:

$$Cov(Z^n) = \Sigma = U^* \widetilde{\Sigma} U, \text{ such that } \widetilde{\Sigma} = diag(\sigma_1, \dots, \sigma_n)$$

Since $Cov(Z^n)$ is positive semi-definite, U is a unitary matrix. Define $\widetilde{X}^n = UX^n$ and $\widetilde{Y}^n = UY^n$, the channel between $\widetilde{X}^n$ and $\widetilde{Y}^n$ is thus

$$\widetilde{Y}^n = \widetilde{X}^n + UZ^n,$$
$$Cov(UZ^n) = UCov(Z^n)U^* = \widetilde{\Sigma}$$

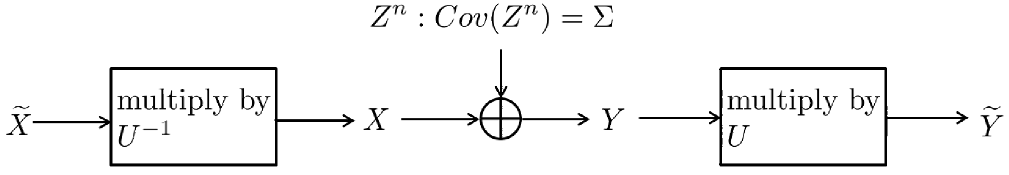
Therefore we have the equivalent channel as follows:

$$\widetilde{Y}^n = \widetilde{X}^n + \widetilde{Z}^n, \ \widetilde{Z}_j^n \sim \mathcal{N}(0, \sigma_j^2) \text{ indep across } j$$

By Theorem 17.8, we have that

$$\widetilde{C} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n \log^+ \frac{T}{\sigma_j^2} = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+ \frac{T}{f_Z(\omega)} d\omega. \text{ (by Szegö, Theorem 5.6)}$$
$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n |T - \sigma_j^2|^+ = P(T)$$

Finally since U is unitary, $C = \widetilde{C}$.



stationary additive Gausian noise channel □

Note: Noise is born white, the colored noise is essentially due to some filtering.

17.6* Additive White Gaussian Noise channel with Intersymbol Interference

Definition 17.9 (AWGN with ISI)**.** An AWGN channel with ISI is defined as follows: $\mathcal{A} = \mathcal{B} = \mathbb{R}$, $\mathsf{c}(x) = x^2$, and the channel law $P_{Y^n|X^n}$ is given by

$$Y_k = \sum_{j=1}^{n} h_{k-j} X_j + Z_k \,, \qquad k = 1, \dots, n$$

where $Z_k \sim \mathcal{N}(0,1)$ is white Gaussian noise, $\{h_k, k = -\infty, \dots, \infty\}$ are coefficients of a discrete-time channel filter.

Theorem 17.10. *Suppose that the sequence $\{h_k\}$ is an inverse Fourier transform of a frequency response $H(\omega)$:*

$$h_k = \frac{1}{2\pi} \int_0^{2\pi} e^{i\omega k} H(\omega) d\omega \,.$$

Assume also that $H(\omega)$ is a continuous function on $[0, 2\pi]$. Then the capacity of the AWGN channel with ISI is given by

$$C(T) = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \log^+ (T|H(\omega)|^2) d\omega$$

$$P(T) = \frac{1}{2\pi} \int_0^{2\pi} \left| T - \frac{1}{|H(\omega)|^2} \right|^+ d\omega$$

Proof. (Sketch) At the decoder apply the inverse filter with frequency response $\omega \mapsto \frac{1}{H(\omega)}$. The equivalent channel then becomes a stationary colored-noise Gaussian channel:

$$\tilde{Y}_j = X_j + \tilde{Z}_j \,,$$

where $\tilde{Z}_j$ is a stationary Gaussian process with spectral density

$$f_{\tilde{Z}}(\omega) = \frac{1}{|H(\omega)|^2} \,.$$

Then apply Theorem 17.9 to the resulting channel.

Remark: to make the above argument rigorous one must simply carefully analyze the non-zero error introduced by truncating the deconvolution filter to finite n. □

17.7* Gaussian channels with amplitude constraints

We have examined some classical results of additive Gaussian noise channels. In the following, we will list some more recent results without proof.

Theorem 17.11 (Amplitude-constrained capacity of AWGN channel)**.** *For an AWGN channel $Y_i = X_i + Z_i$ with amplitude constraint $|X_i| \le A$ and energy constraint $\sum_{i=1}^n X_i^2 \le nP$, we denote the capacity by:*

$$C(A, P) = \max_{P_X : |X| \le A, \mathbb{E}|X|^2 \le P} I(X; X + Z).$$

Capacity achieving input distribution P_X^* is discrete, with finitely many atoms on $[-A, A]$. Moreover, the convergence speed of $\lim_{A \to \infty} C(A, P) = \frac{1}{2}\log(1 + P)$ is of the order $e^{-O(A^2)}$.

For details, see [Smi71] and [PW14, Section III].

17.8* Gaussian channels with fading

Fading channels are often used to model the urban signal propagation with multipath or shadowing. The received signal Y_i is modeled to be affected by multiplicative fading coefficient H_i and additive noise Z_i:

$$Y_i = H_i X_i + Z_i, \quad Z_i \sim \mathcal{N}(0,1)$$

In the coherent case (also known as CSIR – for channel state information at the receiver), the receiver has access to the channel state information of H_i, i.e. the channel output is effectively (Y_i, H_i). Whenever H_j is a stationary ergodic process, we have the channel capacity given by:

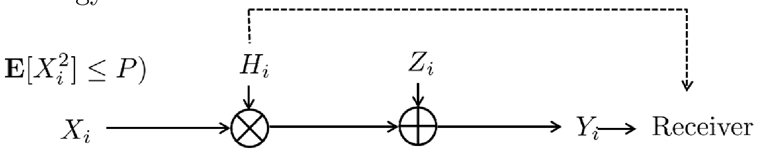
$$C(P) = \mathbb{E}[\frac{1}{2}\log(1 + P|H|^2)]$$

and the capacity achieving input distribution is the usual $P_X = \mathcal{N}(0,P)$. Note that the capacity $C(P)$ is in the order of $\log(P)$ and we call the channel "energy efficient".

In the non-coherent case where the receiver does not have the information of H_i, no simple expression for the channel capacity is known. It is known, however, that the capacity achieving input distribution is discrete, and the capacity

$$C(P) = O(\log\log P), \qquad P \to \infty \tag{17.8}$$

This channel is said to be "energy inefficient".



Fading channel

With introduction of multiple antenna channels, there are endless variations, theoretical open problems and practically unresolved issues in the topic of fading channels. We recommend consulting textbook [TV05] for details.

§ 18. Lattice codes (by O. Ordentlich)

Consider the n-dimensional additive white Gaussian noise (AWGN) channel

$$\mathbf{Y} = \mathbf{X} + \mathbf{Z}$$

where $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I}_{n\times n})$ is statistically independent of the input $\mathbf{X}$. Our goal is to communicate reliably over this channel, under the power constraint

$$\frac{1}{n}\|\mathbf{X}\|^2 \le \mathsf{SNR}$$

where SNR is the *signal-to-noise-ratio.* The capacity of the AWGN channel is

$$C = \tfrac{1}{2}\log(1 + \mathsf{SNR}) \text{ bits/channel use,}$$

and is achieved with high probability by a codebook drawn at random from the Gaussian i.i.d. ensemble. However, a typical codebook from this ensemble has very little structure, and is therefore not applicable for practical systems. A similar problem occurs in discrete additive memoryless stationary channels, e.g., BSC, where most members of the capacity achieving i.i.d. uniform codebook ensemble have no structure. In the discrete case, engineers resort to linear codes to circumvent the lack of structure. Lattice codes are the Euclidean space counterpart of linear codes, and as we shall see, enable to achieve the capacity of the AWGN channel with much more structure than random codes. In fact, we will construct a lattice code with rate that approaches $\frac{1}{2}\log(1+\mathsf{SNR})$ that is guaranteed to achieve small error probability for essentially all additive noise channels with the same noise second moment. More precisely, our scheme will work if the noise vector $\mathbf{Z}$ is *semi norm-ergodic.*

Definition 18.1. We say that a sequence in n of random noise vectors $\mathbf{Z}^{(n)}$ of length n with (finite) effective variance $\sigma^2_{\mathbf{Z}} \triangleq \frac{1}{n}\mathbb{E}\|\mathbf{Z}^{(n)}\|^2$, is *semi norm-ergodic* if for any $\epsilon, \delta > 0$ and n large enough

$$\Pr\left(\mathbf{Z}^{(n)} \notin \mathcal{B}(\sqrt{(1+\delta)n\sigma^2_{\mathbf{Z}}}\right) \le \epsilon, \tag{18.1}$$

where $\mathcal{B}(r)$ is an n-dimensional ball of radius r.

18.1 Lattice Definitions

A lattice Λ is a discrete subgroup of $\mathbb{R}^n$ which is closed under reflection and real addition. Any lattice Λ in $\mathbb{R}^n$ is spanned by some $n \times n$ matrix $\mathbf{G}$ such that

$$\Lambda = \{\mathbf{t} = \mathbf{G}\mathbf{a} : \mathbf{a} \in \mathbb{Z}^n\}.$$

We will assume $\mathbf{G}$ is full-rank. Denote the nearest neighbor quantizer associated with the lattice Λ by

$$Q_\Lambda(\mathbf{x}) \triangleq \arg\min_{\mathbf{t}\in\Lambda} \|\mathbf{x}-\mathbf{t}\|, \tag{18.2}$$

where ties are broken in a systematic manner. We define the modulo operation w.r.t. a lattice Λ as

$$[\mathbf{x}] \bmod \Lambda \triangleq \mathbf{x} - Q_\Lambda(\mathbf{x}),$$

and note that it satisfies the distributive law,

$$\big[[\mathbf{x}] \bmod \Lambda + \mathbf{y}\big] \bmod \Lambda = [\mathbf{x}+\mathbf{y}] \bmod \Lambda.$$

The basic Voronoi region of Λ, denoted by $\mathcal{V}$, is the set of all points in $\mathbb{R}^n$ which are quantized to the zero vector. The systematic tie-breaking in (18.2) ensures that

$$\biguplus_{\mathbf{t}\in\Lambda} (\mathcal{V}+\mathbf{t}) = \mathbb{R}^n,$$

where $\uplus$ denotes disjoint union. Thus, $\mathcal{V}$ is a *fundamental cell* of Λ.

Definition 18.2. A measurable set $S \in \mathbb{R}^n$ is called a *fundamental cell* of Λ if

$$\biguplus_{\mathbf{t}\in\Lambda} (S+\mathbf{t}) = \mathbb{R}^n.$$

We denote the volume of a set $S \in \mathbb{R}^n$ by $\mathrm{Vol}(S)$.

Proposition 18.1. *If S is a fundamental cell of Λ, then* $\mathrm{Vol}(S) = \mathrm{Vol}(\mathcal{V})$. *Furthermore*

$$S \bmod \Lambda = \{[\mathbf{s}] \bmod \Lambda \;:\; \mathbf{s}\in S\} = \mathcal{V}.$$

Proof ([Zam14]). For any $\mathbf{t}\in\Lambda$ define

$$\mathcal{A}_\mathbf{t} \triangleq S \cap (\mathbf{t}+\mathcal{V}); \quad \mathcal{D}_\mathbf{t} \triangleq \mathcal{V} \cap (\mathbf{t}+S).$$

Note that

$$\begin{aligned}\mathcal{D}_\mathbf{t} &= [(-\mathbf{t}+\mathcal{V})\cap S] + \mathbf{t}\\ &= \mathcal{A}_{-\mathbf{t}} + \mathbf{t}.\end{aligned}$$

Thus

$$\mathrm{Vol}(S) = \sum_{\mathbf{t}\in\Lambda} \mathrm{Vol}(\mathcal{A}_\mathbf{t}) = \sum_{\mathbf{t}\in\Lambda} \mathrm{Vol}(\mathcal{A}_{-\mathbf{t}}+\mathbf{t}) = \sum_{\mathbf{t}\in\Lambda} \mathrm{Vol}(\mathcal{D}_\mathbf{t}) = \mathrm{Vol}(\mathcal{V}).$$

Moreover

$$S = \biguplus_{\mathbf{t}\in\Lambda} \mathcal{A}_\mathbf{t} = \biguplus_{\mathbf{t}\in\Lambda} \mathcal{A}_{-\mathbf{t}} = \biguplus_{\mathbf{t}\in\Lambda} \mathcal{D}_\mathbf{t} - \mathbf{t},$$

and therefore

$$[S] \bmod \Lambda = \biguplus_{\mathbf{t}\in\Lambda} \mathcal{D}_\mathbf{t} = \mathcal{V}.$$

□

Corollary 18.1. *If S is a fundamental cell of a lattice Λ with generating matrix $\mathbf{G}$, then* $\mathrm{Vol}(S) = |\det(\mathbf{G})|$. *In Particular,* $\mathrm{Vol}(\mathcal{V}) = |\det(\mathbf{G})|$.

Proof. Let $\mathcal{P} = \mathbf{G}\cdot[0,1)^n$ and note that it is a fundamental cell of Λ as $\mathbb{R}^n = \mathbb{Z}^n + [0,1)^n$. The claim now follows from Proposition 18.1 since $\mathrm{Vol}(\mathcal{P}) = |\det(\mathbf{G})|\cdot\mathrm{Vol}([0,1)^n) = |\det(\mathbf{G})|$. □

Definition 18.3 (Lattice decoder)**.** A lattice decoder w.r.t. the lattice Λ returns for every $\mathbf{y}\in\mathbb{R}^n$ the point $Q_\Lambda(\mathbf{y})$.

Remark 18.1. Recall that for linear codes, the ML decoder merely consisted of mapping syndromes to shifts. Similarly, it can be shown that a lattice decoder can be expressed as

$$Q_\Lambda(\mathbf{y}) = \mathbf{y} - g_{\mathrm{synd}}\left([\mathbf{G}^{-1}\mathbf{y}] \bmod 1\right), \tag{18.3}$$

for some $g_{\mathrm{synd}} : [0,1)^n \mapsto \mathbb{R}^n$, where the $\bmod 1$ operation above is to be understood as componentwise modulo reduction. Thus, a lattice decoder is indeed much more "structured" than ML decoder for a random code.

Note that for an additive channel $\mathbf{Y} = \mathbf{X} + \mathbf{Z}$, if $\mathbf{X}\in\Lambda$ we have that

$$P_e = \Pr\left(Q_\Lambda(\mathbf{Y}) \neq \mathbf{X}\right) = \Pr(\mathbf{Z}\notin\mathcal{V}). \tag{18.4}$$

We therefore see that the resilience of a lattice to additive noise is dictated by its Voronoi region. Since we know that $\mathbf{Z}$ will be inside a ball of radius $\sqrt{n(1+\delta)}$ with high probability, we would like the Voronoi region to be as close as possible to a ball. We define the effective radius of a lattice, denoted $r_{\mathrm{eff}}(\Lambda)$ as the radius of a ball with the same volume as $\mathcal{V}$, namely $\mathrm{Vol}\left(\mathcal{B}\left(r_{\mathrm{eff}}(\Lambda)\right)\right) = \mathrm{Vol}(\mathcal{V})$.

Definition 18.4 (Goodness for coding)**.** A sequence of lattices $\Lambda^{(n)}$ with growing dimension, satisfying

$$\lim_{n\to\infty}\frac{r_{\mathrm{eff}}^2(\Lambda^{(n)})}{n} = \Phi$$

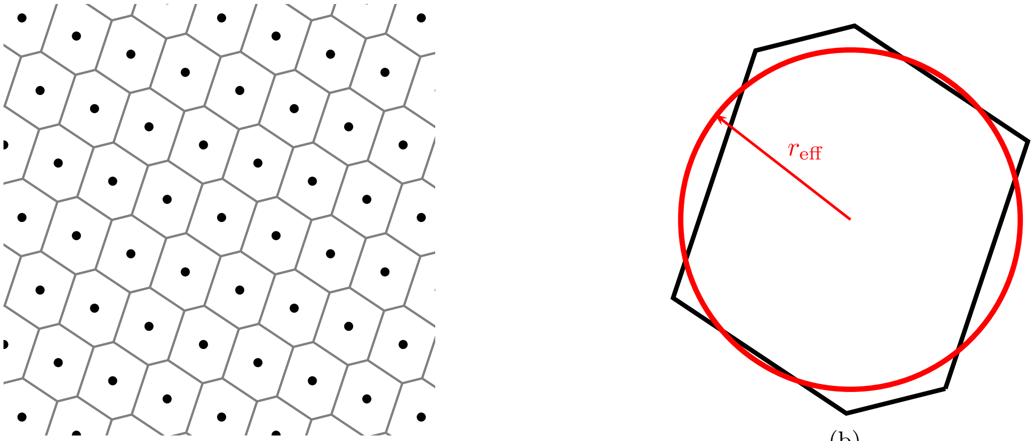
for some $\Phi > 0$, is called *good for channel coding* if for any additive semi norm-ergodic noise sequence $\mathbf{Z}^{(n)}$ with effective variance $\sigma_{\mathbf{Z}}^2 = \frac{1}{n}\mathbb{E}\|\mathbf{Z}\|^2 < \Phi$

$$\lim_{n\to\infty}\Pr\left(\mathbf{Z}^{(n)}\notin\mathcal{V}^{(n)}\right) = 0.$$

An alternative interpretation of this property, is that for a sequence $\Lambda^{(n)}$ that is good for coding, for any $0<\delta<1$ holds

$$\lim_{n\to\infty}\frac{\mathrm{Vol}\left(\mathcal{B}\left((1-\delta)r_{\mathrm{eff}}(\Lambda^{(n)})\right)\cap\mathcal{V}^{(n)}\right)}{\mathrm{Vol}\left(\mathcal{B}\left((1-\delta)r_{\mathrm{eff}}(\Lambda^{(n)})\right)\right)} = 1.$$

Roughly speaking, the Voronoi region of a lattice that is good for coding is as resilient to semi norm-ergodic noise as a ball with the same volume.



(a)

(b)

Figure 18.1: (a) shows a lattice in $\mathbb{R}^2$, and (b) shows its Voronoi region and the corresponding effective ball.

18.2 First Attempt at AWGN Capacity

Assume we have a lattice $\Lambda \subset \mathbb{R}^n$ with $r_{\text{eff}}(\Lambda) = \sqrt{n(1+\delta)}$ that is good for coding, and we would like to use it for communicating over an additive noise channel. In order to meet the power constraint, we must first intersect Λ, or a shifted version of Λ, with some compact set S that enforces the power constraint. The most obvious choice is taking S to be a ball with radius $\sqrt{n\mathsf{SNR}}$, and take some shift $\mathbf{v} \in \mathbb{R}^n$, such that the codebook

$$\mathcal{C} = (\mathbf{v} + \Lambda) \bigcap \mathcal{B}(\sqrt{n\mathsf{SNR}}) \tag{18.5}$$

satisfies the power constraint. Moreover [Loe97], there exist a shift $\mathbf{v}$ such that

$$\begin{aligned} |\mathcal{C}| &\geq \frac{\text{Vol}\,(S)}{\text{Vol}(\mathcal{V})} \\ &= \left(\frac{\sqrt{n\mathsf{SNR}}}{r_{\text{eff}}(\Lambda)}\right)^n \\ &= 2^{\frac{n}{2}(\log(\mathsf{SNR}) - \log(1+\delta))}. \end{aligned}$$

To see this, let $\mathbf{V} \sim \text{Uniform}(\mathcal{V})$, and write the expected size of $|\mathcal{C}|$ as

$$\begin{aligned} \mathbb{E}|\mathcal{C}| &= \mathbb{E} \sum_{\mathbf{t}\in\Lambda} \mathbb{1}((\mathbf{t} + \mathbf{V}) \in S) \\ &= \frac{1}{\text{Vol}(\mathcal{V})} \int_{\mathbf{v}\in\mathcal{V}} \sum_{\mathbf{t}\in\Lambda} \mathbb{1}((\mathbf{t} + \mathbf{v}) \in S) d\mathbf{v} \\ &= \frac{1}{\text{Vol}(\mathcal{V})} \int_{\mathbf{x}\in\mathbb{R}^n} \mathbb{1}(\mathbf{x} \in S) d\mathbf{x} \\ &= \frac{\text{Vol}(S)}{\text{Vol}(\mathcal{V})}. \end{aligned} \tag{18.6}$$

For decoding, we will simply apply the lattice decoder $Q_\Lambda(\mathbf{Y} - \mathbf{v})$ on the shifted output. Since $\mathbf{Y} - \mathbf{v} = \mathbf{t} + \mathbf{Z}$ for some $\mathbf{t} \in \Lambda$, the error probability is

$$P_e = \Pr\left(Q_\Lambda(\mathbf{Y} - \mathbf{v}) \neq \mathbf{t}\right) = \Pr(\mathbf{Z} \notin \mathcal{V}).$$

Since Λ is good for coding and $\frac{r_{\text{eff}}^2(\Lambda)}{n} = (1+\delta) > \frac{1}{n}\mathbb{E}\|\mathbf{Z}\|^2$, the error probability of this scheme over an additive semi norm-ergodic noise channel will vanish with n. Taking $\delta \to 0$ we see that any rate $R < \frac{1}{2}\log(\mathsf{SNR})$ can be achieved reliably. Note that for this coding scheme (encoder+decoder) the average error probability and the maximal error probability are the same.

The construction above gets us close to the AWGN channel capacity. We note that a possible reason for the loss of +1 in the achievable rate is the suboptimality of the lattice decoder for the codebook $\mathcal{C}$. The lattice decoder assumes all points of Λ were equally likely to be transmitted. However, in $\mathcal{C}$ only lattice points inside the ball can be transmitted. Indeed, it was shown [UR98] that if one replaces the lattice decoder with a decoder that takes the shaping region into account, there exist lattices and shifts for which the codebook $(\mathbf{v}+\Lambda)\bigcap\mathcal{B}(\sqrt{n\mathsf{SNR}})$ is capacity achieving. The main drawback of this approach is that the decoder no longer exploits the full structure of the lattice, so the advantages of using a lattice code w.r.t. some typical member of the Gaussian i.i.d. ensemble are not so clear anymore.

18.3 Nested Lattice Codes/Voronoi Constellations

A lattice Λ_c is said to be nested in Λ_f if $\Lambda_c \subset \Lambda_f$. The lattice Λ_c is referred to as the coarse lattice and Λ_f as the fine lattice. The *nesting ratio* is defined as

$$\Gamma(\Lambda_f, \Lambda_c) \triangleq \left(\frac{\text{Vol}(\mathcal{V}_c)}{\text{Vol}(\mathcal{V}_f)}\right)^{1/n} \tag{18.7}$$

A *nested lattice code* (sometimes also called "Voronoi constellation") based on the nested lattice pair $\Lambda_c \subset \Lambda_f$ is defined as [CS83, For89, EZ04]

$$\mathcal{L} \triangleq \Lambda_f \cap \mathcal{V}_c. \tag{18.8}$$

Proposition 18.2.

$$|\mathcal{L}| = \frac{\text{Vol}(\mathcal{V}_c)}{\text{Vol}(\mathcal{V}_f)}.$$

Thus, the codebook $\mathcal{L}$ has rate $R = \frac{1}{n}\log|\mathcal{L}| = \log\Gamma(\Lambda_f, \Lambda_c)$.

Proof. First note that

$$\Lambda_f \triangleq \biguplus_{\mathbf{t}\in\mathcal{L}}(\mathbf{t}+\Lambda_c).$$

Let

$$S \triangleq \biguplus_{\mathbf{t}\in\mathcal{L}}(\mathbf{t}+\mathcal{V}_f),$$

and note that

$$\begin{aligned}\mathbb{R}^n &= \biguplus_{\mathbf{b}\in\Lambda_f}(\mathbf{b}+\mathcal{V}_f) \\ &= \biguplus_{\mathbf{a}\in\Lambda_c}\biguplus_{\mathbf{t}\in\mathcal{L}}(\mathbf{a}+\mathbf{t}+\mathcal{V}_f) \\ &= \biguplus_{\mathbf{a}\in\Lambda_c}\left(\mathbf{a}+\left(\biguplus_{\mathbf{t}\in\mathcal{L}}(\mathbf{t}+\mathcal{V}_f)\right)\right) \\ &= \biguplus_{\mathbf{a}\in\Lambda_c}(\mathbf{a}+S).\end{aligned}$$

Thus, S is a fundamental cell of Λ_c, and we have

$$\mathrm{Vol}(\mathcal{V}_c) = \mathrm{Vol}(S) = |\mathcal{L}| \cdot \mathrm{Vol}(\mathcal{V}_f).$$

□

We will use the codebook $\mathcal{L}$ with a standard lattice decoder, ignoring the fact that only points in $\mathcal{V}_c$ were transmitted. Therefore, the resilience to noise will be dictated mainly by Λ_f. The role of the coarse lattice Λ_c is to perform *shaping*. In order to maximize the rate of the codebook $\mathcal{L}$ without violating the power constraint, we would like $\mathcal{V}_c$ to have the maximal possible volume, under the constraint that the average power of a transmitted codeword is no more than $n\mathsf{SNR}$.

The average transmission power of the codebook $\mathcal{L}$ is related to a quantity called the *second moment of a lattice*. Let $\mathbf{U} \sim \mathrm{Uniform}(\mathcal{V})$. The second moment of Λ is defined as $\sigma^2(\Lambda) \triangleq \frac{1}{n}\mathbb{E}\|\mathbf{U}\|^2$. Let $\mathbf{W} \sim \mathrm{Uniform}(\mathcal{B}(r_{\text{eff}}(\Lambda))$. By the isoperimetric inequality [Zam14]

$$\sigma^2(\Lambda) \geq \frac{1}{n}\mathbb{E}\|\mathbf{W}\|^2 = \frac{r_{\text{eff}}^2(\Lambda)}{n+2}.$$

A lattice Λ exhibits a good tradeoff between average power and volume if its second moment is close to that of $\mathcal{B}(r_{\text{eff}}(\Lambda)$.

Definition 18.5 (Goodness for MSE quantization)**.** A sequence of lattices $\Lambda^{(n)}$ with growing dimension, is called *good for MSE quantization* if

$$\lim_{n\to\infty} \frac{n\sigma^2\left(\Lambda^{(n)}\right)}{r_{\text{eff}}^2\left(\Lambda^{(n)}\right)} = 1.$$

Remark 18.2. Note that both "goodness for coding" and "goodness for quantization" are scale invariant properties: if Λ satisfy them, so does $\alpha\Lambda$ for any $\alpha \in \mathbb{R}$.

Theorem 18.1 ([OE15])**.** *If Λ is good for MSE quantization and $\mathbf{U} \sim \mathrm{Uniform}(\mathcal{V})$, then $\mathbf{U}$ is semi norm-ergodic. Furthermore, if $\mathbf{Z}$ is semi norm-ergodic and statistically independent of $\mathbf{U}$, then for any $\alpha, \beta \in \mathbb{R}$ the random vector $\alpha\mathbf{U} + \beta\mathbf{Z}$ is semi norm-ergodic.*

Theorem 18.2 ([ELZ05, OE15])**.** *For any finite nesting ratio $\Gamma(\Lambda_f, \Lambda_c)$, there exist a nested lattice pair $\Lambda_c \subset \Lambda_f$ where the coarse lattice Λ_c is good for MSE quantization and the fine lattice Λ_f is good for coding.*

We now describe the Mod-Λ coding scheme introduced by Erez and Zamir [EZ04]. Let $\Lambda_c \subset \Lambda_f$ be a nested lattice pair, where the coarse lattice is good for MSE quantization and has $\sigma^2(\Lambda_c) = \mathsf{SNR}(1-\epsilon)$, whereas the fine lattice is good for coding and has $r_{\text{eff}}^2(\Lambda_f) = n\frac{\mathsf{SNR}}{1+\mathsf{SNR}}(1+\epsilon)$. The rate is therefore

$$\begin{aligned} R &= \frac{1}{n}\log\left(\frac{\mathrm{Vol}(\mathcal{V}_c)}{\mathrm{Vol}(\mathcal{V}_f)}\right) \\ &= \frac{1}{2}\log\left(\frac{r_{\text{eff}}^2(\Lambda_c)}{r_{\text{eff}}^2(\Lambda_f)}\right) \\ &\to \frac{1}{2}\log\left(\frac{\mathsf{SNR}(1-\epsilon)}{\frac{\mathsf{SNR}}{1+\mathsf{SNR}}(1+\epsilon)}\right) \qquad (18.9) \\ &\to \frac{1}{2}\log\left(1+\mathsf{SNR}\right), \end{aligned}$$

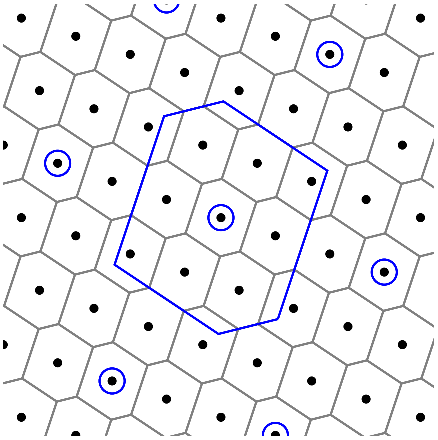


Figure 18.2: An example of a nested lattice code. The points and Voronoi region of Λ_c are plotted in blue, and the points of the fine lattice in black.

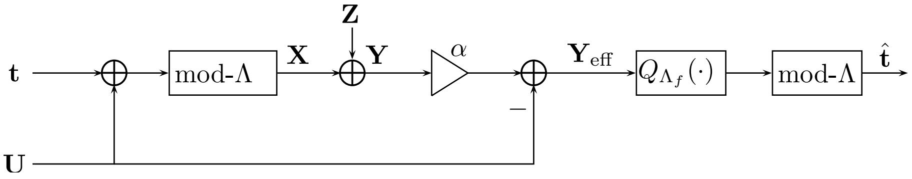


Figure 18.3: Schematic illustration of the Mod-Λ scheme.

where in (18.9) we have used the goodness of Λ_c for MSE quantization, that implies $\frac{r^2_{\text{eff}}(\Lambda_c)}{n} \to \sigma^2(\Lambda_c)$. The scheme also uses common randomness, namely a dither vector $\mathbf{U} \sim \text{Uniform}(\mathcal{V}_c)$ statistically independent of everything, known to both the transmitter and the receiver. In order to transmit a message $w \in [1, \dots, 2^{nR}]$ the encoder maps it to the corresponding point $\mathbf{t} = \mathbf{t}(w) \in \mathcal{L}$ and transmits

$$\mathbf{X} = [\mathbf{t} + \mathbf{U}] \bmod \Lambda. \tag{18.10}$$

Lemma 18.1 (Crypto Lemma)**.** *Let Λ be a lattice in $\mathbb{R}^n$, let $\mathbf{U} \sim \text{Uniform}(\mathcal{V})$ and let $\mathbf{V}$ be a random vector in $\mathbb{R}^n$, statistically independent of $\mathbf{U}$. The random vector $\mathbf{X} = [\mathbf{V} + \mathbf{U}] \bmod \Lambda$ is uniformly distributed over $\mathcal{V}$ and statistically independent of $\mathbf{V}$.*

Proof. For any $\mathbf{v} \in \mathbb{R}^n$ the set $\mathbf{v} + \mathcal{V}$ is a fundamental cell of Λ. Thus, by Proposition 18.1 we have that $[\mathbf{v} + \mathcal{V}] \bmod \Lambda = \mathcal{V}$ and $\text{Vol}(\mathbf{v} + \mathcal{V}) = \text{Vol}(\mathcal{V})$. Thus, for any $\mathbf{v} \in \mathbb{R}^n$

$$\mathbf{X} | \mathbf{V} = \mathbf{v} \sim [\mathbf{v} + \mathbf{U}] \bmod \Lambda \sim \text{Uniform}(\mathcal{V}).$$

□

The Crypto Lemma ensures that $\frac{1}{n}\mathbb{E}\|\mathbf{X}\|^2 = (1 - \epsilon)\mathsf{SNR}$, but our power constraint was $\|\mathbf{X}\|^2 \leq n\mathsf{SNR}$. Since $\mathbf{X}$ is uniformly distributed over $\mathcal{V}_c$ and Λ_c is good for MSE quantization, Theorem 18.1 implies that $\|\mathbf{X}\|^2 \leq n\mathsf{SNR}$ with high probability. Thus, whenever the power constraint is violated we can just transmit $\mathbf{0}$ instead of $\mathbf{X}$, and this will have a negligible effect on the error probability of the scheme.

The receiver scales its observation by a factor $\alpha > 0$ to be specified later, subtracts the dither $\mathbf{U}$ and reduces the result modulo the coarse lattice

$$\begin{aligned}\mathbf{Y}_{\text{eff}} &= [\alpha\mathbf{Y} - \mathbf{U}] \bmod \Lambda_c \\ &= [\mathbf{X} - \mathbf{U} + (\alpha - 1)\mathbf{X} + \alpha\mathbf{Z}] \bmod \Lambda_c \\ &= [\mathbf{t} + (\alpha - 1)\mathbf{X} + \alpha\mathbf{Z}] \bmod \Lambda_c \qquad (18.11) \\ &= [\mathbf{t} + \mathbf{Z}_{\text{eff}}] \bmod \Lambda_c, \qquad (18.12)\end{aligned}$$

where we have used the modulo distributive law in (18.11), and

$$\mathbf{Z}_{\text{eff}} = (\alpha - 1)\mathbf{X} + \alpha\mathbf{Z} \qquad (18.13)$$

is effective noise, that is statistically independent of $\mathbf{t}$, with effective variance

$$\sigma^2_{\text{eff}}(\alpha) \triangleq \frac{1}{n}\mathbb{E}\|\mathbf{Z}_{\text{eff}}\|^2 < \alpha^2 + (1 - \alpha)^2\mathsf{SNR}. \qquad (18.14)$$

Since $\mathbf{Z}$ is semi norm-ergodic, and $\mathbf{X}$ is uniformly distributed over the Voronoi region of a lattice that is good for MSE quantization, Theorem 18.1 implies that $\mathbf{Z}_{\text{eff}}$ is semi norm-ergodic with effective variance $\sigma^2_{\text{eff}}(\alpha)$. Setting $\alpha = \mathsf{SNR}/(1+\mathsf{SNR})$, such as to minimize the upper bound on $\sigma^2_{\text{eff}}(\alpha)$ results in effective variance $\sigma^2_{\text{eff}} < \mathsf{SNR}/(1 + \mathsf{SNR})$.

The receiver next computes

$$\begin{aligned}\hat{\mathbf{t}} &= [Q_{\Lambda_f}(\mathbf{Y}_{\text{eff}})] \bmod \Lambda_c \\ &= \left[Q_{\Lambda_f}(\mathbf{t} + \mathbf{Z}_{\text{eff}})\right] \bmod \Lambda_c, \qquad (18.15)\end{aligned}$$

and outputs the message corresponding to $\hat{\mathbf{t}}$ as its estimate. Since Λ_f is good for coding, $\mathbf{Z}_{\text{eff}}$ is semi norm-ergodic, and

$$\frac{r^2_{\text{eff}}(\Lambda_f)}{n} = (1 + \epsilon)\frac{\mathsf{SNR}}{1 + \mathsf{SNR}} > \sigma^2_{\text{eff}},$$

we have that $\Pr(\hat{\mathbf{t}} \neq \mathbf{t}) \to 0$ as the lattice dimension tends to infinity. Thus, we have proved the following.

Theorem 18.3. *There exist a coding scheme based on a nested lattice pair, that reliably achieves any rate below* $\frac{1}{2}\log(1 + \mathsf{SNR})$ *with lattice decoding for all additive semi norm-ergodic channels. In particular, if the additive noise is AWGN, this scheme is capacity achieving.*

Remark 18.3. In the Mod-Λ scheme the error probability does not depend on the chosen message, such that $P_{e,\max} = P_{e,\text{avg}}$. However, this required common randomness in the form of the dither $\mathbf{U}$. By a standard averaging argument it follows that there exist some fixed shift $\mathbf{u}$ that achieves the same, or better, $P_{e,\text{avg}}$. However, for a fixed shift the error probability is no longer independent of the chosen message.

18.4 Dirty Paper Coding

Assume now that the channel is

$$\mathbf{Y} = \mathbf{X} + \mathbf{S} + \mathbf{Z},$$

where $\mathbf{Z}$ is a unit variance semi norm-ergodic noise, $\mathbf{X}$ is subject to the same power constraint $\|\mathbf{X}\|^2 \leq n\mathsf{SNR}$ as before, and $\mathbf{S}$ is some arbitrary interference vector, known to the transmitter but *not* to the receiver.

Naively, one can think that the encoder can handle the interference $\mathbf{S}$ just by subtracting it from the transmitted codeword. However, if the codebook is designed to exactly meet the power constraint, after subtracting $\mathbf{S}$ the power constraint will be violated. Moreover, if $\|\mathbf{S}\|^2 > n\mathsf{SNR}$, this approach is just not feasible.

Using the Mod-Λ scheme, $\mathbf{S}$ can be cancelled out with no cost in performance. Specifically, instead of transmitting $\mathbf{X} = [\mathbf{t} + \mathbf{U}] \bmod \Lambda_c$, the transmitted signal in the presence of known interference will be

$$\mathbf{X} = [\mathbf{t} + \mathbf{U} - \alpha\mathbf{S}] \bmod \Lambda_c.$$

Clearly, the power constraint is not violated as $\mathbf{X} \sim \text{Uniform}(\mathcal{V}_c)$ due to the Crypto Lemma (now, $\mathbf{U}$ should also be independent of $\mathbf{S}$). The decoder is exactly the same as in the Mod-Λ scheme with no interference. It is easy to verify that the interference is completely cancelled out, and any rate below $\frac{1}{2}\log(1+\mathsf{SNR})$ can still be achieved.

Remark 18.4. When $\mathbf{Z}$ is Gaussian and $\mathbf{S}$ is Gaussian there is a scheme based on random codes that can reliably achieve $\frac{1}{2}\log(1+\mathsf{SNR})$. For arbitrary $\mathbf{S}$, to date, only lattice based coding schemes are known to achieve the interference free capacity. There are many more scenarios where lattice codes can reliably achieve better rates than the best known random coding schemes.

18.5 Construction of Good Nested Lattice Pairs

We now briefly describe a method for constructing nested lattice pairs. Our construction is based on starting with a linear code over a prime finite field, and embedding it periodically in $\mathbb{R}^n$ to form a lattice.

Definition 18.6 (p-ary Construction A)**.** Let p be a prime number, and let $\mathbf{F} \in \mathbb{Z}_p^{k\times n}$ be a $k \times n$ matrix whose entries are all members of the finite field $\mathbb{Z}_p$. The matrix $\mathbf{F}$ generates a linear p-ary code

$$\mathcal{C}(\mathbf{F}) \triangleq \left\{\mathbf{x} \in \mathbb{Z}_p^n \ : \ \mathbf{x} = [\mathbf{w}^T\mathbf{F}] \bmod p \quad \mathbf{w} \in \mathbb{Z}_p^k\right\}.$$

The p-ary Construction A lattice induced by the matrix $\mathbf{F}$ is defined as

$$\Lambda(\mathbf{F}) \triangleq p^{-1}\mathcal{C}(\mathbf{F}) + \mathbb{Z}^n.$$

Note that any point in $\Lambda(\mathbf{F})$ can be decomposed as $\mathbf{x} = p^{-1}\mathbf{c} + \mathbf{a}$ for some $\mathbf{c} \in \mathcal{C}(\mathbf{F})$ (where we identify the elements of $\mathbb{Z}_p$ with the integers $[0,1,\ldots,p-1]$) and $\mathbf{a} \in \mathbb{Z}^n$. Thus, for any $\mathbf{x}_1, \mathbf{x}_2 \in \Lambda(\mathbf{F})$ we have

$$\begin{aligned}
\mathbf{x}_1 + \mathbf{x}_2 &= p^{-1}(\mathbf{c}_1 + \mathbf{c}_2) + \mathbf{a}_1 + \mathbf{a}_2 \\
&= p^{-1}([\mathbf{c}_1 + \mathbf{c}_2] \bmod p + p\mathbf{a}) + \mathbf{a}_1 + \mathbf{a}_2 \\
&= p^{-1}\tilde{\mathbf{c}} + \tilde{\mathbf{a}} \\
&\in \Lambda(\mathbf{F})
\end{aligned}$$

where $\tilde{\mathbf{c}} = [\mathbf{c}_1 + \mathbf{c}_2] \bmod p \in \mathcal{C}(\mathbf{F})$ due to the linearity of $\mathcal{C}(\mathbf{F})$, and $\mathbf{a}$ and $\tilde{\mathbf{a}}$ are some vectors in $\mathbb{Z}^n$. It can be verified similarly that for any $\mathbf{x} \in \Lambda(\mathbf{F})$ it holds that $-\mathbf{x} \in \Lambda(\mathbf{F})$, and that if all codewords in $\mathcal{C}(\mathbf{F})$ are distinct, then $\Lambda(\mathbf{F})$ has a finite minimum distance. Thus, $\Lambda(\mathbf{F})$ is indeed a lattice. Moreover, if $\mathbf{F}$ is full-rank over $\mathbb{Z}_p$, then the number of distinct codewords in $\mathcal{C}(\mathbf{F})$ is p^k. Consequently, the number of lattice points in every integer shift of the unit cube is p^k, so the corresponding Voronoi region must satisfy $\mathrm{Vol}(\mathcal{V}) = p^{-k}$.

Similarly, we can construct a nested lattice pair from a linear code. Let $0 \le k' < k$ and let $\mathbf{F}'$ be the sub-matrix obtained by taking only the first k' rows of $\mathbf{F}$. The matrix $\mathbf{F}'$ generates a linear code $\mathcal{C}'(\mathbf{F}')$ that is nested in $\mathcal{C}(\mathbf{F})$, i.e., $\mathcal{C}'(\mathbf{F}') \subset \mathcal{C}(\mathbf{F})$. Consequently we have that $\Lambda(\mathbf{F}') \subset \Lambda(\mathbf{F})$, and the nesting ratio is

$$\Gamma(\Lambda(\mathbf{F}), \Lambda(\mathbf{F}')) = p^{\frac{k-k'}{n}}.$$

An advantage of this nested lattice construction for Voronoi constellations is that there is a very simple mapping between messages and codewords in $\mathcal{L} = \Lambda_f \cap \mathcal{V}_c$. Namely, we can index our set of $2^{nR} = p^{k-k'}$ messages by all vectors in $\mathbb{Z}_p^{k-k'}$. Then, for each message vector $\mathbf{w} \in \mathbb{Z}_p^{k-k'}$, the corresponding codeword in $\mathcal{L} = \Lambda(\mathbf{F}) \cap \mathcal{V}(\Lambda(\mathbf{F}'))$ is obtained by constructing the vector

$$\tilde{\mathbf{w}}^T = [\underbrace{0 \cdots 0}_{k' \text{ zeros}} \, \mathbf{w}^T] \in \mathbb{Z}_p^k, \tag{18.16}$$

and taking $\mathbf{t} = \mathbf{t}(\mathbf{w}) = \left[[\tilde{\mathbf{w}}^T\mathbf{F}] \bmod p\right] \bmod \Lambda(\mathbf{F}')$. Also, in order to specify the codebook $\mathcal{L}$, only the (finite field) generating matrix $\mathbf{F}$ is needed.

If we take the elements of $\mathbf{F}$ to be i.i.d. and uniform over $\mathbb{Z}_p$, we get a random ensemble of nested lattice codes. It can be shown that if p grows fast enough with the dimension n (taking $p = O(n^{(1+\epsilon)/2})$ suffices) almost all pairs in the ensemble have the property that both the fine and coarse lattice are good for both coding and for MSE quantization [OE15].

Disclaimer: *This text is a very brief and non-exhaustive survey of the applications of lattices in information theory. For a comprehensive treatment, see [Zam14].*

§ 19. Channel coding: energy-per-bit, continuous-time channels

19.1 Energy per bit

Consider the additive Gaussian noise channel:

$$Y_i = X_i + Z_i, \quad Z_i \sim \mathcal{N}(0, \frac{N_0}{2}). \tag{19.1}$$

In the last lecture, we analyzed the maximum number of information bits $(M^*(n, \epsilon, P))$ that can be pumped through for given n time use of the channel under the energy constraint P. Today we shall study the counterpart of it: without any time constraint, in order to send k information bits, what is the minimum energy needed? $(E^*(k, \epsilon))$

Definition 19.1 ($(E, 2^k, \epsilon)$ code)**.** For a channel $W \to X^\infty \to Y^\infty \to \hat{W}$, where $Y^\infty = X^\infty + Z^\infty$, a $(E, 2^k, \epsilon)$ code is a pair of encoder-decoder:

$$\begin{aligned} f : [2^k] \to \mathbb{R}^\infty, &\quad g : \mathbb{R}^\infty \to [2^k], \\ \text{such that 1).} &\quad \forall m, \|f(m)\|_2^2 \le E, \\ \text{2).} &\quad P[g(f(W) + Z^\infty) \ne W] \le \epsilon. \end{aligned}$$

Definition 19.2 (Fundamental limit)**.**

$$E^*(k, \epsilon) = \min\{E : \exists (E, 2^k, \epsilon) \text{ code}\}$$

Note: Operational meaning of $\lim_{\epsilon \to 0} E^*(k, \epsilon)$: it suggests the smallest battery one needs in order to send k bits without any time constraints, below that level reliable communication is impossible.

Theorem 19.1 ($(E_b/N_0)_{min} = -1.6dB$)**.**

$$\lim_{\epsilon \to 0} \limsup_{k \to \infty} \frac{E^*(k, \epsilon)}{k} = \frac{N_0}{\log_2 e}, \quad \frac{1}{\log_2 e} = -1.6dB \tag{19.2}$$

Proof.

1. ("≥" converse part)

$$\begin{aligned}
-h(\epsilon) + \bar{\epsilon}k &\le d((1-\epsilon)\|\frac{1}{M}) \qquad \text{(Fano)} \\
&\le I(W;\hat{W}) \qquad \text{(data processing for divergence)} \\
&\le I(X^\infty;Y^\infty) \qquad \text{(data processing for M.I.)} \\
&\le \sum_{i=1}^{\infty} I(X_i;Y_i) \qquad (\lim_{n\to\infty} I(X^n;U) = I(X^\infty;U)) \\
&\le \sum_{i=1}^{\infty} \frac{1}{2}\log(1+\frac{\mathbb{E}X_i^2}{N_0/2}) \qquad \text{(Gaussian)} \\
&\le \frac{\log e}{2}\sum_{i=1}^{\infty} \frac{\mathbb{E}X_i^2}{N_0/2} \qquad \text{(linearization)} \\
&\le \frac{E}{N_0}\log e \\
\Rightarrow \frac{E^*(k,\epsilon)}{k} &\ge \frac{N_0}{\log e}(\bar{\epsilon} - \frac{h(\epsilon)}{k}).
\end{aligned}$$

2. ("≤" achievability part)
   Notice that a $(n, 2^k, \epsilon, P)$ code for AWGN channel is also a $(nP, 2^k, \epsilon)$ code for the energy problem without time constraint. Therefore,

$$\log_2 M^*_{max}(n,\epsilon,P) \ge k \Rightarrow E^*(k,\epsilon) \le nP.$$

   $\forall P$, take $k_n = \lfloor \log M^*_{max}(n,\epsilon,P) \rfloor$, we have $\frac{E^*(k_n,\epsilon)}{k_n} \le \frac{nP}{k_n}$, $\forall n$, and take the limit:

$$\begin{aligned}
\limsup_{n\to\infty} \frac{E^*(k_n,\epsilon)}{k_n} &\le \limsup_{n\to\infty} \frac{nP}{\log M^*_{max}(n,\epsilon,P)} \\
&= \frac{P}{\liminf_{n\to\infty} \frac{1}{n}\log M^*_{max}(n,\epsilon,P)} \\
&= \frac{P}{\frac{1}{2}\log(1+\frac{P}{N_0/2})}
\end{aligned}$$

   Choose P for the lowest upper bound:

$$\begin{aligned}
\limsup_{n\to\infty} \frac{E^*(k_n,\epsilon)}{k_n} &\le \inf_{P\ge 0} \frac{P}{\frac{1}{2}\log(1+\frac{P}{N_0/2})} \\
&= \lim_{P\to 0} \frac{P}{\frac{1}{2}\log(1+\frac{P}{N_0/2})} \\
&= \frac{N_0}{\log_2 e}
\end{aligned}$$

□

Note: [Remark] In order to send information reliably at $E_b/N_0 = -1.6dB$, infinitely many time slots are needed, and the information rate (spectral efficiency) is thus 0. In order to have non-zero spectral efficiency, one necessarily has to step back from -1.6 dB.

Note: [PPM code] The following code, pulse-position modulation (PPM), is very efficient in terms of E_b/N_0.

$$\text{PPM encoder: } \forall m, f(m) = (0, 0, \dots, \underbrace{\sqrt{E}}_{m\text{-th location}}, \dots) \tag{19.3}$$

It is not hard to derive an upper bound on the probability of error that this code achieves [PPV11, Theorem 2]:

$$\epsilon \le \mathbb{E}\left[\min\left\{MQ\left(\sqrt{\frac{2E}{N_0}} + Z\right), 1\right\}\right], \qquad Z \sim \mathcal{N}(0,1)\,.$$

In fact, the code can be further slightly optimized by subtracting the common center of gravity $(2^{-k}\sqrt{E}, \dots, 2^{-k}\sqrt{E} \dots)$ and rescaling each codeword to satisfy the power constraint. The resulting constellation (simplex code) is conjectured to be non-asymptotic optimum in terms of E_b/N_0 for small ϵ ("simplex conjecture").

19.2 What is N_0?

In the above discussion, we have assumed $Z_i \sim \mathcal{N}(0, N_0/2)$, but how do we determine N_0?

In reality the signals are continuous time (CT) process, the continuous time AWGN channel for the RF signals is modeled as:

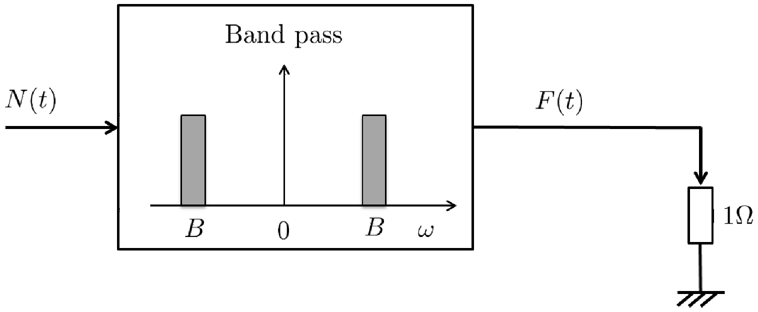
$$Y(t) = X(t) + N(t) \tag{19.4}$$

where noise $N(t)$ (added at the receiver antenna) is a real stationary ergodic process and is assumed to be "white Gaussian noise" with single-sided PSD N_0. Figure 19.1 at the end illustrates the communication architecture. In the following discussion, we shall find the equivalent discrete time (DT) AWGN model for the continuous time (CT) AWGN model in (19.4), and identify the relationship between N_0 in the DT model and $N(t)$ in the CT model.

- Goal: communication in $f_c \pm B/2$ band.
  (the (possibly complex) baseband signal lies in $[-W, +W]$, where $W = B/2$)
- observations:
  1. Any signal band limited to $f_c \pm B/2$ can be produced by this architecture
  2. At the step of C/D conversion, the LPF followed by sampling at B samples/sec is sufficient statistics for estimating $X(t), X_B(t)$, as well as $\{X_i\}$.

First of all, what is $N(t)$ in (19.4)?

Engineers' definition of $N(t)$



Testing whether a process $N(t)$ is "white noise"

Estimate the average power dissipation at the resistor:

$$\lim_{T\to\infty}\frac{1}{T}\int_{t=0}^{T}F_t^2dt \overset{\text{ergodic}}{=} \mathbb{E}[F^2] \overset{(*)}{=} N_0B$$

If for some constant N_0, (*) holds for any narrow band with center frequency f_c and bandwidth B, then $N(t)$ is called a "white noise" with one-sided PSD N_0.

Typically, white noise comes from thermal noise at the receiver antenna. Thus:

$$N_0 \approx k\mathbf{T} \tag{19.5}$$

where $k = 1.38 \times 10^{-23}$ is the Boltzmann constant, and $\mathbf{T}$ is the absolute temperature. The unit of N_0 is $(Watt/Hz = J)$.

An intuitive explanation to (19.5) is as follows: the thermal energy carried by each microscopic degree of freedom (dof) is approximately $\frac{k\mathbf{T}}{2}$; for bandwidth B and duration T, there are in total $2BT$ dof; by "white noise" definition we have the total energy of the noise to be:

$$N_0BT = \frac{k\mathbf{T}}{2}2BT \;\Rightarrow\; N_0 = k\mathbf{T}.$$

Mathematicians' definition of $N(t)$

Denote the set of all real finite energy signals $f(t)$ by $\mathcal{L}_2(\mathbb{R})$, it is a vector space with the inner product of two signals $f(t), g(t)$ defined by

$$< f, g >= \int_{t=-\infty}^{\infty} f(t)g(t)dt.$$

Definition 19.3 (White noise)**.** $N(t)$ is a white noise with two-sided PSD being constant $N_0/2$ if $\forall f, g \in \mathcal{L}_2(\mathbb{R})$ such that $\int_{-\infty}^{\infty} f^2(t)dt = \int_{-\infty}^{\infty} g^2(t)dt = 1$, we have that

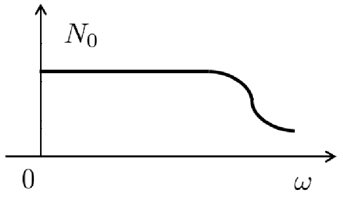
1. $$< f, N >\triangleq \int_{-\infty}^{\infty} f(t)N(t)dt \sim \mathcal{N}(0, \frac{N_0}{2}). \tag{19.6}$$

2. The joint distribution of $(< f, N >, < g, N >)$ is jointly Gaussian with covariance equal to inner product $< f, g >$.

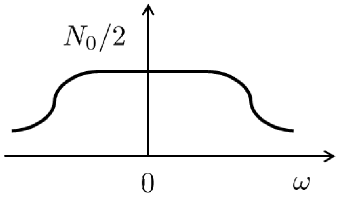
Note: By this definition, $N(t)$ is not a stochastic process, rather it is a collection of linear mappings that map any $f \in \mathcal{L}_2(\mathbb{R})$ to a Gaussian random variable.

Note: Informally, we write:

$N(t)$ is white noise with one-sided PSD N_0(*or* two-sided PSD $N_0/2$) $\Longleftrightarrow \mathbb{E}[N(t)N(s)] = \dfrac{N_0}{2}\delta(t-s)$ (19.7)



Engineers' white noise



Mathematicians' white noise

Note: The concept of one-sided PSD arises when $N(t)$ is necessarily real, since in that case power spectrum density is symmetric around 0, and thus to get the noise power in band $[a,b]$ one can get

$$\text{noise power} = \int_a^b F_{\text{one-sided}}(f)df = \int_a^b + \int_{-b}^{-a} F_{\text{two-sided}}(f)df\,,$$

where $F_{\text{one-sided}}(f) = 2F_{\text{two-sided}}(f)$. In theory of stochastic processes it is uncommon to talk about one-sided PSD, but in engineering it is.

Verify the equivalence between CT /DT models

First, consider the relation between RF signals and baseband signals.

$$X(t) = Re(X_B(t)\sqrt{2}e^{j\omega_c t}),$$
$$Y_B(t) = \sqrt{2}LPF_2(Y(t)e^{j\omega_c t}),$$

where $\omega_c = 2\pi f_c$. The LPF_2 with high cutoff frequency $\sim \frac{3}{4}f_c$ serves to kill the high frequency component after demodulation, and the amplifier of magnitude $\sqrt{2}$ serves to preserve the total energy of the signal, so that in the absence of noise we have that $Y_B(t) = X_B(t)$. Therefore,

$$Y_B(t) = X_B(t) + \widetilde{N}(t) \sim \mathbb{C}$$

where $\widetilde{N}(t)$ is a complex Gaussian white noise and

$$\mathbb{E}[\widetilde{N}(t)\widetilde{N}(s)^*] = N_0\delta(t-s).$$

Notice that after demodulation, the PSD of the noise is $N_0/2$ with $N_0/4$ in the real part and $N_0/4$ in the imaginary part, and after the $\sqrt{2}$ amplifier the PSD of the noise is restored to $N_0/2$ in both real and imaginary part.

Next, consider the equivalent discrete time signals.

$$X_B(t) = \sum_{i=-\infty}^{\infty} X_i sinc_B(t - \frac{i}{B})$$
$$Y_i = \int_{t=-\infty}^{\infty} Y_B(t) sinc_B(t - \frac{i}{B})dt$$
$$Y_i = X_i + Z_i$$

where the additive noise Z_i is given by:

$$Z_i = \int_{t=-\infty}^{\infty} \widetilde{N}(t) sinc_B(t - \frac{i}{B}) dt \sim i.i.d\ \mathbb{C}\mathcal{N}(0, N_0). \qquad \text{(by (19.6))}$$

if we focus on the real part of all signals, it is consistent with the real AWGN channel model in (19.1).

Finally, the energy of the signal is preserved:

$$\sum_{i=-\infty}^{\infty} |X_i|^2 = \|X_B(t)\|_2^2 = \|X(t)\|_2^2.$$

Note: [Punchline]

$$\text{CT AWGN (band limited)} \Longleftrightarrow \text{DT } \mathbb{C}\text{-AWGN}$$

$$\text{two-sided PSD } \frac{N_0}{2} \Longleftrightarrow Z_i \sim \mathbb{C}\mathcal{N}(0, N_0)$$

$$\text{energy=} \int X(t)^2 dt \Longleftrightarrow \text{energy=} \sum |X_i|^2$$

19.3 Capacity of the continuous-time band-limited AWGN channel

Theorem 19.2. *Let $M^*_{CT}(T, \epsilon, P)$ the maximum number of waveforms that can be sent through the channel*

$$Y(t) = X(t) + N(t), \qquad \mathbb{E} N(t)N(s) = \frac{N_0}{2}\delta(t - s)$$

such that:

1. *in the duration $[0, T]$;*
2. *band limited to $[f_c - \frac{B}{2}, f_c + \frac{B}{2}]$ for some large carrier frequency*
3. *input energy constrained to $\int_{t=0}^{T} x^2(t) \le TP$;*
4. *error probability $P[\hat{W} \ne W] \le \epsilon$.*

Then

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} \frac{1}{T} \log M^*_{CT}(T, \epsilon, P) = B \log(1 + \frac{P}{N_0 B}), \tag{19.8}$$

Proof. Consider the DT equivalent $\mathbb{C}$-AWGN channel of this CT model, we have that

$$\frac{1}{T} \log M^*_{CT}(T, \epsilon, P) = \frac{1}{T} \log M^*_{\mathbb{C}\text{-AWGN}}(BT, \epsilon, P/B)$$

This is because:

- in time T we get to choose BT complex samples
- The power constraint in the DT model changed because for blocklength BT we have

  $$\sum_{i=1}^{BT} |X_i|^2 = \|X(t)\|_2^2 \le PT,$$

  thus per-letter power constraint is $\frac{P}{B}$.

Calculate the rate of the equivalent DT AWGN channel and we are done. □

Note the above "theorem" is not rigorous, since conditions 1 and 2 are mutually exclusive: any time limited non-trivial signal cannot be band limited. Rigorously, one should relax 2 by constraining the signal to have a vanishing out-of-band energy as $T \to \infty$. Rigorous approach to this question lead to the theory of prolate spheroidal functions.

19.4 Capacity of the continuous-time band-unlimited AWGN channel

In the limit of large bandwidth B the capacity formula (19.8) yields

$$C_{B=\infty}(P) = \lim_{B\to\infty} B\log(1+\frac{P}{N_0 B}) = \frac{P}{N_0}\log e\,.$$

It turns out that this result is easy to prove rigorously.

Theorem 19.3. *Let $M^*(T,\epsilon,P)$ the maximum number of waveforms that can be sent through the channel*

$$Y(t) = X(t) + N(t)\,, \qquad \mathbb{E}\, N(t)N(s) = \frac{N_0}{2}\delta(t-s)$$

such that each waveform $x(t)$

1. *is non-zero only on $[0,T]$;*
2. *input energy constrained to $\int_{t=0}^{T} x^2(t) \le TP$;*
3. *error probability $P[\hat{W} \ne W] \le \epsilon$.*

Then

$$\lim_{\epsilon\to 0}\liminf_{T\to\infty} \frac{1}{T}\log M^*(T,\epsilon,P) = \frac{P}{N_0}\log e \tag{19.9}$$

Proof. Note that the space of all square-integrable functions on $[0,T]$, denoted $L_2[0,T]$ has countable basis (e.g. sinusoids). Thus, by changing to that basis we may assume that equivalent channel model

$$\tilde{Y}_j = \tilde{X}_j + \tilde{Z}_j\,, \qquad \tilde{Z}_j \sim \mathcal{N}(0,\frac{N_0}{2})\,,$$

and energy constraint (dependent upon duration T):

$$\sum_{j=1}^{\infty} \tilde{X}_j^2 \le PT\,.$$

But then the problem is equivalent to energy-per-bit one and hence

$$\log_2 M^*(T,\epsilon,P) = k \iff E^*(k,\epsilon) = PT\,.$$

Thus,

$$\lim_{\epsilon\to 0}\liminf_{n\to\infty} \frac{1}{T}\log_2 M^*(T,\epsilon,P) = \frac{P}{\lim_{\epsilon\to 0}\limsup_{k\to\infty}\frac{E^*(k,\epsilon)}{k}} = \frac{P}{N_0}\log_2 e\,,$$

where the last step is by Theorem 19.1. □

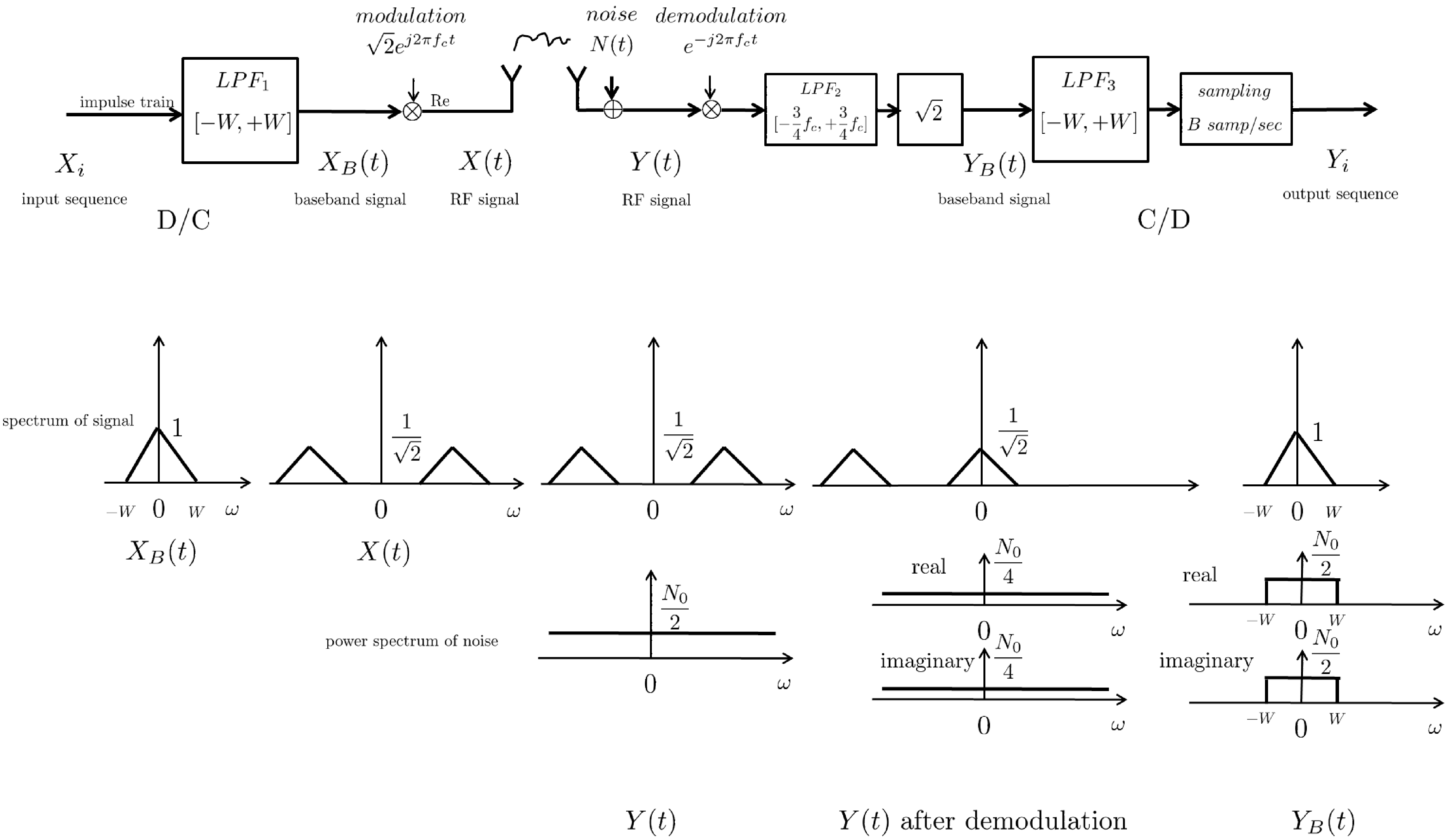


Figure 19.1: DT / CT AWGN model

19.5 Capacity per unit cost

Generalizing the energy-per-bit setting of Theorem 19.1 we get the problem of *capacity per unit cost*:

1. Given a random transformation $P_{Y^\infty|X^\infty}$ and cost function $\mathsf{c} : \mathcal{X} \to \mathbb{R}_+$, we let

$$M^*(E,\epsilon) = \max\{M : (E,M,\epsilon)\text{-code}\}\,,$$

   where (E,M,ϵ)-code is defined as a map $[M] \to \mathcal{X}^\infty$ with every codeword x^∞ satisfying

$$\sum_{t=1}^{\infty} \mathsf{c}(x_t) \le E\,. \tag{19.10}$$

2. Capacity per unit cost is defined as

$$C_{puc} \triangleq \lim_{\epsilon\to 0} \liminf_{E\to\infty} \frac{1}{E} \log M^*(E,\epsilon)\,.$$

3. Let $C(P)$ be the capacity-cost function of the channel (in the usual sense of capacity, as defined in (17.1). Assuming $P_0 = 0$ and $C(0) = 0$ it is not hard to show that:

$$C_{puc} = \sup_P \frac{C(P)}{P} = \lim_{P\to 0} \frac{C(P)}{P} = \left.\frac{d}{dP}\right|_{P=0} C(P)\,.$$

4. The surprising discovery of Verdú is that one can avoid computing $C(P)$ and derive the C_{puc} directly. This is a significant help, as for many practical channels $C(P)$ is unknown. Additionally, this gives a yet another fundamental meaning to KL-divergence.

Theorem 19.4. *For a stationary memoryless channel $P_{Y^\infty|X^\infty} = \prod P_{Y|X}$ with $P_0 = \mathsf{c}(x_0) = 0$ (i.e. there is a symbol of zero cost), we have*

$$C_{puc} = \sup_{x\ne x_0} \frac{D(P_{Y|X=x} \| P_{Y|X=x_0})}{\mathsf{c}(x)}\,.$$

In particular, $C_{puc} = \infty$ if there exists $x_1 \ne x_0$ with $\mathsf{c}(x_1) = 0$.

Proof. Let

$$C_V = \sup_{x\ne x_0} \frac{D(P_{Y|X=x} \| P_{Y|X=x_0})}{\mathsf{c}(x)}\,.$$

Converse: Consider a (E,M,ϵ) code $W \to X^\infty \to Y^\infty \to \hat{W}$. Introduce an auxiliary distribution $Q_{W,X^\infty,Y^\infty,\hat{W}}$, where a channel is a useless one

$$Q_{Y^\infty|X^\infty} = Q_{Y^\infty} \triangleq P^\infty_{Y|X=x_0}\,.$$

That is, the overall factorization is

$$Q_{W,X^\infty,Y^\infty,\hat{W}} = P_W P_{X^\infty|W} Q_{Y^\infty} P_{\hat{W}|Y^\infty}\,.$$

Then, as usual we have from the data-processing for divergence

$$
\begin{aligned}
(1-\epsilon)\log M + h(\epsilon) &\le d(1-\epsilon\|\frac{1}{M}) && (19.11)\\
&\le D(P_{W,X^\infty,Y^\infty,\hat{W}}\|Q_{W,X^\infty,Y^\infty,\hat{W}}) && (19.12)\\
&= D(P_{Y^\infty|X^\infty}\|Q_{Y^\infty}|P_{X^\infty}) && (19.13)\\
&= \mathbb{E}\left[\sum_{t=1}^{\infty} d(X_t)\right], && (19.14)
\end{aligned}
$$

where we denoted for convenience

$$d(x) \triangleq D(P_{Y|X=x}\|P_{Y|X=x_0}).$$

By the definition of C_V we have

$$d(x) \le \mathsf{c}(x) C_V .$$

Thus, continuing (19.14) we obtain

$$(1-\epsilon)\log M + h(\epsilon) \le C_V\,\mathbb{E}\left[\sum_{t=1}^{\infty} \mathsf{c}(X_t)\right] \le C_V\cdot E,$$

where the last step is by the cost constraint (19.10). Thus, dividing by E and taking limits we get

$$C_{puc} \le C_V .$$

Achievability: We generalize the PPM code (19.3). For each $x_1 \in \mathcal{X}$ and $n \in \mathbb{Z}_+$ we define the encoder f as follows:

$$
\begin{aligned}
f(1) &= (\underbrace{x_1, x_1, \dots, x_1}_{n\text{-times}}, \underbrace{x_0, \dots, x_0}_{n(M-1)\text{-times}}) && (19.15)\\
f(2) &= (\underbrace{x_0, x_0, \dots, x_0}_{n\text{-times}}, \underbrace{x_1, \dots, x_1}_{n\text{-times}}, \underbrace{x_0, \dots, x_0}_{n(M-2)\text{-times}}) && (19.16)\\
&\cdots && (19.17)\\
f(M) &= (\underbrace{x_0, \dots, x_0}_{n(M-1)\text{-times}}, \underbrace{x_1, x_1, \dots, x_1}_{n\text{-times}}) && (19.18)
\end{aligned}
$$

Now, by Stein's lemma there exists a subset $S \subset \mathcal{Y}^n$ with the property that

$$
\begin{aligned}
&\mathbb{P}[Y^n \in S | X^n = (x_1, \dots, x_1)] \ge 1-\epsilon_1 && (19.19)\\
&\mathbb{P}[Y^n \in S | X^n = (x_0, \dots, x_0)] \le \exp\{-nD(P_{Y|X=x_1}\|P_{Y|X=x_0}) + o(n)\}. && (19.20)
\end{aligned}
$$

Therefore, we propose the following (suboptimal!) decoder:

$$
\begin{aligned}
Y^n \in S &\implies \hat{W} = 1 && (19.21)\\
Y_{n+1}^{2n} \in S &\implies \hat{W} = 2 && (19.22)\\
&\cdots && (19.23)
\end{aligned}
$$

From the union bound we find that the overall probability of error is bounded by

$$\epsilon \le \epsilon_1 + M\exp\{-nD(P_{Y|X=x_1}\|P_{Y|X=x_0}) + o(n)\}.$$

At the same time the total cost of each codeword is given by $n\mathsf{c}(x_1)$. Thus, taking $n \to \infty$ and after straightforward manipulations, we conclude that

$$C_{puc} \ge \frac{D(P_{Y|X=x_1} \| P_{Y|X=x_0})}{\mathsf{c}(x_1)}\,.$$

This holds for any symbol $x_1 \in \mathcal{X}$, and so we are free to take supremum over x_1 to obtain $C_{puc} \ge C_V$, as required. □

19.5.1 Energy-per-bit for AWGN channel subject to fading

Consider a stationary memoryless Gaussian channel with fading H_j (unknown at the receiver). Namely,

$$Y_j = H_j X_j + Z_j, \qquad H_j \sim \mathcal{N}(0,1) \perp\!\!\!\perp Z_j \sim \mathcal{N}(0, \frac{N_0}{2})\,.$$

The cost function is the usual quadratic one $\mathsf{c}(x) = x^2$. As we discussed previously, cf. (17.8), the capacity-cost function $C(P)$ is unknown in closed form, but is known to behave drastically different from the case of non-fading AWGN (i.e. when $H_j = 1$). So here previous theorem comes handy, as we cannot just compute $C'(0)$. Let us perform a simple computation required, cf. (1.16):

$$C_{puc} = \sup_{x \ne 0} \frac{D(\mathcal{N}(0, x^2 + \frac{N_0}{2}) \| \mathcal{N}(0, \frac{N_0}{2}))}{x^2} \tag{19.24}$$

$$= \frac{1}{N_0} \sup_{x \ne 0} \left(\log e - \frac{\log(1 + \frac{2x^2}{N_0})}{\frac{2x^2}{N_0}} \right) \tag{19.25}$$

$$= \frac{\log e}{N_0} \tag{19.26}$$

Comparing with Theorem 19.1 we discover that surprisingly, the capacity-per-unit-cost is unaffected by the presence of fading. In other words, the random multiplicative noise which is so detrimental at high SNR, appears to be much more benign at low SNR (recall that $C_{puc} = C'(0)$). There is one important difference, however. It should be noted that the supremization over x in (19.25) is solved at $x = \infty$. Following the proof of the converse bound, we conclude that any code hoping to achieve optimal C_{puc} must satisfy a strange constraint:

$$\sum_t x_t^2 1\{|x_t| \ge A\} \approx \sum_t x_t^2 \qquad \forall A > 0$$

I.e. the total energy expended by each codeword must be almost entirely concentrated in very large spikes. Such a coding method is called “flash signalling”. Thus, we can see that unlike non-fading AWGN (for which due to rotational symmetry all codewords can be made “mellow”), the only hope of achieving full C_{puc} in the presence of fading is by signalling in huge bursts of energy.

This effect manifests itself in the speed of convergence to C_{puc} with increasing constellation sizes. Namely, the energy-per-bit $\frac{E^*(k,\epsilon)}{k}$ behaves as

$$\frac{E^*(k,\epsilon)}{k} = (-1.59\ dB) + \sqrt{\frac{\text{const}}{k}} Q^{-1}(\epsilon) \qquad \text{(AWGN)} \tag{19.27}$$

$$\frac{E^*(k,\epsilon)}{k} = (-1.59\ dB) + \sqrt[3]{\frac{\log k}{k}} \left(Q^{-1}(\epsilon)\right)^2 \qquad \text{(fading)} \tag{19.28}$$

Fig. 19.2 shows numerical details.

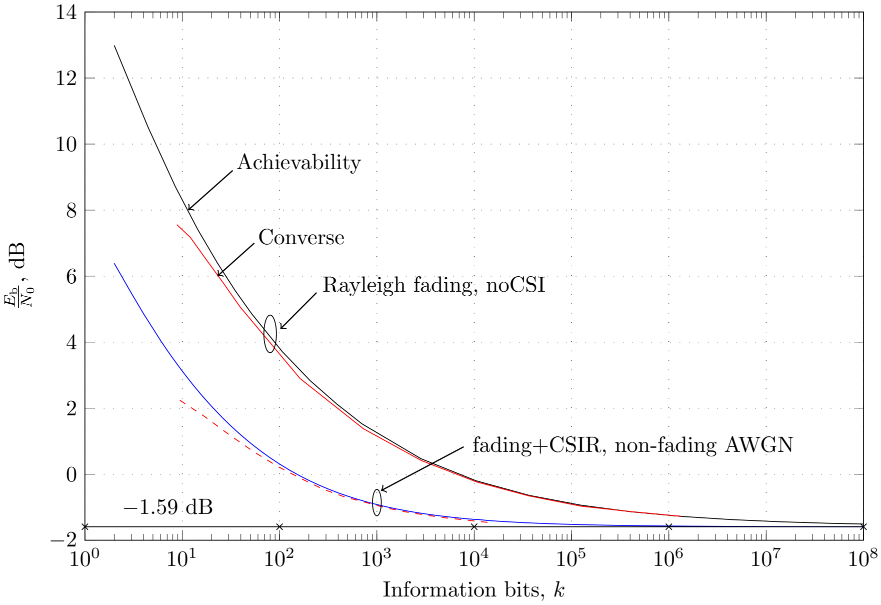


Figure 19.2: Comparing the energy-per-bit required to send a packet of k-bits for different channel models (curves represent upper and lower bounds on the unknown optimal value $\frac{E^*(k,\epsilon)}{k}$). As a comparison: to get to -1.5 dB one has to code over $6 \cdot 10^4$ data bits when the channel is non-fading AWGN or fading AWGN with H_j known perfectly at the receiver. For fading AWGN without knowledge of H_j (noCSI), one has to code over at least $7 \cdot 10^7$ data bits to get to the same -1.5 dB. Plot generated via [Spe15].

§ 20. Advanced channel coding. Source-Channel separation.

Topics: Strong Converse, Channel Dispersion, Joint Source Channel Coding (JSCC)

20.1 Strong Converse

We begin by stating the main theorem.

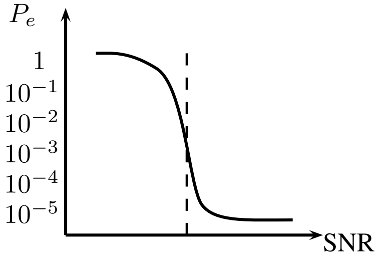
Theorem 20.1. *For any stationary memoryless channel with either $|\mathcal{A}| < \infty$ or $|\mathcal{B}| < \infty$ we have $C_\epsilon = C$ for $0 < \epsilon < 1$.*

Remark: In Theorem 16.4, we showed that $C \le C_\epsilon \le \frac{C}{1-\epsilon}$. Now we are asserting that equality holds for every ϵ. Our previous converse arguments showed that communication with an arbitrarily small error probability is possible only when using rate $R < C$; the strong converse shows that when you try to communicate with any rate above capacity $R > C$, then the probability of error will go to 1 (typically with exponential speed in n). In other words,

$$\epsilon^*(n, \exp(nR)) \to \begin{cases} 0 & R < C \\ 1 & R > C \end{cases}$$

where $\epsilon^*(n, M)$ is the inverse of $M^*(n, \epsilon)$ defined in (16.3).

In practice, engineers observe this effect in the form of *waterfall plots*, which depict the dependence of a given communication system (code+modulation) on the SNR.



Below a certain SNR, the probability of error shoots up to 1, so that the receiver will only see garbage.

Proof. We will give a sketch of the proof. Take an (n, M, ϵ)-code for channel $P_{Y|X}$. The main trick is to consider an auxiliary channel $Q_{Y|X}$ which is easier to analyze.

$$\begin{tikzcd} W \arrow[r] & X^n \arrow[r, "P_{Y^n|X^n}"] \arrow[r, bend right, "Q_{Y^n|X^n}"'] & Y^n \arrow[r] & \hat{W} \end{tikzcd}$$

Sketch 1: Here, we take $Q_{Y^n|X^n} = (P_Y^*)^n$, where P_Y^* is the capacity-achieving output distribution (caod) of the channel $P_{Y|X}$.[1] Note that for communication purposes, $Q_{Y^n|X^n}$ is a useless channel; it ignores the input and randomly picks a member of the output space according to $(P_Y^*)^n$, so that X^n and Y^n are decoupled (independent). Consider the probability of error under each channel:

$$\mathbb{Q}[\hat{W} = W] = \frac{1}{M} \quad \text{(Blindly guessing the sent codeword)}$$
$$\mathbb{P}[\hat{W} = W] = 1 - \epsilon$$

Since the random variable $\mathbf{1}_{\{\hat{W}=W\}}$ has a huge mass under $\mathbb{P}$ and small mass under $\mathbb{Q}$, this looks like a great binary hypothesis test to distinguish the two distributions, $P_{WX^nY^n\hat{W}}$ and $Q_{WX^nY^n\hat{W}}$. Since any hypothesis test can't beat the optimal Neyman-Pearson test, we get the upper bound

$$\beta_{1-\epsilon}(P_{WX^nY^n\hat{W}}, Q_{WX^nY^n\hat{W}}) \le \frac{1}{M} \tag{20.1}$$

(Recall that $\beta_\alpha(P,Q) = \inf_{P[E]\ge\alpha} Q[E]$). Since the likelihood ratio is a sufficient statistic for this hypothesis test, we can test only between

$$\frac{P_{WX^nY^n\hat{W}}}{Q_{WX^nY^n\hat{W}}} = \frac{P_W P_{X^n|W} P_{Y^n|X_n} P_{\hat{W}|Y^n}}{P_W P_{X^n|W} (P_Y^*)^n P_{\hat{W}|Y^n}} = \frac{P_{W|X^n} P_{X^nY^n} P_{\hat{W}|Y^n}}{P_{W|X^n} P_{X^n} (P_Y^*)^n P_{\hat{W}|Y^n}} = \frac{P_{X^nY^n}}{P_{X^n}(P_Y^*)^n}$$

Therefore, inequality above becomes

$$\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(P_Y^*)^n) \le \frac{1}{M} \tag{20.2}$$

Computing the LHS of this bound need not be easy, since generally we know $P_{Y|X}$ and P_Y^*, but can't assume anything about P_{X^n} which depends on the code. (Note that X^n is the output of the encoder and uniformly distributed on the codebook for deterministic encoders). Certain tricks are needed to remove the dependency on codebook. However, in case the channel is "symmetric" the dependence on the codebook disappears: this is shown in the following example for the BSC. To treat the general case one simply decomposes the channel into symmetric subchannels (for example, by considering constant composition subcodes).

Example. For a BSC$(\delta)^n$, recall that

$$P_{Y^n|X^n}(y^n|x^n) = P_Z^n(y^n - x^n), \quad Z^n \sim \text{Bern}(\delta)^n$$
$$(P_Y^*)^n(y^n) = 2^{-n}$$

From the Neyman Pearson test, the optimal HT takes the form

$$\beta_\alpha(\underbrace{P_{X^nY^n}}_{\mathbb{P}}, \underbrace{P_{X^n}(P_Y^*)^n}_{\mathbb{Q}}) = \mathbb{Q}\left[\log\frac{P_{X^nY^n}}{P_{X^n}(P_Y^*)^n} \ge \gamma\right] \;\text{ where }\; \alpha = \mathbb{P}\left[\log\frac{P_{X^nY^n}}{P_{X^n}(P_Y^*)^n} \ge \gamma\right]$$

For the BSC, this becomes

$$\log\frac{P_{X^nY^n}}{P_{X^n}(P_Y^*)^n} = \log\frac{P_{Z^n}(y^n - x^n)}{2^{-n}}$$

[1] Recall from Theorem 4.5 that the caod of a random transformation *always exists and is unique*, whereas a caid may not exist.

So under each hypothesis $\mathbb{P}$ and $\mathbb{Q}$, the difference $Y^n - X^n$ takes the form

$$\mathbb{Q} : Y^n - X^n \sim \text{Bern}(\frac{1}{2})^n$$
$$\mathbb{P} : Y^n - X^n \sim \text{Bern}(\delta)^n$$

Now all the relevant distributions are known, so we can compute β_α

$$\begin{aligned}
\beta_\alpha(P_{X^n Y^n}, P_{X^n}(P_Y^*)^n) &= \beta_\alpha(\text{Bern}(\delta)^n, \text{Bern}(\frac{1}{2})^n) \\
&= 2^{-nD(\text{Bern}(\delta)\|\text{Bern}(\frac{1}{2}))+o(n)} && \text{(Stein's Lemma Theorem 11.1)} \\
&= 2^{-nd(\delta\|\frac{1}{2})+o(n)}
\end{aligned}$$

Putting this all together, we see that any (n, M, ϵ) code for the BSC satisfies

$$2^{-nd(\delta\|\frac{1}{2})+o(n)} \le \frac{1}{M} \implies \log M \le nd(\delta\|\frac{1}{2}) + o(n)$$

Since this is satisfied for all codes, it is also satisfied for the optimal code, so we get the converse bound

$$\liminf_{n\to\infty} \frac{1}{n} \log M^*(n, \epsilon) \le d(\delta\|\frac{1}{2}) = \log 2 - h(\delta)$$

For a general channel, this computation can be much more difficult. The expression for β in this case is

$$\beta_{1-\epsilon}(P_{X^n} P_{Y^n|X^n}, P_{X^n}(P_Y^*)^n) = 2^{-nD(P_{Y|X}\|P_Y^*|\bar{P}_X)+o(n)} \le \frac{1}{M} \tag{20.3}$$

where $\bar{P}_X$ is unknown (depending on the code).

Explanation of (20.3): A statistician observes sequences of (X^n, Y^n):

$$\begin{array}{r|c|cc|cc|ccc}
\cline{2-2}\cline{5-6}
X^n = & [\,0 & 1 & 2 & 0 & 0 & 1 & 2 & 2\,] \\
Y^n = & [\,a & b & b & a & c & c & a & b\,] \\
\cline{2-2}\cline{5-6}
\end{array}$$

On the marked three blocks, test between iid samples of $P_{Y|X=0}$ vs P_Y^*, which has exponent $D(P_{Y|X=0}\|P_Y^*)$. Thus, intuitively averaging over the composition of the codeword we get that the exponent of β is given by (20.3).

Recall that from the saddle point characterization of capacity (Theorem 4.4) for any distribution $\bar{P}_X$ we have

$$D(P_{Y|X}\|P_Y^*|\bar{P}_X) \le C\,. \tag{20.4}$$

Thus from (20.3) and (20.1):

$$\log M \le nD(P_{Y|X}\|P_Y^*|\bar{P}_X) + o(n) \le nC + o(n)$$

Sketch 2: (More formal) Again, we will choose a dummy auxiliary channel $Q_{Y^n|X^n} = (Q_Y)^n$. However, choice of Q_Y will depend on one of the two cases:

1. If $|\mathcal{B}| < \infty$ we take $Q_Y = P_Y^*$ (the caod) and note that from (16.16) we have

$$\sum_y P_{Y|X}(y|x_0) \log^2 P_{Y|X}(y|x_0) \le \log^2 |\mathcal{B}| \qquad \forall x_0 \in \mathcal{A}$$

   and since $\min_y P_Y^*(y) > 0$ (without loss of generality), we conclude that for any distribution of X on $\mathcal{A}$ we have

$$\operatorname{Var}\left[\log \frac{P_{Y|X}(Y|X)}{Q_Y(Y)} | X\right] \le K < \infty \qquad \forall P_X\,. \tag{20.5}$$

   Furthermore, we also have from (20.4) that

$$\mathbb{E}\left[\log \frac{P_{Y|X}(Y|X)}{Q_Y(Y)} | X\right] \le C \qquad \forall P_X\,. \tag{20.6}$$

2. If $|\mathcal{A}| < \infty$, then for each codeword $c \in \mathcal{A}^n$ we define its *composition* as

$$\hat{P}_c(x) \triangleq \frac{1}{n} \sum_{j=1}^n 1\{c_j = x\}\,.$$

   By simple counting it is clear that from any (n, M, ϵ) code, it is possible to select an (n, M', ϵ) subcode, such that a) all codeword have the same composition P_0; and b) $M' > \frac{M}{n^{|\mathcal{A}|}}$. Note that, $\log M = \log M' + O(\log n)$ and thus we may replace M with M' and focus on the analysis of the chosen subcode. Then we set $Q_Y = P_{Y|X} \circ P_0$. In this case, from (16.9) we have

$$\operatorname{Var}\left[\log \frac{P_{Y|X}(Y|X)}{Q_Y(Y)} | X\right] \le K < \infty \qquad X \sim P_0\,. \tag{20.7}$$

   Furthermore, we also have

$$\mathbb{E}\left[\log \frac{P_{Y|X}(Y|X)}{Q_Y(Y)} | X\right] = D(P_{Y|X} \| Q_Y | P_0) = I(X;Y) \le C \qquad X \sim P_0\,. \tag{20.8}$$

Now, proceed as in (20.2) to get

$$\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(Q_Y)^n) \le \frac{1}{M}\,. \tag{20.9}$$

We next apply the lower bound on β from Theorem 10.5:

$$\gamma\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(Q_Y)^n) \ge \mathbb{P}\left[\log \frac{dP_{Y^n|X^n}(Y^n|X^n)}{d\prod Q_Y(Y_i)} \le \log\gamma\right] - \epsilon$$

Set $\log\gamma = nC + K'\sqrt{n}$ with K' to be chosen shortly and denote for convenience

$$S_n \triangleq \log \frac{dP_{Y^n|X^n}(Y^n|X^n)}{d\prod Q_Y(Y_i)} = \sum_{j=1}^n \log \frac{dP_{Y|X}(Y_j|X_j)}{dQ_Y(Y_j)}$$

Conditioning on X^n and using (20.6) and (20.8) we get

$$\mathbb{P}\left[S_n \le nC + K'\sqrt{n} | X^n\right] \ge \mathbb{P}\left[S_n \le n\,\mathbb{E}[S_n|X^n] + K'\sqrt{n} | X^n\right]$$

From here, we apply Chebyshev inequality and (20.5) or (20.7) to get

$$\mathbb{P}\Big[S_n \le n\,\mathbb{E}[S_n|X^n] + K'\sqrt{n}|X^n\Big] \ge 1 - \frac{K'^2}{K}\,.$$

If we set K' so large that $1 - \frac{K'^2}{K} > 2\epsilon$ then overall we get that

$$\log\beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(Q_Y)^n) \ge -nC - K'\sqrt{n} - \log\epsilon\,.$$

Consequently, from (20.9) we conclude that

$$\log M^*(n,\epsilon) \le nC + O(\sqrt{n})\,,$$

implying the strong converse. □

In summary, the take-away points for the strong converse are

1. Strong converse can be proven by using binary hypothesis testing.
2. The capacity saddle point (20.4) is key.

In the homework, we will explore in detail proofs of the strong converse for the BSC and the AWGN channel.

20.2 Stationary memoryless channel without strong converse

It may seem that the strong converse should hold for an arbitrary stationary memoryless channel (it was only showed for the *discrete* ones above). However, it turns out that there exist counterexamples. We construct one next.

Let output alphabet be $\mathcal{B} = [0,1]$. The input $\mathcal{A}$ is going to be countable, it will be convenient to define it as

$$\mathcal{A} = \{(j,m) : j,m \in \mathbb{Z}_+, 0 \le j \le m\}\,.$$

The single-letter channel $P_{Y|X}$ is defined in terms of probability density function as

$$p_{Y|X}(y|(j,m)) = \begin{cases} a_m, & \frac{j}{m} \le y \le \frac{j+1}{m}, , \\ b_m, & \text{otherwise}\,, \end{cases}$$

where a_m, b_m are chosen to satisfy

$$\frac{1}{m}a_m + (1-\frac{1}{m})b_m = 1 \tag{20.10}$$

$$\frac{1}{m}a_m\log a_m + (1-\frac{1}{m})b_m\log b_m = C\,, \tag{20.11}$$

where $C > 0$ is an arbitary fixed constant. Note that for large m we have

$$a_m = \frac{mC}{\log m}(1 + O(\frac{1}{\log m}))\,, \tag{20.12}$$

$$b_m = 1 - \frac{C}{\log m} + O(\frac{1}{\log^2 m}) \tag{20.13}$$

It is easy to see that $P_Y^* = \text{Unif}[0,1]$ is the capacity-achieving output distribution and

$$\sup_{P_X} I(X;Y) = C\,.$$

Thus by Theorem 16.6 the capacity of the corresponding stationary memoryless channel is C. We next show that nevertheless the ϵ-capacity can be strictly greater than C.

Indeed, fix blocklength n and consider a *single letter* distribution P_X assigning equal weights to all atoms (j,m) with $m = \exp\{2nC\}$. It can be shown that in this case, the distribution of a single-letter information density is given by

$$i(X;Y) \approx \begin{cases} 2nC, & w.p. \frac{1}{2n} \\ 0, & w.p. 1 - \frac{1}{2n} \end{cases}$$

Thus, for blocklength-n density we have

$$\frac{1}{n} i(X^n;Y^n) \to 2C\text{Poisson}(1/2)\,.$$

Therefore, from Theorem 15.1 we get that for $\epsilon > 1 - e^{-1/2}$ there exist (n,M,ϵ)-codes with

$$\log M \geq 2nC\,.$$

In particular,

$$C_\epsilon \geq 2C \quad \forall \epsilon > 1 - e^{-1/2}$$

20.3 Channel Dispersion

The strong converse tells us that $\log M^*(n,\epsilon) = nC + o(n)$ $\forall \epsilon \in (0,1)$. An engineer sees this, and estimates $\log M^* \approx nC$. However, this doesn't give any information about the dependence of $\log M^*$ on the error probability ϵ, which is hidden in the $o(n)$ term. We unravel this in the following theorem.

Theorem 20.2. *Consider one of the following channels:*

1. *DMC*
2. *DMC with cost constraint*
3. *AWGN or parallel AWGN*

The following expansion holds for a fixed $0 < \epsilon < 1/2$ *and* $n \to \infty$

$$\log M^*(n,\epsilon) = nC - \sqrt{nV} Q^{-1}(\epsilon) + O(\log n)$$

where Q^{-1} *is the inverse of the complementary standard normal CDF, the channel capacity is* $C = I(X^*;Y^*) = \mathbb{E}[i(X^*;Y^*)]$*, and the channel dispersion*[2] *is* $V = \text{Var}[i(X^*;Y^*)|X^*]$.

[2] There could be multiple capacity-achieving input distributions, in which case P_{X^*} should be chosen as the one that minimizes $\text{Var}[i(X^*;Y^*)|X^*]$. See [PPV10] for more details.

Proof. For achievability, we have shown (Theorem 16.7) that $\log M^*(n,\epsilon) \ge nC - \sqrt{nV}Q^{-1}(\epsilon)$ by refining the proof of the noisy channel coding theorem using the CLT.

The converse statement is $\log M^* \le -\log \beta_{1-\epsilon}(P_{X^nY^n}, P_{X^n}(P_Y^*)^n)$. For the BSC, we showed that the RHS of the previous expression is

$$-\log \beta_{1-\epsilon}(\text{Bern}(\delta)^n, \text{Bern}(\frac{1}{2})^n) = nd(\delta\|\frac{1}{2}) + \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n})$$

(see homework) where the dispersion is

$$V = \text{Var}_{Z\sim\text{Bern}(\delta)}\left[\log \frac{\text{Bern}(\delta)}{\text{Bern}(\frac{1}{2})}(Z)\right].$$

The general proof is omitted. □

Remark: This expansion only applies for certain channels (as described in the theorem). If, for example, $\text{Var}[i(X;Y)] = \infty$, then the theorem need not hold and there are other stable (non-Gaussian) distributions that we might converge to instead. Also notice that for DMC without cost constraint

$$\text{Var}[i(X^*;Y^*)|X^*] = \text{Var}[i(X^*;Y^*)]$$

since (capacity saddle point!) $\mathbb{E}[i(X^*;Y^*)|X^* = x] = C$ for P_{X^*}-almost all x.

20.3.1 Applications

As stated earlier, direct computation of $M^*(n,\epsilon)$ by exhaustive search doubly exponential in complexity, and thus is infeasible in most cases. However, we can get an easily computable approximation using the channel dispersion via

$$\log M^*(n,\epsilon) \approx nC - \sqrt{nV}Q^{-1}(\epsilon)$$

Consider a BEC ($n = 500, \delta = 1/2$) as an example of using this approximation. For this channel, the capacity and dispersion are

$$C = 1 - \delta$$
$$V = \delta\bar{\delta}$$

Where $\bar{\delta} = 1 - \delta$. Using these values, our approximation for this BEC becomes

$$\log M^*(500, 10^{-3}) \approx nC - \sqrt{nV}Q^{-1}(\epsilon) = n\bar{\delta} - \sqrt{n\delta\bar{\delta}}Q^{-1}(10^{-3}) \approx 215.5 \text{ bits}$$

In the homework, for the BEC(500, 1/2) we obtained bounds $213 \le \log M^*(500, 10^{-3}) \le 217$, so this approximation falls in the middle of these bounds.

Examples of Channel Dispersion

For a few common channels, the dispersions are

$$\text{BEC: } V(\delta) = \delta\bar{\delta}\log^2 2$$

$$\text{BSC: } V(\delta) = \delta\bar{\delta}\log^2 \frac{\bar{\delta}}{\delta}$$

$$\text{AWGN: } V(P) = \frac{P(P+2)}{2(P+1)^2}\log^2 e \ \text{(Real)} \qquad \frac{P(P+2)}{(P+1)^2}\log^2 e \ \text{(Complex)}$$

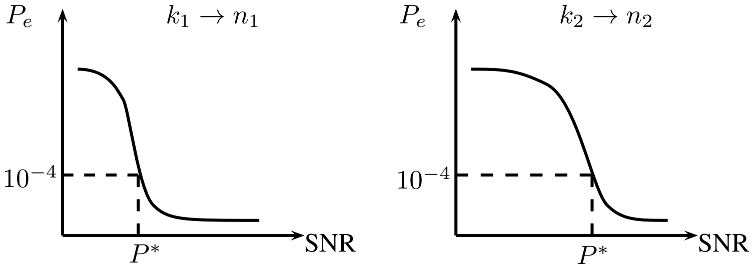
$$\text{Parallel AWGN: } V(\mathbf{P},\sigma^2) = \sum_{j=1}^{L} V_{AWGN}\left(\frac{P_j}{\sigma_j^2}\right) = \frac{\log^2 e}{2}\sum_{j=1}^{L}\left|1-\left(\frac{\sigma_j^2}{T}\right)^2\right|^+$$

$$\text{where } \sum_{j=1}^{L}|T-\sigma_j^2|^+ = P \text{ is the water-filling solution of the parallel AWGN}$$

Punchline: Although the only machinery needed for this approximation is the CLT, the results produced are incredibly useful. Even though $\log M^*$ is nearly impossible to compute on its own, by only finding C and V we are able to get a good approximation that is easily computable.

20.4 Normalized Rate

Suppose you're given two codes $k_1 \to n_1$ and $k_2 \to n_2$, how do you fairly compare them? Perhaps they have the following waterfall plots



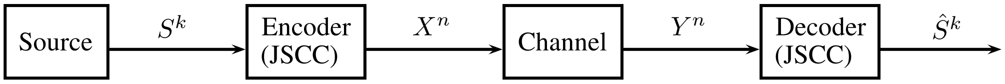
After inspecting these plots, one may believe that the $k_1 \to n_1$ code is better, since it requires a smaller SNR to achieve the same error probability. However, there are many factors, such as blocklength, rate, etc. that don't appear on these plots. To get a fair comparison, we can use the notion of *normalized rate*. To each $(n, 2^k, \epsilon)$-code, define

$$R_{\text{norm}} = \frac{k}{\log_2 M^*_{AWGN}(n,\epsilon,P)} \approx \frac{k}{nC(P) - \sqrt{nV(P)}Q^{-1}(\epsilon)}$$

Take $\epsilon = 10^{-4}$, and P (SNR) according to the water fall plot corresponding to $P_e = 10^{-4}$, and we can compare codes directly (see Fig. 20.1). This normalized rate gives another motivation for the expansion given in Theorem 20.2.

20.5 Joint Source Channel Coding

Now we will examine a slightly different information transmission scenario called *Joint Source Channel Coding*



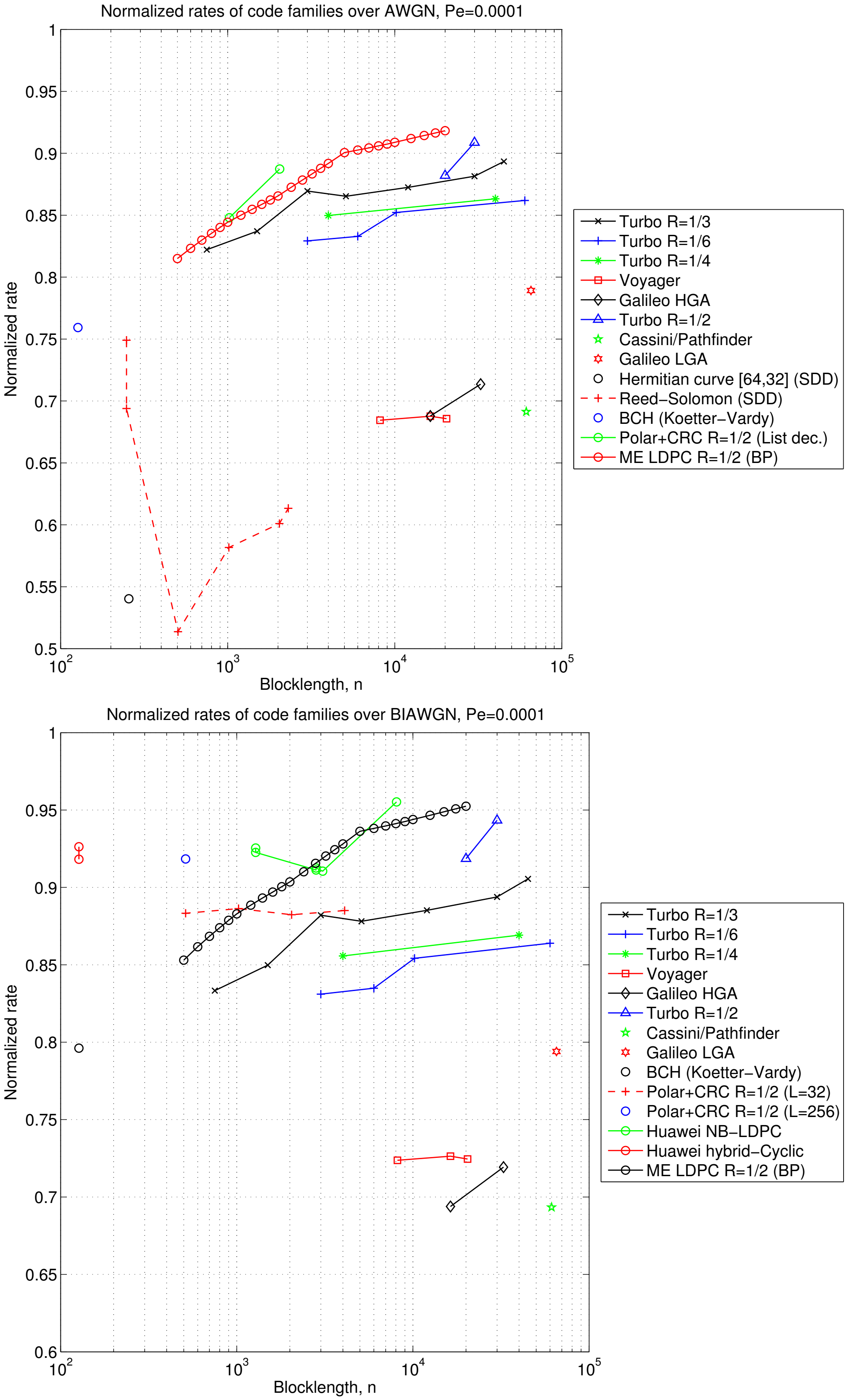


Figure 20.1: Normalized rates for various codes. Plots generated via [Spe15].

Definition 20.1. For a Joint Source Channel Code

- Goal: $\mathbb{P}[S^k \neq \hat{S}^k] \leq \epsilon$
- Encoder: $f : \mathcal{A}^k \to \mathcal{X}^n$
- Decoder: $g : \mathcal{Y}^n \to \mathcal{A}^k$
- Fundamental Limit (Optimal probability of error): $\epsilon^*_{JSCC}(k,n) = \inf_{f,g} \mathbb{P}[S^k \neq \hat{S}^k]$

where the rate is $R = \frac{k}{n}$ (symbol per channel use).

Note: In channel coding we are interested in transmitting M messages and all messages are born equal. Here we want to convey the source realizations which might not be equiprobable (has redundancy). Indeed, if S^k is uniformly distributed on, say, $\{0,1\}^k$, then we are back to the channel coding setup with $M = 2^k$ under average probability of error, and $\epsilon^*_{JSCC}(k,n)$ coincides with $\epsilon^*(n, 2^k)$ defined in Section 20.1.
Note: Here, we look for a clever scheme to directly encode k symbols from $\mathcal{A}$ into a length n channel input such that we achieve a small probability of error over the channel. This feels like a mix of two problems we've seen: compressing a source and coding over a channel. The following theorem shows that compressing and channel coding separately is optimal. This is a relief, since it implies that we do not need to develop any new theory or architectures to solve the Joint Source Channel Coding problem. As far as the leading term in the asymptotics is concerned, the following two-stage scheme is optimal: First use the optimal compressor to eliminate all the redundancy in the source, then use the optimal channel code to add redundancy to combat the noise in the transmission.

Theorem 20.3. *Let the source $\{S_k\}$ be stationary memoryless on a finite alphabet with entropy H. Let the channel be stationary memoryless with finite capacity C. Then*

$$\epsilon^*_{JSCC}(nR,n) \begin{cases} \to 0 & R < C/H \\ \not\to 0 & R > C/H \end{cases} \quad n \to \infty.$$

Note: Interpretation: Each source symbol has information content (entropy) H bits. Each channel use can convey C bits. Therefore to reliably transmit k symbols over n channel uses, we need $kH \leq nC$.

Proof. **Achievability.** The idea is to separately compress our source and code it for transmission. Since this is a feasible way to solve the JSCC problem, it gives an achievability bound. This separated architecture is

$$S^k \xrightarrow{f_1} W \xrightarrow{f_2} X^n \overset{P_{Y^n|X^n}}{\longrightarrow} Y^n \xrightarrow{g_2} \hat{W} \xrightarrow{g_1} \hat{S}^k$$

Where we use the optimal compressor (f_1, g_1) and optimal channel code (<u>maximum</u> probability of error) (f_2, g_2). Let W denote the output of the compressor which takes at most M_k values. Then

$$\text{(From optimal compressor) } \frac{1}{k}\log M_k > H + \delta \implies \mathbb{P}[\hat{S}^k \neq S^k(W)] \leq \epsilon \quad \forall k \geq k_0$$

$$\text{(From optimal channel code) } \frac{1}{n}\log M_k < C - \delta \implies \mathbb{P}[\hat{W} \neq m | W = m] \leq \epsilon \quad \forall m, \forall k \geq k_0$$

Using both of these,

$$\begin{aligned}\mathbb{P}[S^k \neq \hat{S}^k(\hat{W})] &\leq \mathbb{P}[S^k \neq \hat{S}^k, W = \hat{W}] + \mathbb{P}[W \neq \hat{W}] \\ &\leq \mathbb{P}[S^k \neq \hat{S}^k(W)] + \mathbb{P}[W \neq \hat{W}] \leq \epsilon + \epsilon\end{aligned}$$

And therefore if $R(H+\delta) < C - \delta$, then $\epsilon^* \to 0 \overset{\delta\to 0}{\Longrightarrow} R > C/H$.

Converse: channel-substitution proof. Let $Q_{S^k\hat{S}^k} = U_{S^k}P_{\hat{S}^k}$ where U_{S^k} is the uniform distribution. Using data processing

$$D(P_{S^k\hat{S}^k}\|Q_{S^k\hat{S}^k}) = D(P_{S^k}\|U_{S^k}) + D(P_{\hat{S}|S^k}\|P_{\hat{S}}|P_{S^k}) \geq d(1-\epsilon\|\frac{1}{|\mathcal{A}|^k})$$

Rearranging this gives

$$\begin{aligned}I(S^k;\hat{S}^k) &\geq d(1-\epsilon\|\frac{1}{|\mathcal{A}|^k}) - D(P_{S^k}\|U_{S^k}) \\ &\geq -\log 2 + k\bar{\epsilon}\log|\mathcal{A}| + H(S^k) - k\log|\mathcal{A}| \\ &= H(S^k) - \log 2 - k\epsilon\log|\mathcal{A}|\end{aligned}$$

Which follows from expanding out the terms. Now, normalizing and taking the sup of both sides gives

$$\frac{1}{n}\sup_{X^n} I(X^n;Y^n) \geq \frac{1}{n}H(S^k) - \epsilon\frac{k}{n}\log|A| + o(1)$$

letting $R = k/n$, this shows

$$C \geq RH - \epsilon R\log|A| \implies \epsilon \geq \frac{RH - C}{R\log|A|} > 0$$

where the last expression is positive when $R > C/H$.

Converse: usual proof. Any JSCC encoder/decoder induces a Markov chain

$$S^k \to X^n \to Y^n \to \hat{S}^k.$$

Applying data processing for mutual information

$$I(S^k;\hat{S}^k) \leq I(X^n;Y^n) \leq \sup_{P_{X^n}} I(X^n;Y^n) = nC.$$

On the other hand, since $\mathbb{P}[S^k \neq \hat{S}^k] \leq \epsilon_n$, Fano's inequality yields

$$I(S^k;\hat{S}^k) = H(S^k) - H(S^k|\hat{S}^k) \geq kH - \epsilon_n\log|\mathcal{A}|^k - \log 2.$$

Combining the two gives

$$nC \geq kH - \epsilon_n\log|\mathcal{A}|^k - \log 2.$$

Since $R = \frac{k}{n}$, dividing both sides by n and sending $n \to \infty$ yields

$$\liminf_{n\to\infty}\epsilon_n \geq \frac{RH - C}{R\log|\mathcal{A}|}.$$

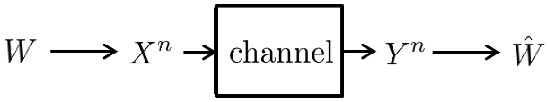
Therefore ϵ_n does not vanish if $R > C/H$. □

§ 21. Channel coding with feedback

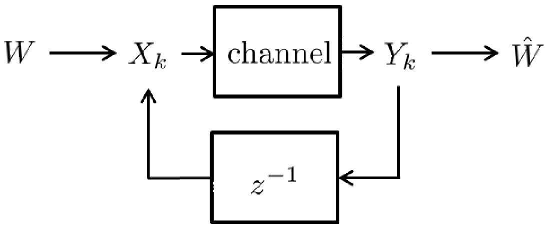
Criticism: Channels without feedback don't exist (except storage).

Motivation: Consider the communication channel of the downlink transmission from a satellite to earth. Downlink transmission is very expensive (power constraint at the satellite), but the uplink from earth to the satellite is cheap which makes virtually noiseless feedback readily available at the transmitter (satellite). In general, channel with noiseless feedback is interesting when such asymmetry exists between uplink and downlink.

In the first half of our discussion, we shall follow Shannon to show that feedback gains "nothing" in the conventional setup, while in the second half, we look at situations where feedback gains a lot.



channel w/o feedback



causality constraint
noiseless / Shannon / ideal feedback channel

channel with feedback

21.1 Feedback does not increase capacity for stationary memoryless channels

Definition 21.1 (Code with feedback)**.** An (n, M, ϵ)-code with feedback is specified by the encoder-decoder pair (f, g) as follows:

- Encoder: (time varying)

$$\begin{aligned} &f_1 : [M] \to \mathcal{A} \\ &f_2 : [M] \times \mathcal{B} \to \mathcal{A} \\ &\vdots \\ &f_n : [M] \times \mathcal{B}^{n-1} \to \mathcal{A} \end{aligned}$$

- Decoder:

$$g : \mathcal{B}^n \to [M]$$

such that $\mathbb{P}[W \neq \hat{W}] \leq \epsilon$.

Note: [Probability space]

$$W \sim \text{ uniform on } [M]$$
$$\left.\begin{array}{l} X_1 = f_1(W) \xrightarrow{P_{Y|X}} Y_1 \\ \vdots \\ X_n = f_n(W, Y_1^{n-1}) \xrightarrow{P_{Y|X}} Y_n \end{array}\right\} \longrightarrow \hat{W} = g(Y^n)$$

Definition 21.2 (Fundamental limits)**.**

$$\begin{aligned} M^*_{fb}(n,\epsilon) &= \max\{M : \exists (n, M, \epsilon) \text{ code with feedback.}\} \\ C_{fb,\epsilon} &= \liminf_{n\to\infty} \frac{1}{n} \log M^*_{fb}(n,\epsilon) \\ C_{fb} &= \lim_{\epsilon\to 0} C_{fb,\epsilon} && \text{(Shannon capacity with feedback)} \end{aligned}$$

Theorem 21.1 (Shannon 1956)**.** *For a stationary memoryless channel,*

$$C_{fb} = C = C_i = \sup_{P_X} I(X;Y)$$

Proof. Achievability: Although it is obvious that $C_{fb} \ge C$, we wanted to demonstrate that in fact constructing codes achieving capacity with *full feedback* can be done directly, without appealing to a (much harder) problem of non-feedback codes. Let $\pi_t(\cdot) \triangleq P_{W|Y^t}(\cdot|Y^t)$ with the (random) posterior distribution after t steps. It is clear that due to the knowledge of Y^t on both ends, transmitter and receiver have perfectly synchronized knowledge of π_t. Now consider how the transmission progresses:

1. Initialize $\pi_0(\cdot) = \frac{1}{M}$
2. At $(t+1)$-th step, having knowledge of π_t all messages are partitioned into classes $\mathcal{P}_a$, according to the values $f_{t+1}(\cdot, Y^t)$:
$$\mathcal{P}_a \triangleq \{j \in [M] : f_{t+1}(j, Y^t) = a\} \qquad a \in \mathcal{A}\,.$$
Then transmitter, possessing the knowledge of the true message W, selects a letter $X_{t+1} = f_{t+1}(W, Y^t)$.
3. Channel perturbs X_{t+1} into Y_{t+1} and both parties compute the updated posterior:
$$\pi_{t+1}(j) \triangleq \pi_t(j) B_{t+1}(j)\,, \qquad B_{t+1}(j) \triangleq \frac{P_{Y|X}(Y_{t+1}|f_{t+1}(j, Y^t))}{\sum_{a\in\mathcal{A}} \pi_t(\mathcal{P}_a)}\,.$$
Notice that (this is the crucial part!) the random multiplier satisfies:
$$\mathbb{E}[\log B_{t+1}(W)|Y^t] = \sum_{a\in\mathcal{A}} \sum_{y\in\mathcal{B}} \pi_t(\mathcal{P}_a) \log \frac{P_{Y|X}(y|a)}{\sum_{a\in\mathcal{A}} \pi_t(\mathcal{P}_a) a} = I(\tilde{\pi}_t, P_{Y|X}) \tag{21.1}$$
where $\tilde{\pi}_t(a) \triangleq \pi_t(\mathcal{P}_a)$ is a (random) distribution on $\mathcal{A}$.

The goal of the code designer is to come up with such a partitioning $\{\mathcal{P}_a, a \in \mathcal{A}\}$ that the speed of growth of $\pi_t(W)$ is maximal. Now, analyzing the speed of growth of a random-multiplicative process is best done by taking logs:

$$\log \pi_t(j) = \sum_{s=1}^{t} \log B_s + \log \pi_0(j)\,.$$

Intutively, we expect that the process $\log \pi_t(W)$ resembles a random walk starting from $-\log M$ and having a positive drift. Thus to estimate the time it takes for this process to reach value 0 we need to estimate the upward drift. Appealing to intuition and the law of large numbers we approximate

$$\log \pi_t(W) - \log \pi_0(W) \approx \sum_{s=1}^{t} \mathbb{E}[\log B_s] .$$

Finally, from (21.1) we conclude that the best idea is to select partitioning at each step in such a way that $\tilde{\pi}_t \approx P_X^*$ (caid) and this obtains

$$\log \pi_t(W) \approx tC - \log M ,$$

implying that the transmission terminates in time $\approx \frac{\log M}{C}$. The important lesson here is the following: *The optimal transmission scheme should map messages to channel inputs in such a way that the induced input distribution $P_{X_{t+1}|Y^t}$ is approximately equal to the one maximizing $I(X;Y)$.* This idea is called *posterior matching* and explored in detail in [SF11].[1]

Converse: we are left to show that $C_{fb} \le C_i$.

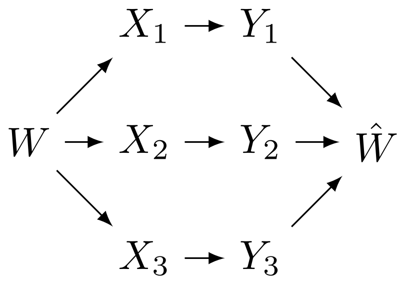
Recall the key in proving weak converse for channel coding without feedback: Fano's inequality plus the graphical model

$$W \to X^n \to Y^n \to \hat{W}. \tag{21.2}$$

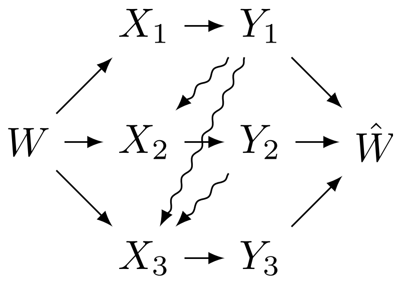
Then

$$h(\epsilon) + \bar{\epsilon} \log M \le I(W;\hat{W}) \le I(X^n;Y^n) \le nC_i.$$

With feedback the probabilistic picture becomes more complicated as the following figure shows for $n = 3$ (dependence introduced by the extra squiggly arrows):



without feedback



with feedback

So, while the Markov chain realtion in (21.2) is still true, we also have

$$P_{Y^n|X^n}(y^n|x^n) \ne \prod_{j=1}^{n} P_{Y|X}(y_j|x_j) \qquad (!)$$

(This is easy to see from the example where $X_2 = Y_1$ and thus $P_{Y_1|X^2}$ has no randomness.) There is still a large degree of independence in the channel, though. Namely, we have

$$(Y^{i-1}, W) \to X_i \to Y_i, \quad i = 1, \dots, n \tag{21.3}$$

$$W \to Y^n \to \hat{W} \tag{21.4}$$

[1] Note that the magic of Shannon's theorem is that this optimal partitioning can also be done blindly. I.e. it is possible to preselect partitions $\mathcal{P}_a$ in a way independent of π_t (but dependent on t) and so that the $\pi_t(\mathcal{P}_a) \approx P_X^*(a)$ with overwhelming probability and for all $t \in [1,n]$.

Then

$$
\begin{aligned}
h(\epsilon) + \bar{\epsilon}\log M &\le I(W;\hat{W}) && \text{(Fano)}\\
&\le I(W;Y^n) && \text{(Data processing applied to (21.4))}\\
&= \sum_{i=1}^n I(W;Y_i|Y^{i-1}) && \text{(Chain rule)}\\
&\le \sum_{i=1}^n I(W,Y^{i-1};Y_i) && (I(W;Y_i|Y^{i-1}) = I(W,Y^{i-1};Y_i) - I(Y^{i-1};Y_i))\\
&\le \sum_{i=1}^n I(X_i;Y_i) && \text{(Data processing applied to (21.3))}\\
&\le nC_i && \square
\end{aligned}
$$

The following result (without proof) suggests that feedback does not even improve the speed of approaching capacity either (under fixed-length block coding) and can at most improve smallish $\log n$ terms:

Theorem 21.2 (Dispersion with feedback)**.** *For weakly input-symmetric DMC (e.g. additive noise, BSC, BEC) we have:*

$$\log M^*_{fb}(n,\epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)$$

(The meaning of this is that for such channels feedback can at most improve smallish $\log n$ terms.)

21.2* Alternative proof of Theorem 21.1 and Massey's directed information

The following alternative proof emphasizes on data processing inequality and the comparison idea (auxiliary channel) as in Theorem 19.1.

Proof. It is obvious that $C_{fb} \ge C$, we are left to show that $C_{fb} \le C_i$.

1. Recap of the steps of showing the strong converse of $C \le C_i$ in the last lecture: take any (n, M, ϵ) code, compare the two distributions:

$$P : W \to X^n \to Y^n \to \hat{W} \tag{21.5}$$

$$Q : W \to X^n \quad Y^n \to \hat{W} \tag{21.6}$$

   two key observations:

   a) Under Q, $W \perp\!\!\!\perp W$, so that $\mathbb{Q}[W = \hat{W}] = \frac{1}{M}$ while $\mathbb{P}[W = \hat{W}] \ge 1 - \epsilon$.

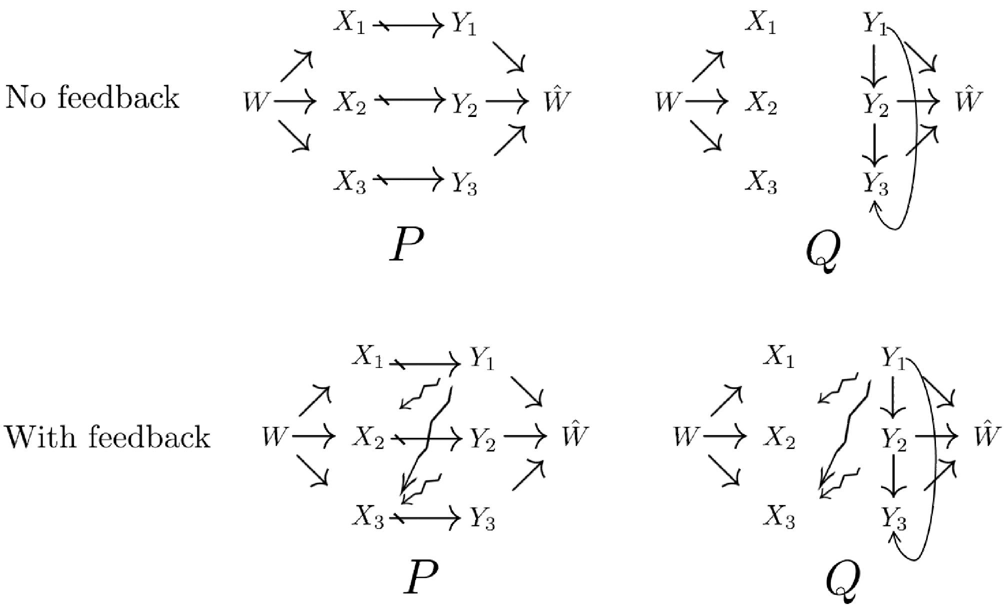
   b) The two graphical models give the factorization:

$$P_{W,X^n,Y^n,\hat{W}} = P_{W,X^n}P_{Y^n|X^n}P_{\hat{W}|Y^n}, \qquad Q_{W,X^n,Y^n,\hat{W}} = P_{W,X^n}P_{Y^n}P_{\hat{W}|Y^n}$$

   thus $D(P\|Q) = I(X^n;Y^n)$ measures the information flow through the links $X^n \to Y^n$.

$$h(\epsilon) + \bar{\epsilon}\log M = d(1-\epsilon\|\frac{1}{M}) \overset{\text{d-proc ineq}}{\le} D(P\|Q) = I(X^n;Y^n) \overset{mem-less,stat}{=} \sum_{i=1}^n I(X;Y) \le nC_i \tag{21.7}$$

2. Notice that when feedback is present, $X^n \to Y^n$ is not memoryless due to the transmission protocol, let's unfold the probability space over time to see the dependence. As an example, the graphical model for $n = 3$ is given below:



If we define Q similarly as in the case without feedback, we will encounter a problem at the second last inequality in (21.7), as with feedback $I(X^n;Y^n)$ can be significantly larger than $\sum_{i=1}^n I(X;Y)$. Consider the example where $X_2 = Y_1$, we have $I(X^n;Y^n) = +\infty$ independent of $I(X;Y)$.

We also make the observe that if Q is defined in (21.6), $D(P\|Q) = I(X^n;Y^n)$ measures the information flow through all the $\nrightarrow$ and $\rightsquigarrow$ links. This motivates us to find a proper Q such that $D(P\|Q)$ only captures the information flow through all the $\nrightarrow$ links $\{X_i \to Y_i : i = 1,\dots,n\}$, so that $D(P\|Q)$ closely relates to nC_i, while still guarantees that $W \perp\!\!\!\perp W$, so that $\mathbb{Q}[W \neq \hat{W}] = \frac{1}{M}$.

3. Formally, we shall restrict $Q_{W,X^n,Y^n,\hat{W}} \in \mathcal{Q}$, where $\mathcal{Q}$ is the set of distributions that can be factorized as follows:

$$Q_{W,X^n,Y^n,\hat{W}} = Q_W Q_{X_1|W} Q_{Y_1} Q_{X_2|W,Y_1} Q_{Y_2|Y_1} \cdots Q_{X_n|W,Y^{n-1}} Q_{Y_n|Y^{n-1}} Q_{\hat{W}|Y^n} \tag{21.8}$$

$$P_{W,X^n,Y^n,\hat{W}} = P_W P_{X_1|W} P_{Y_1|X_1} P_{X_2|W,Y_1} P_{Y_2|X_2} \cdots P_{X_n|W,Y^{n-1}} P_{Y_n|X_n} P_{\hat{W}|Y^n} \tag{21.9}$$

Verify that $W \perp\!\!\!\perp W$ under Q: W and $\hat{W}$ are d-separated by X^n.

Notice that in the graphical models, when removing $\nrightarrow$ we also added the directional links between the Y_is, these links serve to maximally preserve the dependence relationships between variables when $\nrightarrow$ are removed, so that Q is the "closest" to P while $W \perp\!\!\!\perp W$ is satisfied.

Now we have that for $Q \in \mathcal{Q}$, $d(1-\epsilon\|\frac{1}{M}) \leq D(P\|Q)$, in order to obtain the least upper bound, in Lemma 21.1 we shall show that:

$$\begin{aligned}\inf_{Q\in\mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}}\|Q_{W,X^n,Y^n,\hat{W}}) &= \sum_{k=1}^{n} I(X_k;Y_k|Y^{k-1})\\ &= \sum_{k=1}^{n} \mathbb{E}_{Y^{k-1}}[I(P_{X_k|Y^{k-1}},P_{Y|X})]\\ &\le \sum_{k=1}^{n} I(\mathbb{E}_{Y^{k-1}}[P_{X_k|Y^{k-1}}],P_{Y|X}) \quad (\text{concavity of } I(P_X,P_{Y|X}) \text{ in } P_X)\\ &= \sum_{k=1}^{n} I(P_{X_k},P_{Y|X})\\ &\le nC_i.\end{aligned}$$

Following the same procedure as in (a) we have

$$h(\epsilon)+\bar{\epsilon}\log M \le nC_i \Rightarrow \log M \le \frac{nC+h(\epsilon)}{1-\epsilon} \Rightarrow C_{fb,\epsilon} \le \frac{C}{1-\epsilon} \Rightarrow C_{fb} \le C.$$

4. Notice that the above proof is also valid even when cost constraint is present.

□

Lemma 21.1.

$$\inf_{Q\in\mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}}\|Q_{W,X^n,Y^n,\hat{W}}) = \sum_{k=1}^{n} I(X_k;Y_k|Y^{k-1}) \tag{21.10}$$

$$(\triangleq \vec{I}(X^n;Y^n),\ \textbf{\textit{directed information}})$$

Proof. By chain rule, we can show that the minimizer $Q\in\mathcal{Q}$ must satisfy the following equalities:

$$\begin{aligned}Q_{X,W} &= P_{X,W},\\ Q_{X_k|W,Y^{k-1}} &= P_{X_k|W,Y^{k-1}}, \quad (\text{check!})\\ Q_{\hat{W}|Y^n} &= P_{W|Y^n}.\end{aligned}$$

and therefore

$$\begin{aligned}&\inf_{Q\in\mathcal{Q}} D(P_{W,X^n,Y^n,\hat{W}}\|Q_{W,X^n,Y^n,\hat{W}})\\ &= D(P_{Y_1|X_1}\|Q_{Y_1}|X_1)+D(P_{Y_2|X_2,Y_1}\|Q_{Y_2|Y_1}|X_2,Y_1)+\cdots+D(P_{Y_n|X_n,Y^{n-1}}\|Q_{Y_n|Y^{n-1}}|X_n,Y^{n-1})\\ &= I(X_1;Y_1)+I(X_2;Y_2|Y_1)+\cdots+I(X_n;Y_n|Y^{n-1})\end{aligned}$$

□

21.3 When is feedback really useful?

Theorems 21.1 and 21.2 state that feedback does not improve communication rate neither asymptotically nor for moderate blocklengths. In this section, we shall examine three cases where feedback turns out to be very useful.

21.3.1 Code with very small (e.g. zero) error probability

Theorem 21.3 (Shannon '56). *For any DMC $P_{Y|X}$,*

$$C_{fb,0} = \max_{P_X} \min_{y\in\mathcal{B}} \log \frac{1}{P_X(S_y)} \tag{21.11}$$

where

$$S_y = \{a \in \mathcal{A} : P_{Y|X}(y|a) > 0\}$$

denotes the set of input symbols that can lead to the output symbol y.

Note: For stationary memoryless channel,

$$C_0 \overset{\text{def.}}{\le} C_{fb,0} \overset{\text{def.}}{\le} C_{fb} = \lim_{\epsilon\to0} C_{fb,\epsilon} \overset{\text{Thm 21.1}}{=} C = \lim_{\epsilon\to0} C_\epsilon \overset{\text{Shannon}}{=} C_i = \sup_{P_X} I(X;Y)$$

All capacity quantities above are defined with (fixed-length) block codes.

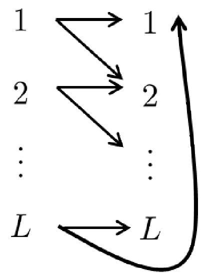
Observations:

1. In DMC for both zero-error capacities (C_0 and $C_{fb,0}$) only the support of the transition matrix $P_{Y|X}$, i.e., whether $P_{Y|X}(b|a) > 0$ or not, matters. The value of $P_{Y|X}(b|a) > 0$ is irrelevant. That is, C_0 and $C_{fb,0}$ are functions of a bipartite graph between input and output alphabets. Furthermore, the C_0 (but not $C_{fb,0}$!) is a function of the *confusability graph* – a simple undirected graph on $\mathcal{A}$ with $a \ne a'$ connected by an edge iff $\exists b \in \mathcal{B}$ s.t. $P_{Y|X}(b|a)P_{Y|X}(b|a') > 0$.

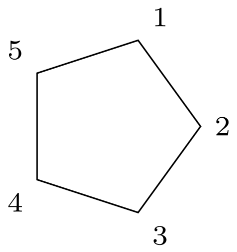
2. That $C_{fb,0}$ is not a function of the confusability graph alone is easily seen from comparing the polygon channel (next remark) with $L = 3$ (for which $C_{fb,0} = \log\frac{3}{2}$) and the useless channel with $\mathcal{A} = \{1,2,3\}$ and $\mathcal{B} = \{1\}$ (for which $C_{fb,0} = 0$). Clearly in both cases confusability graph is the same – a triangle.

3. Usually C_0 is very hard to compute, but $C_{fb,0}$ can be obtained in closed form as in (21.11).

   Example: (Polygon channel)



Bipartite graph



Confusability graph

- Zero-error capacity C_0:
  - $L = 3$: $C_0 = 0$
  - $L = 5$: $C_0 = \frac{1}{2}\log 5$ (Shannon '56-Lovasz '79).
    Achievability:
    a) blocklength one: $\{1,3\}$, rate = 1 bit.
    b) blocklength two: $\{(1,1),(2,3),(3,5),(4,2),(5,4)\}$, rate = $\frac{1}{2}\log 5$ bit – optimal!

  - $L = 7$: $3/5 \log 7 \le C_0 \le \log 3.32$ (Exact value unknown to this day)
  - Even $L = 2k$: $C_0 = \log \frac{L}{2}$ for all k (Why? Homework.).
  - Odd $L = 2k+1$: $C_0 = \log \frac{L}{2} + o(1)$ as $k \to \infty$ (Bohman '03)
- Zero-error capacity with feedback (proof: exercise!)

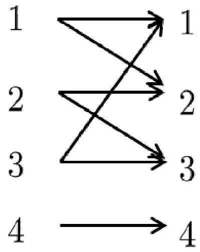
$$C_{fb,0} = \log \frac{L}{2}, \quad \forall L,$$

which can be strictly bigger than C_0.

4. Notice that $C_{fb,0}$ is not necessarily equal to $C_{fb} = \lim_{\epsilon \to 0} C_{fb,\epsilon} = C$. Here is an example when

$$C_0 < C_{fb,0} < C_{fb} = C$$

Example:



Then

$$\begin{aligned} C_0 &= \log 2 \\ C_{fb,0} &= \max_{\delta} -\log \max(\frac{2}{3}\delta, 1-\delta) && (P_X^* = (\delta/3, \delta/3, \delta/3, \bar{\delta})) \\ &= \log \frac{5}{2} > C_0 && (\delta^* = \frac{3}{5}) \end{aligned}$$

On the other hand, Shannon capacity $C = C_{fb}$ can be made arbitrarily close to $\log 4$ by picking the cross-over probability arbitrarily close to zero, while the confusability graph stays the same.

Proof of Theorem 21.3. 1. Fix any $(n, M, 0)$-code. Denote the confusability set of all possible messages that could have produced the received signal $y^t = (y_1, \ldots, y_t)$ for all $t = 0, 1, \ldots, n$ by:

$$E_t(y^t) \triangleq \{m \in [M] : f_1(m) \in S_{y_1}, f_2(m, y_1) \in S_{y_2}, \ldots, f_n(m, y^{t-1}) \in S_{y_t}\}$$

Notice that zero-error means no ambiguity:

$$\epsilon = 0 \Leftrightarrow \forall y^n \in \mathcal{B}^n, |E_n(y^n)| = 1 \text{ or } 0. \tag{21.12}$$

2. The key quantities in the proof are defined as follows:

$$\theta_{fb} = \min_{P_X} \max_{y \in \mathcal{B}} P_X(S_y),$$
$$P_X^* = \operatorname*{argmin}_{P_X} \max_{y \in \mathcal{B}} P_X(S_y)$$

By definition, we have

$$\forall P_X, \exists y \in \mathcal{B}, \text{ such that } P_X(S_y) \geq \theta_{fb} \tag{21.13}$$

Notice the minimizer distribution P_X^* is usually not the caid in the usual sense. This definition also sheds light on how the encoding and decoding should be proceeded and serves to lower bound the uncertainty reduction at each stage of the decoding scheme.

3. "≤" (converse): Let P_{X^n} be he joint distribution of the codewords. Denote $E_0 = [M]$ – original message set.

   $\underline{t = 1}$: For P_{X_1}, by (21.13), $\exists y_1^*$ such that:

   $$P_{X_1}(S_{y_1^*}) = \frac{|\{m : f_1(m) \in S_{y_1^*}\}|}{|\{m \in [M]\}|} = \frac{|E_1(y_1^*)|}{|E_0|} \geq \theta_{fb}.$$

   $\underline{t = 2}$: For $P_{X_2|X_1 \in S_{y_1^*}}$, by (21.13), $\exists y_2^*$ such that:

   $$P_{X_2}(S_{y_2^*}|X_1 \in S_{y_1^*}) = \frac{|\{m : f_1(m) \in S_{y_1^*}, f_2(m, y_1^*) \in S_{y_2^*}\}|}{|\{m : f_1(m) \in S_{y_1^*}\}|} = \frac{|E_2(y_1^*, y_2^*)|}{|E_1(y_1^*)|} \geq \theta_{fb},$$

   $\underline{t = n}$: Continue the selection process up to y_n^* which satisfies that:

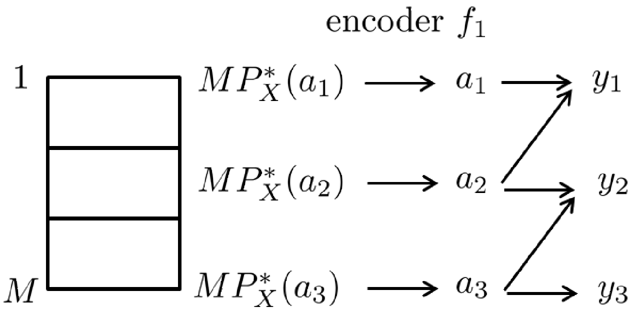
   $$P_{X_n}(S_{y_n^*}|X_k \in S_{y_k^*} \text{ for } k = 1, \ldots, n-1) = \frac{|E_n(y_1^*, \ldots, y_n^*)|}{|E_{n-1}(y_1^*, \ldots, y_{n-1}^*)|} \geq \theta_{fb}.$$

   Finally, by (21.12) and the above selection procedure, we have

   $$\begin{aligned} \frac{1}{M} &\geq \frac{|E_n(y_1^*, \ldots, y_n^*)|}{|E_0|} \geq \theta_{fb}^n \\ \Rightarrow M &\leq -n \log \theta_{fb} \\ \Rightarrow C_{fb,0} &\leq -\log \theta_{fb} \end{aligned}$$

4. "≥" (achievability)

   Let's construct a code that achieves $(M, n, 0)$.



The above example with $|\mathcal{A}| = 3$ illustrates that the encoder f_1 partitions the space of all messages to 3 groups. The encoder f_1 at the first stage encodes the groups of messages into a_1, a_2, a_3 correspondingly. When channel outputs y_1 and assume that $S_{y_1} = \{a_1, a_2\}$, then the decoder can eliminate a total number of $MP_X^*(a_3)$ candidate messages in this round. The "confusability set" only contains the remaining $MP_X^*(S_{y_1})$ messages. By definition of P_X^* we know that $MP_X^*(S_{y_1}) \le M\theta_{fb}$. In the second round, f_2 partitions the remaining messages into three groups, send the group index and repeat.

By similar arguments, each interaction reduces the uncertainty by a factor of *at least* θ_{fb}. After n iterations, the size of "confusability set" is upper bounded by $M\theta_{fb}^n$, if $M\theta_{fb}^n \le 1$,[2] then zero error probability is achieved. This is guaranteed by choosing $\log M = -n \log \theta_{fb}$. Therefore we have shown that $-n \log \theta_{fb}$ bits can be reliably delivered with $n + O(1)$ channel uses with feedback, thus

$$C_{fb,0} \ge -\log \theta_{fb}$$

□

21.3.2 Code with variable length

Consider the example of BEC(δ) with feedback, send k bits in the following way: repeat sending each bit until it gets through the channel correctly. The expected number of channel uses for sending k bits is given by

$$l = \mathbb{E}[n] = \frac{k}{1-\delta}$$

We state the result for **variable-length feedback** (VLF) code without proof:

$$\log M_{VLF}^*(l, 0) \ge lC$$

Notice that compared to the scheme without feedback, there is the improvement of $\sqrt{nV}Q^{-1}(\epsilon)$ in the order of $O(\sqrt{n})$, which is stronger than the result in Theorem 21.2.

This is also true in general:

$$\log M_{VLF}^*(l, \epsilon) = \frac{lC}{1-\epsilon} + O(\log l)$$

Example: For BSC(0.11), without feedback, $n = 3000$ is needed to achieve 90% of capacity C, while with VLF code $l = \mathbb{E}n = 200$ is enough to achieve that.

21.3.3 Code with variable power

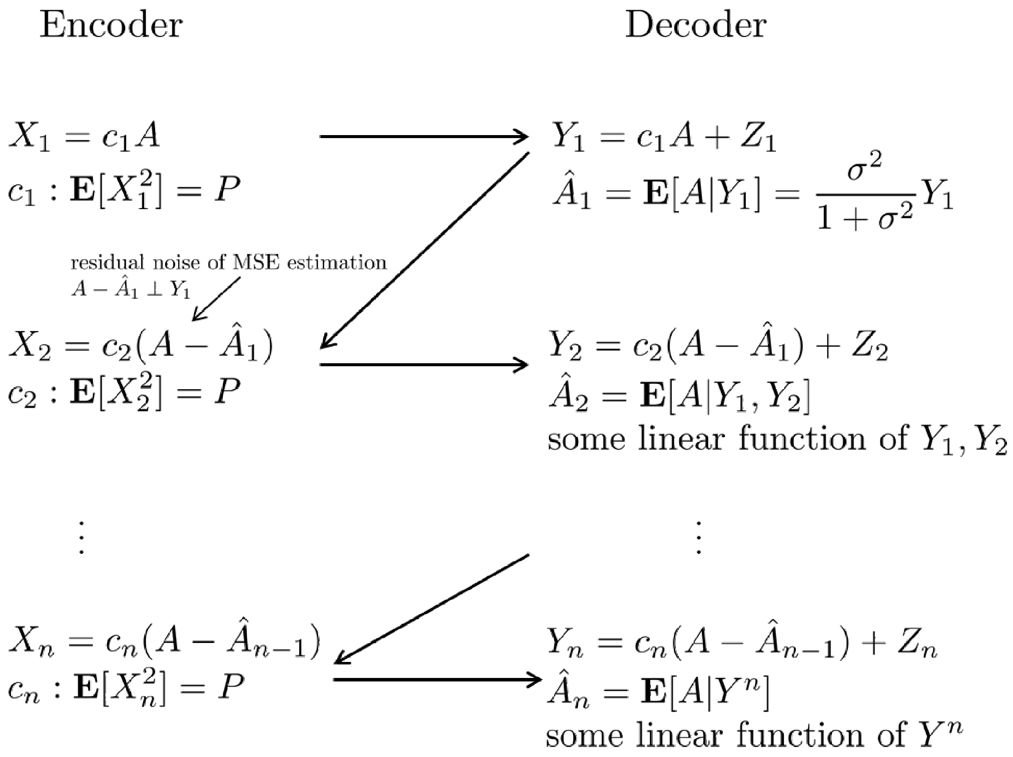
Elias' scheme of sending a number A drawn from a Gaussian distribution $\mathcal{N}(0, \operatorname{Var} A)$ with linear processing.

AWGN setup:

$$Y_k = X_k + Z_k, \quad Z_k \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$
$$\mathbb{E}[X_k^2] \le P, \quad \text{power constraint in expectation}$$

Note: If we insist the codeword satisfies power constraint almost surely instead on average, i.e., $\sum_{k=1}^n X_k^2 \le nP$ a.s., then the scheme below does not work!

[2] Some rounding-off errors need to be corrected in a few final steps (because P_X^* may not be closely approximable when very few messages are remaining). This does not change the asymptotics though.


Encoder
Decoder
$X_1 = c_1 A$
$c_1 : \mathbf{E}[X_1^2] = P$
$Y_1 = c_1 A + Z_1$
$\hat{A}_1 = \mathbf{E}[A|Y_1] = \frac{\sigma^2}{1+\sigma^2} Y_1$
residual noise of MSE estimation
$A - \hat{A}_1 \perp Y_1$
$X_2 = c_2(A - \hat{A}_1)$
$c_2 : \mathbf{E}[X_2^2] = P$
$Y_2 = c_2(A - \hat{A}_1) + Z_2$
$\hat{A}_2 = \mathbf{E}[A|Y_1, Y_2]$
some linear function of Y_1, Y_2
$X_n = c_n(A - \hat{A}_{n-1})$
$c_n : \mathbf{E}[X_n^2] = P$
$Y_n = c_n(A - \hat{A}_{n-1}) + Z_n$
$\hat{A}_n = \mathbf{E}[A|Y^n]$
some linear function of Y^n

According to the orthogonality principle of the mininum mean-square estimation (MMSE) of A at receiver side in every step:

$$A = \hat{A}_n + N_n, \quad N_n \perp Y^n.$$

Morever, since all operations are lienar and everything is jointly Gaussian, $N_n \perp\!\!\!\perp Y^n$. Since $X_n \propto N_{n-1} \perp\!\!\!\perp Y^{n-1}$, the codeword we are sending at each time slot is independent of the history of the channel output ("innovation"), in order to maximize information transfer.

Note that $Y^n \to \hat{A}_n \to A$, and the optimal estimator $\hat{A}_n$ (a linear combination of Y^n) is a sufficient statistic of Y^n for A under Gaussianity. Then

$$\begin{aligned} I(A;Y^n) =& I(A;\hat{A}_n, Y^n) \\ &= I(A;\hat{A}_n) + I(A;Y^n|\hat{A}_n) \\ &= I(A;\hat{A}_n) \\ &= \frac{1}{2}\log\frac{\mathrm{Var}(A)}{\mathrm{Var}(N_n)}. \end{aligned}$$

where the last equality uses the fact that N follows a normal distribution. $\mathrm{Var}(N_n)$ can be computed directly using standard linear MMSE results. Instead, we determine it information theoretically: Notice that we also have

$$\begin{aligned} I(A;Y^n) &= I(A;Y_1) + I(A;Y_2|Y_1) + \cdots + I(A;Y_n|Y^{n-1}) \\ &= I(X_1;Y_1) + I(X_2;Y_2|Y_1) + \cdots + I(X_n;Y_n|Y^{n-1}) \\ &\overset{\text{key}}{=} I(X_1;Y_1) + I(X_2;Y_2) + \cdots + I(X_n;Y_n) \\ &= n\frac{1}{2}\log(1+P) = nC \end{aligned}$$

Therefore, with Elias' scheme of sending $A \sim \mathcal{N}(0, \operatorname{Var} A)$, after the n-th use of the AWGN(P) channel with feedback,

$$\operatorname{Var} N_n = \operatorname{Var}(\hat{A}_n - A) = 2^{-2nC} \operatorname{Var} A = \left(\frac{P}{P+\sigma^2}\right)^n \operatorname{Var} A,$$

which says that the reduction of uncertainty in the estimation is exponential fast in n.

Schalkwijk-Kailath: Elias' scheme can also be used to send digital data.

Let $W \sim$ uniform on M-PAM constellation in $\in [-1,1]$, i.e., $\{-1, -1+\frac{2}{M}, \cdots, -1+\frac{2k}{M}, \cdots, 1\}$. In the very first step W is sent (after scaling to satisfy the power constraint):

$$X_0 = \sqrt{P}W, \quad Y_0 = X_0 + Z_0$$

Since Y_0 and X_0 are both known at the encoder, it can compute Z_0. Hence, to describe W it is sufficient for the encoder to describe the noise realization Z_0. This is done by employing the Elias' scheme ($n-1$ times). After $n-1$ channel uses, and the MSE estimation, the equivalent channel output:

$$\widetilde{Y}_0 = X_0 + \widetilde{Z}_0, \quad \operatorname{Var}(\widetilde{Z}_0) = 2^{-2(n-1)C}$$

Finally, the decoder quantizes $\widetilde{Y}_0$ to the nearest PAM point. Notice that

$$\begin{aligned}
\epsilon \le \mathbb{P}\left[|\widetilde{Z}_0| > \frac{1}{2M}\right] = \mathbb{P}\left[2^{-(n-1)C}|Z| > \frac{\sqrt{P}}{2M}\right] = 2Q\left(\frac{2^{(n-1)C}\sqrt{P}}{2M}\right)\\
\Rightarrow \log M \ge (n-1)C + \log\frac{\sqrt{P}}{2} - \log Q^{-1}\left(\frac{\epsilon}{2}\right)\\
= nC + O(1).
\end{aligned}$$

Hence if the rate is strictly less than capacity, the error probability decays doubly exponentially fast as n increases. More importantly, we gained an $\sqrt{n}$ term in terms of $\log M$, since for the case without feedback we have

$$\log M^*(n,\epsilon) = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)\,.$$

Example: $P = 1 \;\Rightarrow$ channel capacity $C = 0.5$ bit per channel use. To achieve error probability 10^{-3}, $2Q\left(\frac{2^{(n-1)C}}{2M}\right) \approx 10^{-3}$, so $\frac{e^{(n-1)C}}{2M} \approx 3$, and $\frac{\log M}{n} \approx \frac{n-1}{n}C - \frac{\log 8}{n}$. Notice that the capacity is achieved to within 99% in as few as $n = 50$ channel uses, whereas the best possible block codes without feedback require $n \approx 2800$ to achieve 90% of capacity.

Take-away message:

Feedback is best harnessed with *adaptive* strategies. Although it does not increase capacity under block coding, feedback greatly boosts reliability as well as reduces coding complexity.

§ 22. Capacity-achieving codes via Forney concatenation

Shannon's Noisy Channel Theorem assures us the existence of capacity-achieving codes. However, exhaustive search for the code has double-exponential complexity: Search over all codebook of size 2^{nR} over all possible $|\mathcal{X}|^n$ codewords.

Plan for today: Constructive version of Shannon's Noisy Channel Theorem. The goal is to show that for BSC, it is possible to achieve capacity in polynomial time. Note that we need to consider three aspects of complexity

- Encoding
- Decoding
- Construction of the codes

22.1 Error exponents

Recall we have defined the fundamental limit

$$M^*(n,\epsilon) = \max\{M : \exists (n,M,\epsilon)\text{-code}\}$$

For notational convenience, let us define its functional inverse

$$\epsilon^*(n,M) = \inf\{\epsilon : \exists (n,M,\epsilon)\text{-code}\}$$

Shannon's theorem shows that for stationary memoryless channels, $\epsilon_n \triangleq \epsilon^*(n,\exp(nR)) \to 0$ for any $R < C = \sup_X I(X;Y)$. Now we want to know how fast it goes to zero as $n \to \infty$. It turns out the speed is exponential, i.e., $\epsilon_n \approx \exp(-nE(R))$ for some error exponent $E(R)$ as a function R, which is also known as the reliability function of the channel. Determining $E(R)$ is one of the most long-standing open problems in information theory. What we know are

- Lower bound on $E(R)$ (achievability): Gallager's random coding bound (which analyzes the ML decoder, instead of the suboptimal decoder as in Shannon's random coding bound or DT bound).
- Upper bound on $E(R)$ (converse): Sphere-packing bound (Shannon-Gallager-Berlekamp), etc.

It turns out there exists a number $R_{\text{crit}} \in (0,C)$, called the critical rate, such that the lower and upper bounds meet for all $R \in (R_{\text{crit}}, C)$, where we obtain the value of $E(R)$. For $R \in (0, R_{\text{crit}})$, we do not even know the existence of the exponent!

Deriving these bounds is outside the scope of this lecture. Instead, we only need the *positivity* of error exponent, i.e., for any $R < C$, $E(R) > 0$. On the other hand, it is easy to see that $E(C-) = 0$ as a consequence of weak converse. Since as the rate approaches capacity from below, the communication becomes less reliable. The next theorem is a simple application of large deviation.

Theorem 22.1. *For any DMC, for any $R < C = \sup_X I(X;Y)$,*

$$\epsilon^*(n, \exp(nR)) \le \exp(-nE(R)), \quad \text{for some } E(R) > 0.$$

Proof. Fix $R < C$ so that $C - R > 0$. Let P_X^* be the capacity-achieving input distribution, i.e., $C = I(X^*;Y^*)$. Recall Shannon's random coding bound (DT/Feinstein work as well):

$$\epsilon \le P(i(X;Y) \le \log M + \tau) + \exp(-\tau).$$

As usual, we apply this bound with iid $P_{X^n} = (P_X^*)^n$, $\log M = nR$ and $\tau = \frac{n(C-R)}{2}$, to conclude the achievability of

$$\epsilon_n \le P\left(\frac{1}{n} i(X^n;Y^n) \le \frac{C+R}{2}\right) + \exp\left(-\frac{n(C-R)}{2}\right).$$

Since $i(X^n;Y^n) = \sum i(X_k;Y_k)$ is an iid sum, and $\mathbb{E}i(X;Y) = C > (C+R)/2$, the first term is upper bounded by $\exp(-n\psi_T^*(\frac{R+C}{2}))$ where $T = i(X;Y)$. The proof is complete since ϵ_n is smaller than the sum of two exponentially small terms. □

Note: Better bound can be obtained using DT bound. But to get the best lower bound on $E(R)$ we know (Gallager's random coding bound), we have to analyze the ML decoder.

22.2 Achieving polynomially small error probability

In the sequel we focus on BSC channel with cross-over probability δ, which is an additive-noise DMC. Fix $R < C = 1 - h(\delta)$ `bits`. Let the block length be n. Our goal is to achieve error probability $\epsilon_n \le n^{-\alpha}$ for arbitrarily large $\alpha > 0$ in polynomial time.

To this end, fix some $b > 1$ to be specified later and pick $m = b \log n$ and divide the block into $\frac{n}{m}$ sub-blocks of m bits. Applying Theorem 22.1, we can find [later on how to find] an $(m, \exp(Rm), \epsilon_m)$-code such that

$$\epsilon_m \le \exp(-mE(R)) = n^{-bE(R)}$$

where $E(R) > 0$. Apply this code to each m-bit sub-block and apply ML decoding to each block. The encoding/decoding complexity is at most $\frac{n}{m}\exp(O(m)) = n^{O(1)}$. To analyze the probability of error, use union bound:

$$P_e \le \frac{n}{m}\epsilon_m \le n^{-bE(R)+1} \le n^{-\alpha},$$

if we choose $b \ge \frac{\alpha+1}{E(R)}$.

Remark 22.1. The final question boils down to how to find the shorter code of blocklength m in poly(n)-time. This will be done if we can show that we can find good code (satisfying the Shannon random coding bound) for BSC of blocklenth m in exponential time. To this end, let us go through the following strategies:

1. Exhaustive search: A codebook is a subset of cardinality 2^{Rm} out of 2^m possible codewords. Total number of codebooks: $\binom{2^m}{2^{Rm}} = \exp(\Omega(m2^{Rm})) = \exp(\Omega(n^c \log n))$. The search space is too big.

2. Linear codes: In Lecture 16 we have shown that for additive-noise channels on finite fields we can focus on linear codes. For BSC, each linear code is parameterized by a generator matrix, with Rm^2 entries. Then there are a total of $2^{Rm^2} = n^{\Omega(\log n)}$ – still superpolynomial and we cannot afford the search over all linear codes.

3. Toeplitz generator matrices: In Homework 8 we see that it does not lose generality to focus on linear codes with **Toeplitz** generator matrices, i.e., G such that $G_{ij} = G_{i-1,j-1}$ for all $i, j > 1$. Toeplitz matrices are determined by diagonals. So there are at most $2^{2m} = n^{O(1)}$ and we can find the optimal one in poly(n)-time.

Since the channel is additive-noise, linear codes + syndrome decoder leads to the same maximal probability of error as average (Lecture 16).

Remark 22.2. Remark on de-randomization; randomness as a resource, coin flips and cooking (brown both sides of onions)...

22.3 Concatenated codes

Forney introduced the idea of concatenated codes in 1965 to build longer codes from shorter codes with manageable complexity. It consists of an inner code and an outer code:

1. $C_{\text{in}} : \{0,1\}^k \to \{0,1\}^n$, with rate $\frac{k}{n}$

2. $C_{\text{out}} : B^K \to B^N$ for some alphabet B of cardinality 2^k, with rate $\frac{K}{N}$.

The concatenated code $C : \{0,1\}^{kK} \to \{0,1\}^{nN}$ works as follows (Fig. 22.1):

1. Collect the kK message bits into K symbols in the alphabet B, apply C_{out} componentwise to get a vector in B^N

2. Map each symbol in B into k bits and apply C_{in} componentwise to get a nN-bit codeword.

The rate of the concatenated code is the product of the rates of the inner and outer codes: $R = \frac{k}{n}\frac{K}{N}$.

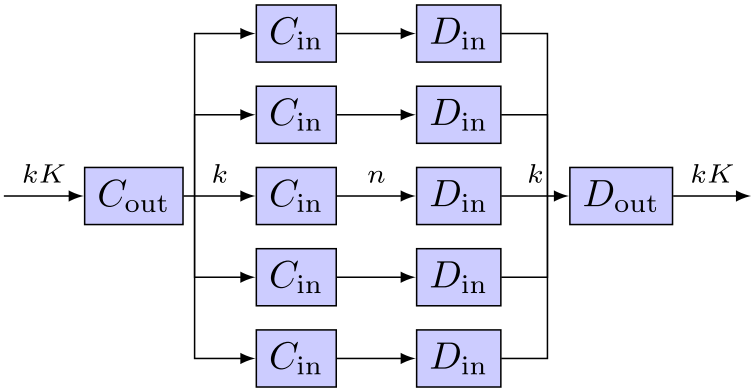


Figure 22.1: Concatenated code, where there are N inner encoder-decoder pairs.

22.4 Achieving exponentially small error probability

Forney proposed the following idea:

- Use an optimal code as the inner code
- Use a Reed-Solomon code as the outer code which can correct a constant fraction of errors.

Reed-Solomon (RS) codes are linear codes from $\mathbb{F}_q^K \to \mathbb{F}_q^N$ where the block length $N = q - 1$ and the message length is K. Similar to the Reed-Muller code, the RS code treats the input $(a_0, a_1, \dots, a_{K-1})$ as a polynomial $p(x) = \sum_{i=0}^{K-1} a_i z^i$ over $\mathbb{F}_q$ of degree at most $K - 1$, and encodes it by its values at all non-zero elements. Therefore the RS codeword is a vector $(p(\alpha) : \alpha \in \mathbb{F}_q \backslash \{0\}) \in \mathbb{F}_q^N$. Therefore the generator matrix of RS code is a Vandermonde matrix.

The RS code has the following advantages:

1. The minimum distance of RS code $N - K + 1$. So if we choose $K = (1 - \epsilon)N$, then RS code can correct $\frac{\epsilon N}{2}$ errors.
2. The encoding and decoding (e.g., Berlekamp-Massey decoding algorithm) can be implemented in poly(N) time.

In fact, as we will see later, any efficient code which can correct a constant fraction of errors will suffice as the outer code for our purpose.

Now we show that we can achieve any rate below capacity and exponentially small probability of error in polynomial time: Fix $\eta, \epsilon > 0$ arbitrary.

- Inner code: Let $k = (1 - h(\delta) - \eta)n$. By Theorem 22.1, there exists a $C_{\text{in}} : \{0,1\}^k \to \{0,1\}^n$, which is a linear $(n, 2^k, \epsilon_n)$-code and *maximal* error probability $\epsilon_n \le 2^{-nE(\eta)}$. By Remark 22.1, C_{in} can be chosen to be a linear code with Toeplitz generator matrix, which can be found in 2^n time. The inner decoder is ML, which we can afford since n is small.
- Outer code: We pick the RS code with field size $q = 2^k$ with blocklength $N = 2^k - 1$. Pick the number of message bits to be $K = (1 - \epsilon)N$. Then we have $C_{\text{out}} : \mathbb{F}_{2^k}^K \to \mathbb{F}_{2^k}^N$.

Then we obtain a concatenated code $C : \{0,1\}^{kK} \to \{0,1\}^{nN}$ with blocklength $L = nN = n2^{Cn}$ for some constant C and rate $R = (1 - \epsilon)(1 - h(\delta) - \eta)$. It is clear that the code can be constructed in $2^n = \text{poly}(L)$ time and all encoding/decoding operations are poly(L) time.

Now we analyze the probability of error: Let us conditioned on the message bits (input to C_{out}). Since the outer code can correct $\frac{\epsilon N}{2}$ errors, an error happens only if the number of erroneous inner encoder-decoder pairs exceeds $\frac{\epsilon N}{2}$. Since the channel is memoryless, each of the N pairs makes an error independently[1] with probability at most ϵ_n. Therefore the number of errors is stochastically smaller than Binom(N, ϵ_n), and we can upper bound the total probability of error using Chernoff bound:

$$P_e \le \mathbb{P}\left[\text{Binom}(N, \epsilon_n) \ge \frac{\epsilon N}{2}\right] \le \exp\left(-N d(\epsilon/2 \| \epsilon_n)\right) = \exp\left(-\Omega(N \log N)\right) = \exp(-\Omega(L)).$$

where we have used $\epsilon_n \le \exp(-\Omega(n))$ and $d(\epsilon/2\|\epsilon_n) \ge \frac{\epsilon}{2} \log \frac{\epsilon}{2\epsilon_n} = \Omega(n) = \Omega(\log N)$.

[1] Here controlling the *maximal* error probability of inner code is the key. If we only have average error probability, then given a uniform distributed input to the RS code, the output symbols (which are the inputs to the inner encoders) need *not* be independent, and Chernoff bound is not necessarily applicable.

Note: For more details see the excellent exposition by Spielman [Spi97]. For modern constructions using sparse graph codes which achieve the same goal in *linear* time, see, e.g., [Spi96].

Part V

Lossy data compression

§ 23. Rate-distortion theory

Big picture so far:

1. Lossless data compression: Given a discrete ergodic source S^k, we know how to encode to pure bits $W \in [2^k]$.

2. Binary HT: Given two distribution P and Q, we know how to distinguish them optimally.

3. Channel coding: How to send bits over a channel $[2^k] \ni W \to X \to Y$.

4. JSCC: how to send discrete data optimally over a noisy channel.

Next topic, lossy data compression: Given X, find a k-bit representation W, $X \to W \to \hat{X}$, such that $\hat{X}$ is a good reconstruction of X.

Real-world examples: codecs consist of a compressor and a decompressor

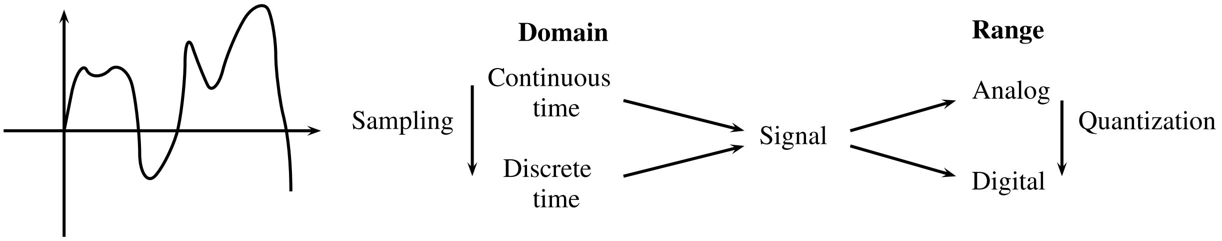
- Image: JPEG...

- Audio: MP3, CD...

- Video: MPEG...

23.1 Scalar quantization

Problem: Data isn't discrete! Often, a signal (function) comes from voltage levels or other continuous quantities. The question of how to map (naturally occurring) continuous time/analog signals into (electronics friendly) discrete/digital signals is known as *quantization*, or in information theory, as *rate distortion theory*.

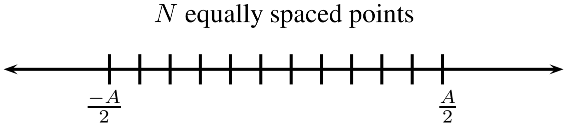


We will look at several ways to do quantization in the next few sections.

23.1.1 Scalar Uniform Quantization

The idea of qunatizing an inherently continuous-valued signal was most explicitly expounded in the patenting of Pulse-Coded Modulation (PCM) by A. Reeves, cf. [Ree65] for some interesting historical notes. His argument was that unlike AM and FM modulation, quantized (digital) signals could be sent over long routes without the detrimental accumulation of noise. Some initial theoretical analysis of the PCM was undertaken in 1947 by Oliver, Pierce, and Shannon (same Shannon), cf. [OPS48].

For a random variable $X \in [-A/2, A/2] \subset \mathbb{R}$, the scalar uniform quantizer $q_U(X)$ with N quantization points partitions the interval $[-A/2, A/2]$ uniformly

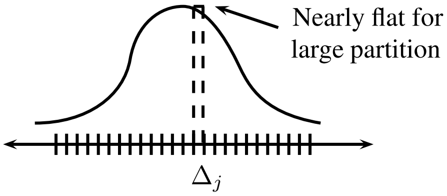


where the points are in $\{\frac{-A}{2} + \frac{kA}{N}, k = 0, \dots, N-1\}$.

What is the *quality* (or fidelity) of this quantization? Most of the time, mean squared error is used as the quality criterion:

$$D(N) = \mathbb{E}|X - q_U(X)|^2$$

where D denotes the average *distortion*. Often $R = \log_2 N$ is used instead of N, so that we think about the number of bits we can use for quantization instead of the number of points. To analyze this scalar uniform quantizer, we'll look at the high-rate regime ($R \gg 1$). The key idea in the high rate regime is that (assuming a smooth density P_X), each quantization interval Δ_j looks nearly flat, so conditioned on Δ_j, the distribution is accurately approximately by a uniform distribution.



Let c_j be the j-th quantization point, and Δ_j be the j-th quantization interval. Here we have

$$\begin{aligned}
\mathbb{E}|X - q_U(X)|^2 &= \sum_{j=1}^{N} \mathbb{E}[|X - c_j|^2 | X \in \Delta_j] \mathbb{P}[X \in \Delta_j] \\
\text{(high rate approximation)} \quad &\approx \sum_{j=1}^{N} \frac{|\Delta_j|^2}{12} \mathbb{P}[X \in \Delta_j] \\
&= \frac{(\frac{A}{N})^2}{12} = \frac{A^2}{12} 2^{-2R}
\end{aligned}$$

How much do we gain per bit?

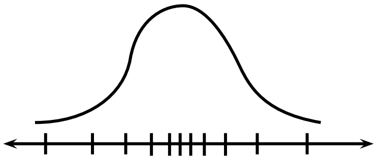
$$\begin{aligned}
10 \log_{10} SNR &= 10 \log_{10} \frac{Var(X)}{\mathbb{E}|X - q_U(X)|^2} \\
&= 10 \log_{10} \frac{12 Var(X)}{A^2} + (20 \log_{10} 2) R \\
&= \text{constant} + (6.02 dB) R
\end{aligned}$$

For example, when X is uniform on $[-\frac{A}{2}, \frac{A}{2}]$, the constant is 0. Every engineer knows the rule of thumb "6dB per bit"; adding one more quantization bit gets you 6 dB improvement in SNR. However, here we can see that this rule of thumb is valid only in the high rate regime. (Consequently, widely articulated claims such as "16-bit PCM (CD-quality) provides 96 dB of SNR" should be taken with a grain of salt.)

Note: The above deals with X with a bounded support. When X is unbounded, a wise thing to do is to allocate the quantization points to the range of values that are more likely and saturate the large values at the dynamic range of the quantizer. Then there are two contributions, known as the granular distortion and overload distortion. This leads us to the question: Perhaps instead of uniform quantization optimal?

23.1.2 Scalar Non-uniform Quantization

Since our source has density p_X, a good idea might be to use more quantization points where p_X is larger, and less where p_X is smaller.

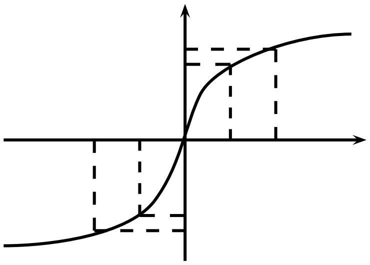


Often the way such quantizers are implemented is to take a monotone transformation of the source $f(X)$, perform uniform quantization, then take the inverse function:

$$\begin{tikzcd} X \arrow[r, "f"] \arrow[d, dashed, "q"] & U \arrow[d, "q_U"] \\ \hat{X} & q_U(U) \arrow[l, "f^{-1}"'] \end{tikzcd} \tag{23.1}$$

i.e., $q(X) = f^{-1}(q_U(f(X)))$. The function f is usually called the *compander* (compressor+expander). One of the choice of f is the CDF of X, which maps X into uniform on $[0,1]$. In fact, this compander architecture is optimal in the high-rate regime (fine quantization) but the optimal f is not the CDF (!). We defer this discussion till Section 23.1.4.

In terms of practical considerations, for example, the human ear can detect sounds with volume as small as 0 dB, and a painful, ear-damaging sound occurs around 140 dB. Achieveing this is possible because the human ear inherently uses logarithmic companding function. Furthermore, many natural signals (such as *differences* of consecutive samples in speech or music (but not samples themselves!)) have an approximately Laplace distribution. Due to these two factors, a very popular and sensible choice for f is the μ-companding function



$$f(X) = \text{sign}(X)\frac{\ln(1+\mu|X|)}{\ln(1+\mu)}$$

which compresses the dynamic range, uses more bits for smaller $|X|$'s, e.g. $|X|$'s in the range of human hearing, and less quantization bits outside this region. This results in the so-called μ-law which is used in the digital telecommunication systems in the US, while in Europe they use a slightly different compander called the A-law.

23.1.3 Optimal Scalar Quantizers

Now we look for the optimal scalar quantizer given R bits for reconstruction. Formally, this is

$$D_{scalar}(R) = \min_{q:|\text{Im } q|\le 2^R} \mathbb{E}|X - q(X)|^2$$

Intuitively, we would think that the optimal quantization regions should be contiguous; otherwise, given a point c_j, our reconstruction error will be larger. Therefore quantizers are piecewise constant:

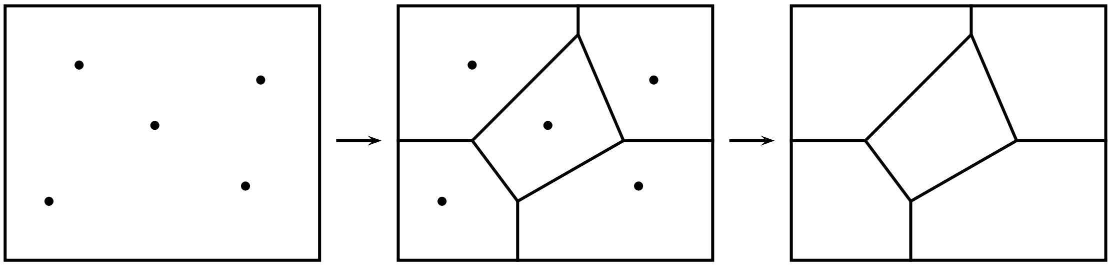
$$q(x) = c_j \mathbf{1}_{T_j \le x \le T_{j+1}}$$

for some $c_j \in [T_j, T_{j+1}]$.

Simple example: One-bit quantization of $X \sim \mathcal{N}(0,\sigma^2)$. Then optimal quantization points are $c_1 = \mathbb{E}[X|X \ge 0] = \sqrt{\frac{2}{\pi}}\sigma$, $c_2 = \mathbb{E}[X|X \le 0] = -\sqrt{\frac{2}{\pi}}\sigma$.

With ideas like this, in 1982 Stuart Lloyd developed an algorithm (called *Lloyd's algorithm*) for iteratively finding optimal quantization regions and points. This works for both the scalar and vector cases, and goes as follows:

1. Pick any $N = 2^k$ points
2. Draw the Voronoi regions around the chosen quantization points (aka minimum distance tessellation, or set of points closest to c_j), which forms a partition of the space.
3. Update the quantization points by the centroids ($\mathbb{E}[X|X \in D]$) of each Voronoi region.
4. Repeat.



Steps of Lloyd's algorithm

Lloyd's clever observation is that the centroid of each Voronoi region is (in general) different than the original quantization points. Therefore, iterating through this procedure gives the *Centroidal Voronoi Tessellation* (CVT - which are very beautiful objects in their own right), which can be viewed as the fixed point of this iterative mapping. The following theorem gives the results about Lloyd's algorithm

Theorem 23.1 (Lloyd)**.**

1. *Lloyd's algorithm always converges to a Centroidal Voronoi Tessellation.*

2. *The optimal quantization strategy is always a CVT.*

3. *CVT's are non-unique, and the algorithm may converge to non-global optima.*

Remark: The third point tells us that Lloyd's algorithm isn't always guaranteed to give the optimal quantization strategy.[1] One sufficient condition for uniqueness of a CVT is the log-concavity of the density of X [Fleischer '64]. Thus, for Gaussian P_X, Lloyd's algorithm outputs the optimal quantizer, but even for Gaussian, if $N > 3$, optimal quantization points are not known in closed form! So it's hard to say too much about optimal quantizers. Because of this, we next look for an approximation in the regime of huge number of points.

23.1.4 Fine quantization

[Panter-Dite '51] Now we look at the high SNR approximation. For this, introduce the probability density function $\lambda(x)$, which represents the density of our quantization points and allows us to approximate summations by integrals[2]. Then the number of quantization points in any interval $[a,b]$ is $\approx N\int_a^b \lambda(x)dx$. For any point x, denote its distance to the closest quantization point by $\Delta(x)$. Then $N\lambda(x)\Delta(x) \approx 1 \implies \Delta(x) \approx \frac{1}{N\lambda(x)}$. With this approximation, the quality of reconstruction is

$$\begin{aligned}\mathbb{E}|X-q(X)|^2 &= \sum_{j=1}^{N} \mathbb{E}[|X-c_j|^2 | X \in \Delta_j]\mathbb{P}[X \in \Delta_j] \\ &\approx \sum_{j=1}^{N} \mathbb{P}[X \in \Delta_j]\frac{|\Delta_j|^2}{12} \approx \int p(x)\frac{\Delta^2(x)}{12}dx \\ &= \frac{1}{12N^2}\int p(x)\lambda^{-2}(x)dx\end{aligned}$$

To find the optimal density λ that gives the best reconstruction (minimum MSE) when X has density p, we use Hölder's inequality: $\int p^{1/3} \le (\int p\lambda^{-2})^{1/3}(\int \lambda)^{2/3}$. Therefore $\int p\lambda^{-2} \ge (\int p^{1/3})^3$, with equality iff $p\lambda^{-2} \propto \lambda$. Hence the optimizer is $\lambda^*(x) = \frac{f^{1/3}(x)}{\int f^{1/3}dx}$. Therefore when $N = 2^R$,[3]

$$D_{scalar}(R) \approx \frac{1}{12}2^{-2R}\left(\int p^{1/3}(x)dx\right)^3$$

So our optimal quantizer density in the high rate regime is proportional to the cubic root of the density of our source. This approximation is called the *Panter-Dite approximation*. For example, when $X \sim \mathcal{N}(0,\sigma^2)$, this gives

$$D_{scalar}(R) \approx \sigma^2 2^{-2R}\frac{\pi\sqrt{3}}{2}$$

Note: In fact, in *scalar* case the optimal non-uniform quantizer can be realized using the compander architecture (23.1) that we discussed in Section 23.1.2: As an exercise, use Taylor expansion to

[1] As a simple example one may consider $P_X = \frac{1}{3}\phi(x-1) + \frac{1}{3}\phi(x) + \frac{1}{3}\phi(x+1)$ where $\phi(\cdot)$ is a very narrow pdf, symmetric around 0. Here the CVT with centers $\pm\frac{2}{3}$ is not optimal among binary quantizers (just compare to any quantizer that quantizes two adjacent spikes to same value).

[2] This argument is easy to make rigorous. We only need to define reconstruction points c_j as solutions of

$$\int_{-\infty}^{c_j} \lambda(x)\,dx = \frac{j}{N}\,.$$

[3] In fact when $R \to \infty$, "$\approx$" can be replaced by "$= 1 + o(1)$" [Zador '56].

analyze the quantization error of (23.1) when $N \to \infty$. The optimal compander $f : \mathbb{R} \to [0,1]$ turns out to be $f(x) = \frac{\int_{-\infty}^{t} p^{1/3}(t)dt}{\int_{-\infty}^{\infty} p^{1/3}(t)dt}$ [Bennett '48, Smith '57].

23.1.5 Fine quantization and variable rate

So far we were considering quantization with restriction on the cardinality of the image of $q(\cdot)$. If one, however, intends to further compress the values $q(X)$ via noiseless compressor, a more natural constraint is to bound $H(q(X))$.

Koshelev [Kos63] discovered in 1963 that in the high rate regime uniform quantization is asymptotically optimal under the entropy constraint. Indeed, if q_Δ is a uniform quantizer with cell size Δ, then it is easy to see that

$$H(q_\Delta(X)) = h(X) - \log \Delta + o(1)\,, \tag{23.2}$$

where $h(X) = -\int p_X(x) \log p_X(x)\, dx$ is the differential entropy of X. So a uniform quantizer with $H(q(X)) = R$ achieves

$$D = \frac{\Delta^2}{12} \approx 2^{-2R} \frac{2^{2h(X)}}{12}\,.$$

On the other hand, any quantizer with unnormalized point density function $\Lambda(x)$ (i.e. smooth function such that $\int_{-\infty}^{c_j} \Lambda(x) dx = j$) can be shown to achieve (assuming $\Lambda \to \infty$ pointwise)

$$D \approx \frac{1}{12} \int p_X(x) \frac{1}{\Lambda^2(x)} dx \tag{23.3}$$

$$H(q(X)) \approx \int p_X(x) \log \frac{\Lambda(x)}{p_X(x)}\, dx \tag{23.4}$$

Now, from Jensen's inequality we have

$$\frac{1}{12} \int p_X(x) \frac{1}{\Lambda^2(x)} dx \ge \frac{1}{12} \exp\{-2 \int p_X(x) \log \Lambda(x)\, dx\} \approx 2^{-2H(q(X))} \frac{2^{2h(X)}}{12}\,,$$

concluding that uniform quantizer is asymptotically optimal.

Furthermore, it turns out that for any source, even the optimal vector quantizers (to be considered next) can not achieve distortion better that $2^{-2R} \frac{2^{2h(X)}}{2\pi e}$ – i.e. the maximal improvement they can gain (on any iid source!) is 1.53 dB (or 0.255 bit/sample). This is one reason why scalar uniform quantizers followed by lossless compression is an overwhelmingly popular solution in practice.

23.2 Information-theoretic vector quantization

By doing vector quantization (namely, compressing $(X_1, \ldots, X_n) \to 2^{nR}$ points), rate-distortion theory tells us that when n is large, we can achieve the per-coordinate MSE:

$$D_{vec}(R) = \sigma^2 2^{-2R}$$

which saves 4.35 dB (or 0.72 bit/sample). This should be rather surprising, so we repeat it again: even when $X_1, \ldots, X_n$ are iid, we can get better performance by quantizing X_i jointly. One instance of this surprising effect is the following:

Hamming Game: Given 100 unbiased bits, we want to look at them and scribble something down on a piece of paper that can store 50 bits at most. Later we will be asked to guess the original 100 bits, with the goal of maximizing the number of correctly guessed bits. What is the best strategy? Intuitvely, the optimal strategy would be to store half of the bits then guess on the rest, which gives 25% BER. However, as we will show in the next few lectures, the optimal strategy amazingly achieves a BER of 11%. Note does this happen? After all we are guessing independent bits and the utility function (BER) treats all bits equally. Some intuitive explanation:

1. Applying scalar quantization componentwise results in quantization region that are hypercubes, which might not be efficient for covering.
2. Concentration of measures removes many source realizations that are highly unlikely. For example, if we think about quantizing a single Gaussian X, then we need to cover large portion of $\mathbb{R}$ in order to cover the cases of significant deviations of X from 0. However, when we are quantizing many $(X_1, \dots, X_n)$ together, the law of large numbers makes sure that many X_j's cannot conspire together and all produce large values. Thus, we may exclude large portions of the $\mathbb{R}^n$ from consideration.

Math Formalism: A lossy compressor is an encoder/decoder pair (f, g) where

$$X \xrightarrow{f} W \xrightarrow{g} \hat{X}$$

- $X \in \mathcal{X}$ - continuous source
- W - discrete data
- $\hat{X} \in \hat{\mathcal{X}}$ - reproduction

A *distortion metric* is a function $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R} \cup \{+\infty\}$ (loss function). There are various formulations of the lossy compression problem:

1. Fixed length (fixed rate), average distortion: $W \in [M]$, minimize $\mathbb{E}[d(X, \hat{X})]$.
2. Fixed length, excess distortion: $W \in [M]$, minimize $\mathbb{P}[d(X, \hat{X}) > D]$.
3. Variable length, max distortion: $W \in \{0,1\}^*$, $d(X, \hat{X}) \le D$ a.s., minimize $\mathbb{E}[\text{length}(W)]$ or $H(\hat{X}) = H(W)$.

Note: In this course we focus on fixed length and average distortion loss compression. The difference between average distortion and excess distortion is analogous to average risk bound and high-probability bound in statistics/machine learning.

Definition 23.1. Rate-distortion problem is characterized by a pair of alphabets $\mathcal{A}$, $\hat{\mathcal{A}}$, a single-letter distortion function $d(\cdot, \cdot) : \mathcal{A} \times \hat{\mathcal{A}} \to \mathbb{R} \cup \{+\infty\}$ and a source – a sequence of $\mathcal{A}$-valued r.v.'s $(S_1, S_2, \dots)$. A separable distortion metric is defined for n-letter vectors by averaging the single-letter distortions:

$$d(a^n, \hat{a}^n) \triangleq \frac{1}{n} \sum d(a_i, \hat{a}_i)$$

An (n, M, D)-code is

- Encoder $f : \mathcal{A}^n \to [M]$
- Decoder $g : [M] \to \hat{\mathcal{A}}^n$

- Average distortion: $\mathbb{E}[d(S^n, g(f(S^n)))] \le D$

Fundamental limit:

$$\begin{aligned} M^*(n,D) &= \min\{M : \exists (n,M,D)\text{-code}\} \\ R(D) &= \limsup_{n\to\infty} \frac{1}{n} \log M^*(n,D) \end{aligned}$$

Now that we have the definition, we give the (surprisingly simple) general converse

Theorem 23.2 (General Converse)**.** *For all lossy codes $X \to W \to \hat{X}$ such that $\mathbb{E}[d(X,\hat{X})] \le D$, we have*

$$\log M \ge \varphi_X(D) \triangleq \inf_{P_{Y|X}:\mathbb{E}[d(X,Y)]\le D} I(X;Y)$$

where $W \in [M]$.

Proof.

$$\log M \ge H(W) \ge I(X;W) \ge I(X;\hat{X}) \ge \varphi_X(D)$$

where the last inequality follows from the fact that $P_{\hat{X}|X}$ is a feasible solution (by assumption). □

Theorem 23.3 (Properties of φ_X)**.**

1. *φ_X is convex, non-increasing.*

2. *φ_X continuous on (D_0,∞), where $D_0 = \inf\{D : \varphi_X(D) < \infty\}$.*

3. *If*

$$d(x,y) = \begin{cases} D_0 & x = y \\ > D_0 & x \ne y \end{cases}$$

   Then $\varphi_X(D_0) = I(X;X)$.

4. *Let*

$$D_{\max} = \inf_{\hat{x}\in\hat{\mathcal{X}}} \mathbb{E}d(X,\hat{x}).$$

   Then $\varphi_X(D) = 0$ for all $D > D_{\max}$. If $D_0 > D_{max}$ then also $\varphi_X(D_{max}) = 0$.

Note: If $D_{\max} = \mathbb{E}d(X,\hat{x})$ for some $\hat{x}$, then $\hat{x}$ is the "default" reconstruction of X, i.e., the best estimate when we have no information about X. Therefore $D \ge D_{\max}$ can be achieved for free. This is the reason for the notation $D_{\max}$ despite that it is defined as an infimum.
Example: (Gaussian with MSE distortion) For $X \sim \mathcal{N}(0,\sigma^2)$ and $d(x,y) = (x-y)^2$, we have $\varphi_X(D) = \frac{1}{2}\log^+ \frac{\sigma^2}{D}$. In this case $D_0 = 0$ which is not attained; $D_{\max} = \sigma^2$ and if $D \ge \sigma^2$, we can simply output $\hat{X} = 0$ as the reconstruction which requires zero bits.

Proof.

1. Convexity follows from the convexity of $P_{Y|X} \mapsto I(P_X, P_{Y|X})$.

2. Continuity on interior of the domain follows from convexity.

3. The only way to satisfy the constraint is to take $X = Y$.

4. For any $D > D_{max}$ we can set $\hat{X} = \hat{x}$ deterministically. Thus $I(X;\hat{x}) = 0$. The second claim follows from continuity. □

In channel coding, we looked at the capacity and the information capacity. We define the *Information Rate-Distortion function* in an analogous way here, which by itself is *not* an operational quantity.

Definition 23.2. The Information Rate-Distortion function for a source is

$$R_i(D) = \limsup_{n\to\infty} \frac{1}{n}\varphi_{S^n}(D) \text{ where } \varphi_{S^n}(D) = \inf_{P_{\hat{S}^n|S^n}:\mathbb{E}[d(S^n,\hat{S}^n)]\le D} I(S^n;\hat{S}^n)$$

And $D_0 = \inf\{D : R_i(D) < \infty\}$.

The reason for defining $R_i(D)$ is because from Theorem 23.2 we immediately get:

Corollary 23.1. $\forall D$, $R(D) \ge R_i(D)$.

Naturally, the information rate-distortion function inherit the properties of φ:

Theorem 23.4 (Properties of R_i)**.**

1. *$R_i(D)$ is convex, non-increasing*

2. *$R_i(D)$ is continuous on (D_0, ∞), where $D_0 \triangleq \inf\{D : R_i(D) < \infty\}$.*

3. *If*

$$d(x,y) = \begin{cases} D_0 & x = y \\ > D_0 & x \neq y \end{cases}$$

Then for stationary ergodic $\{S^n\}$, $R_i(D) = \mathcal{H}$ (entropy rate) or $+\infty$ if S_k is not discrete.

4. *$R_i(D) = 0$ for all $D > D_{\max}$, where*

$$D_{\max} \triangleq \limsup_{n\to\infty} \inf_{\hat{x^n}\in\hat{\mathcal{X}}} \mathbb{E}d(X^n, \hat{x^n}) .$$

If $D_0 < D_{\max}$, then $R_i(D_{\max}) = 0$ too.

5. *(Single letterization) If the source $\{S_i\}$ is i.i.d., then*

$$R_i(D) = \phi_{S_1}(D) = \inf_{P_{\hat{S}|S}:\mathbb{E}[d(S,\hat{S})]\le D} I(S;\hat{S})$$

Proof. Properties 1-4 follow directly from corresponding properties of ϕ_{S^n} and property 5 will be established in the next section. □

23.3* Converting excess distortion to average

Finally, we discuss how to build a compressor for average distortion if we have a compressor for excess distortion, which we will not discuss in details in class.

Assumption D_p. Assume that for (S, d), there exists $p > 1$ such that $D_p < \infty$, where

$$D_p \triangleq \sup_n \inf_{\hat{x}} (\mathbb{E}|d(S^n, \hat{x})|^p)^{1/p} < +\infty$$

i.e. that our separable distortion metric d doesn't grow too fast. Note that (by Minkowski's inequality) for stationary memoryless sources we have a single-letter bound:

$$D_p \leq \inf_{\hat{x}} (\mathbb{E}|d(S, \hat{x})|^p)^{1/p} \tag{23.5}$$

Theorem 23.5 (Excess-to-Average)**.** *Suppose there exists $X \to W \to \hat{X}$ such that $W \in [M]$ and $\mathbb{P}[d(X, \hat{X}) > D] \leq \epsilon$. Suppose for some $p \geq 1$ and $\hat{x}_0 \in \hat{\mathcal{X}}$, $(\mathbb{E}[d(X, \hat{x}_0)]^p)^{1/p} = D_p < \infty$. Then there exists $X \to W' \to \hat{X}'$ code such that $W' \in [M+1]$ and*

$$\mathbb{E}[d(X, \hat{X}')] \leq D(1 - \epsilon) + D_p \epsilon^{1-1/p} \tag{23.6}$$

Remark 23.1. Theorem is only useful for $p > 1$, since for $p = 1$ the right-hand side of (23.6) does not converge to 0 as $\epsilon \to 0$.

Proof. We transform the first code into the second by adding one codeword:

$$f'(x) = \begin{cases} f(x) & d(x, g(f(x))) \leq D \\ M+1 & \text{o/w} \end{cases}$$

$$g'(j) = \begin{cases} g(j) & j \leq M \\ \hat{x}_0 & j = M+1 \end{cases}$$

Then

$$\begin{aligned} \mathbb{E}[d(X, g' \circ f'(X)) &\leq \mathbb{E}[d(X, \hat{X}) | \hat{W} \neq M+1](1 - \epsilon) + \mathbb{E}[d(X, x_0) \mathbf{1}\{\hat{W} = M+1\}] \\ \text{(Hölders Inequality)} \quad &\leq D(1 - \epsilon) + D_p \epsilon^{1-1/p} \end{aligned}$$

□

§ 24. Rate distortion: achievability bounds

24.1 Recap

Compute $R(D)$.

Recall from the last lecture:

$$R(D) = \limsup_{n\to\infty} \frac{1}{n} \log M^*(n,D), \qquad \text{(rate distortion function)}$$
$$R_i(D) = \limsup_{n\to\infty} \frac{1}{n} \varphi_{S^n}(D), \qquad \text{(information rate distortion function)}$$

and

$$\varphi_S(D) \triangleq \inf_{P_{\hat{S}|S}:\mathbb{E}[d(S,\hat{S})]\le D} I(S;\hat{S})$$
$$\varphi_{S^n}(D) = \inf_{P_{\hat{S}^n|S^n}:\mathbb{E}[d(S^n,\hat{S}^n)]\le D} I(S^n;\hat{S}^n)$$

Also, we showed the general converse: For any (M,D)-code $X \to W \to \hat{X}$ we have

$$\begin{aligned} &\log M \ge \varphi_X(D) \\ \Longrightarrow\ &\log M^*(n,D) \ge \varphi_{S^n}(D) \\ \Longrightarrow\ &R(D) \ge R_i(D) \end{aligned}$$

In this lecture, we will prove the achievability bound and establish the identity $R(D) = R_i(D)$ for stationary memoryless sources.

First we show that $R_i(D)$ can be easily calculated for memoryless source without going through the multi-letter optimization problem.

Theorem 24.1 (Single-letterization)**.** *For stationary memoryless source S^n and separable distortion d,*

$$R_i(D) = \varphi_S(D)$$

Proof. By definition we have that $\varphi_{S^n}(D) \le n\varphi_S(D)$ by choosing a product channel: $P_{\hat{S}^n|S^n} = (P_{\hat{S}|S})^n$. Thus $R_i(D) \le \varphi_S(D)$.

For the converse, take any $P_{\hat{S}^n|S^n}$ such that the constraint $\mathbb{E}[d(S^n,\hat{S}^n)] \le D$ is satisfied, we have

$$\begin{aligned}
I(S^n;\hat{S}^n) &\ge \sum_{j=1}^n I(S_j,\hat{S}_j) && (S^n \text{ independent})\\
&\ge \sum_{j=1}^n \varphi_S(\mathbb{E}[d(S_j,\hat{S}_j)])\\
&\ge n\varphi_S\left(\frac{1}{n}\sum_{j=1}^n \mathbb{E}[d(S_j,\hat{S}_j)]\right) && (\text{convexity of } \varphi_S)\\
&\ge n\varphi_S(D) && (\varphi_S \text{ non-increasing})
\end{aligned}$$

□

24.2 Shannon's rate-distortion theorem

Theorem 24.2. *Let the source S^n be stationary and memoryless, $S^n \overset{i.i.d.}{\sim} P_S$, and suppose that distortion metric d and the target distortion D satisfy:*

1. *$d(s^n,\hat{s}^n)$ is non-negative and separable*
2. *$D > D_0$*
3. *$D_{\max}$ is finite, i.e.*

$$D_{\max} \triangleq \inf_{\hat{s}} \mathbb{E}[d(S,\hat{s})] < \infty.$$

Then

$$R(D) = R_i(D) = \inf_{P_{\hat{S}|S}:\mathbb{E}[d(S,\hat{S})]\le D} I(S;\hat{S}). \tag{24.1}$$

Remarks:

- Note that $D_{\max} < \infty$ does not imply that $d(\cdot,\cdot)$ only takes values in $\mathbb{R}$, i.e. theorem permits $d(a,\hat{a}) = \infty$.
- It should be remarked that when $D_{\max} = \infty$ typically $R(D) = \infty$. Indeed, suppose that $d(\cdot,\cdot)$ is a metric (i.e. finite valued and satisfies triangle inequality). Then, for any $x_0 \in \mathcal{A}^n$ we have

  $$d(X,\hat{X}) \ge d(X,x_0) - d(x_0,\hat{X})\,.$$

  Thus, for any finite codebook $\{c_1,\dots,c_M\}$ we have $\max_j d(x_0,c_j) < \infty$ and therefore

  $$\mathbb{E}[d(X,\hat{X})] \ge \mathbb{E}[d(X,x_0)] - \max_j d(x_0,c_j) = \infty\,.$$

  So that $R(D) = \infty$ for any finite D. This observation, however, should not be interpreted as absolute impossibility of compression for such sources. It is just not possible with fixed-rate codes. As an example, for quadratic distortion and Cauchy-distributed S, $D_{\max} = \infty$ since S has infinite second-order moments. But it is easy to see that $R_i(D) < \infty$ for any $D \in (0,\infty)$. In fact, in this case $R_i(D)$ is a hyperbola-like curve that never touches either axis. A non-trivial compression can be attained with compressors $S^n \to W$ of bounded entropy $H(W)$ (but unbounded alphabet of W). Indeed if we take W to be a Δ-quantized version of S and notice that differential entropy of S is finite, we get from (23.2) that $R_i(\Delta) \le H(W) < \infty$. Interesting question: Is $H(W) = nR_i(D) + o(n)$ attainable?

- Techniques in proving (24.1) for memoryless sources can be applied to prove it for "stationary ergodic" sources with changes similar to those we have discussed in channel coding.

Before giving a formal proof, we illustrate the intuition non-rigorously.

24.2.1 Intuition

Try to throw in M points $\mathcal{C} = \{c_1, \dots, c_M\} \in \hat{\mathcal{A}}^n$ which are drawn i.i.d. according to a product distribution $Q^n_{\hat{S}}$ where $Q_{\hat{S}}$ is some distribution on $\hat{\mathcal{A}}$. Examine the simple encoder and decoder pair:

$$\text{encoder : } f(s^n) = \operatorname*{argmin}_{j\in[M]} d(s^n, c_j) \tag{24.2}$$

$$\text{decoder : } g(j) = c_j \tag{24.3}$$

The basic idea is the following: Since the codewords are generated independently of the source, the probability that a given codeword offers good reconstruction is (exponentially) small, say, ϵ. However, since we have many codewords, the chance that there exists a good one can be of high probability. More precisely, the probability that no good codeword exist is $(1-\epsilon)^M$, which can be very close to zero as long as M grows faster than $\frac{1}{\epsilon}$.

To explain the intuition further, let us consider the excess distortion of this code: $\mathbb{P}[d(S^n, \hat{S}^n) > D]$. Define

$$P_{success} \triangleq \mathbb{P}[\exists c \in \mathcal{C}, \text{ s.t. } d(S^n, c) \le D]$$

Then

$$\begin{aligned}
P_{failure} &\triangleq \mathbb{P}[\forall c_i \in \mathcal{C}, d(S^n, c) > D] && (24.4)\\
&\approx \mathbb{P}[\forall c_i \in \mathcal{C}, d(S^n, c) > D | S^n \in T_n] && (24.5)\\
&\quad (\ T_n \text{ is the set of typical strings with empirical distribution } \hat{P}_{S^n} \approx P_S\ )\\
&= \mathbb{P}[d(S^n, \hat{S}^n) > D | S^n \in T_n]^M \qquad (P_{S^n, \hat{S}^n} = P^n_S Q^n_{\hat{S}}) && (24.6)\\
&= (1 - \underbrace{\mathbb{P}[d(S^n, \hat{S}^n) \le D | S^n \in T_n]}_{\text{since } S^n \perp\!\!\!\perp \hat{S}^n \text{, this should be small}}\ )^M && (24.7)\\
&\approx (1 - 2^{-nE(Q_{\hat{S}})})^M \qquad \text{(large deviation!)} && (24.8)
\end{aligned}$$

where it can be shown (similar to information projection) that

$$E(Q_{\hat{S}}) = \min_{P_{\hat{S}|S}: \mathbb{E}[d(S,\hat{S})] \le D} D(P_{\hat{S}|S} \| Q_{\hat{S}} | P_S) \tag{24.9}$$

Thus we conclude that $\forall Q_{\hat{S}}, \forall \delta > 0$ we can pick $M = 2^{n(E(Q_{\hat{S}})+\delta)}$ and the above code will have arbitrarily small excess distortion:

$$P_{failure} = \mathbb{P}[\forall c \in \mathcal{C}, d(S^n, c) > D] \to 0 \text{ as } n \to \infty.$$

We optimize $Q_{\hat{S}}$ to get the smallest possible M:

$$\begin{aligned}
\min_{Q_{\hat{S}}} E(Q_{\hat{S}}) &= \min_{P_{\hat{S}|S}: \mathbb{E}[d(S,\hat{S})] \le D} \min_{Q_{\hat{S}}} D(P_{\hat{S}|S} \| Q_{\hat{S}} | P_S) && (24.10)\\
&= \min_{P_{\hat{S}|S}: \mathbb{E}[d(S,\hat{S})] \le D} I(S; \hat{S})\\
&= \varphi_S(D)
\end{aligned}$$

24.2.2 Proof of Theorem 24.2

Theorem 24.3 (Performance bound of average-distortion codes)**.** *Fix P_X and suppose $d(x,\hat{x}) \geq 0$ for all $x, \hat{x}$. $\forall P_{Y|X}$, $\forall \gamma > 0$, $\forall y_0 \in \hat{\mathcal{A}}$, there exists a code $X \to W \to \hat{X}$, where $W \in [M+1]$ and*

$$\begin{aligned} \mathbb{E}[d(X,\hat{X})] &\leq \mathbb{E}[d(X,Y)] + \mathbb{E}[d(X,y_0)]e^{-M/\gamma} + \mathbb{E}[d(X,y_0)\mathbf{1}_{\{i(X;Y)>\log\gamma\}}] \\ d(X,\hat{X}) &\leq d(X,y_0) \quad a.s. \end{aligned}$$

Notes:

- This theorem says that from an arbitrary $P_{Y|X}$ such that $\mathbb{E}d(X,Y) \leq D$, we can extract a good code with average distortion D plus some extra terms which will vanish in the asymptotic regime.
- The proof uses the random coding argument. The role of the deterministic y_0 is a "fail-safe" codeword (think of y_0 as the default reconstruction with $D_{\max} = \mathbb{E}[d(X,y_0)]$). We add y_0 to the random codebook for damage control, to hedge the (highly unlikely and unlucky) event that we end up with a horrible codebook.

Proof. Similar to the previous intuitive argument, we apply random coding and generate the codewords randomly and independently of the source:

$$\mathcal{C} = \{c_1, \ldots, c_M\} \overset{\text{i.i.d.}}{\sim} P_Y \perp\!\!\!\perp X$$

and add the "fail-safe" codeword $c_{M+1} = y_0$. We adopt the same encoder-decoder pair (24.2) – (24.3) and let $\hat{X} = g(f(X))$. Then by definition,

$$d(X,\hat{X}) = \min_{j \in [M+1]} d(X, c_j) \leq d(X, y_0).$$

To simplify notation, let $\overline{Y}$ be an independent copy of Y (similar to the idea of introducing unsent codeword $\overline{X}$ in channel coding):

$$P_{X,Y,\overline{Y}} = P_{X,Y} P_{\overline{Y}}$$

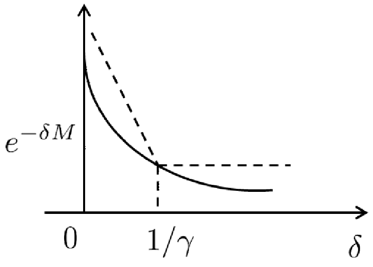
where $P_{\overline{Y}} = P_Y$. Recall the formula for computing the expectation of a random variable $U \in [0,a]$: $\mathbb{E}[U] = \int_0^a \mathbb{P}[U \geq u] du$. Then the average distortion is

$$\begin{aligned} \mathbb{E}d(X,\hat{X}) &= \mathbb{E} \min_{j \in [M+1]} d(X, c_j) && (24.11) \\ &= \mathbb{E}_X \mathbb{E}\Big[\min_{j \in [M+1]} d(X,c_j) \Big| X\Big] && (24.12) \\ &= \mathbb{E}_X \int_0^{d(X,y_0)} \mathbb{P}\Big[\min_{j \in [M+1]} d(X,c_j) > u \Big| X\Big] du && (24.13) \\ &\leq \mathbb{E}_X \int_0^{d(X,y_0)} \mathbb{P}\Big[\min_{j \in [M]} d(X,c_j) > u \Big| X\Big] du && (24.14) \\ &= \mathbb{E}_X \int_0^{d(X,y_0)} \mathbb{P}[d(X,\overline{Y}) > u | X]^M du && (24.15) \\ &= \mathbb{E}_X \int_0^{d(X,y_0)} (1 - \underbrace{\mathbb{P}[d(X,\overline{Y}) \leq u | X]}_{\triangleq \delta(X,u)})^M du && (24.16) \end{aligned}$$

Next we upper bound $(1-\delta(X,u))^M$ as follows:

$$(1-\delta(X,u))^M \le e^{-M/\gamma} + |1-\gamma\delta(X,u)|^+ \tag{24.17}$$
$$= e^{-M/\gamma} + \left|1-\gamma\mathbb{E}[\exp\{-i(X;Y)\}\mathbf{1}_{\{d(X,Y)\le u\}}|X]\right|^+ \tag{24.18}$$
$$\le e^{-M/\gamma} + \mathbb{P}[i(X;Y) > \log\gamma|X] + \mathbb{P}[d(X,Y) > u|X] \tag{24.19}$$

where

- (24.17) uses the following trick in dealing with $(1-\delta)^M$ for $\delta \ll 1$ and $M \gg 1$. First, recall the standard rule of thumb:
$$(1-\epsilon_n)^n \approx \begin{cases} 0, & \epsilon_n n \gg 1 \\ 1, & \epsilon_n n \ll 1 \end{cases}$$
In order to argue firm bounds of similar flavor, consider
$$\begin{aligned} 1-\delta M \overset{\text{union bound}}{\le} (1-\delta)^M &\le e^{-\delta M} && (\log(1-\delta)\le-\delta) \\ &\le e^{-M/\gamma}(\gamma\delta\wedge 1) + |1-\gamma\delta|^+ && (\forall\gamma>0) \\ &\le e^{-M/\gamma} + |1-\gamma\delta|^+ \end{aligned}$$


upper bound $e^{-\delta M}$
- (24.18) is simply change of measure using $i(x;y) = \log\frac{P_Y(y)}{P_{Y|X}(y|x)}$ (i.e., conditioning-unconditioning trick for information density, cf. Proposition 15.1.
- (24.19):
$$\begin{aligned} 1-\gamma\mathbb{E}[\exp\{-i(X;Y)\}\mathbf{1}_{\{d(X,Y)\le u\}}|X] &\le 1-\gamma\mathbb{E}[\exp\{-i(X;Y)\}\mathbf{1}_{\{d(X,Y)\le u, i(X;Y)\le\log\gamma\}}|X] \\ &\le 1-\mathbb{E}[\mathbf{1}_{\{d(X,Y)\le u, i(X;Y)\le\log\gamma\}}|X] \\ &= \mathbb{P}[d(X,Y)>u \text{ or } i(X;Y)>\log\gamma|X] \\ &\le \mathbb{P}[d(X,Y)>u|X] + \mathbb{P}[i(X;Y)>\log\gamma|X] \end{aligned}$$

Plugging (24.19) into (24.16), we have

$$\begin{aligned} \mathbb{E}[d(X,\hat{X})] &\le \mathbb{E}_X\int_0^{d(X,y_0)}(e^{-M/\gamma} + \mathbb{P}[i(X;Y)>\log\gamma|X] + \mathbb{P}[d(X,Y)>u|X])du \\ &\le \mathbb{E}[d(X,y_0)]e^{-M/\gamma} + \mathbb{E}[d(X,y_0)\mathbb{P}[i(X;Y)>\log\gamma|X]] + \mathbb{E}_X\int_0^\infty \mathbb{P}[d(X,Y)>u|X])du \\ &= \mathbb{E}[d(X,y_0)]e^{-M/\gamma} + \mathbb{E}[d(X,y_0)\mathbf{1}_{\{i(X;Y)>\log\gamma\}}] + \mathbb{E}[d(X,Y)] \end{aligned}$$

□

As a side product, we have the following achievability for excess distortion.

Theorem 24.4 (Performance bound of excess-distortion codes). $\forall P_{Y|X}$, $\forall \gamma > 0$, *there exists a code* $X \to W \to \hat{X}$, *where* $W \in [M]$ *and*

$$\mathbb{P}[d(X,\hat{X}) > D] \le e^{-M/\gamma} + \mathbb{P}[\{d(X,Y) > D\} \cup \{i(X;Y) > \log\gamma\}]$$

Proof. Proceed exactly as in the proof of Theorem 24.3, replace (24.11) by $\mathbb{P}[d(X,\hat{X}) > D] = \mathbb{P}[\forall j \in [M], d(X,c_j) > D] = \mathbb{E}_X[(1 - \mathbb{P}[d(X,\overline{Y}) \le D|X])^M]$, and continue similarly. □

Finally, we are able to prove Theorem 24.2 rigorously by applying Theorem 24.3 to iid sources $X = S^n$ and $n \to \infty$:

Proof of Theorem 24.2. Our goal is the achievability: $R(D) \le R_i(D) = \varphi_S(D)$.

WLOG we can assume that $D_{\max} = \mathbb{E}[d(S,\hat{s}_0)]$ achieved at some fixed $\hat{s}_0$ – this is our default reconstruction; otherwise just take any other fixed sequence so that the expectation is finite. The default reconstruction for S^n is $\hat{s}_0^n = (\hat{s}_0, \dots, \hat{s}_0)$ and $\mathbb{E}[d(S^n,\hat{s}_0^n)] = D_{\max} < \infty$ since the distortion is separable.

Fix some small $\delta > 0$. Take any $P_{\hat{S}|S}$ such that $\mathbb{E}[d(S,\hat{S})] \le D - \delta$. Apply Theorem 24.3 to $(X,Y) = (S^n,\hat{S}^n)$ with

$$\begin{aligned}
P_X &= P_{S^n} \\
P_{Y|X} &= P_{\hat{S}^n|S^n} = (P_{\hat{S}|S})^n \\
\log M &= n(I(S;\hat{S}) + 2\delta) \\
\log\gamma &= n(I(S;\hat{S}) + \delta) \\
d(X,Y) &= \frac{1}{n}\sum_{j=1}^n d(S_j,\hat{S}_j) \\
y_0 &= \hat{s}_0^n
\end{aligned}$$

we conclude that there exists a compressor $f : \mathcal{A}^n \to [M+1]$ and $g : [M+1] \to \hat{\mathcal{A}}^n$, such that

$$\begin{aligned}
\mathbb{E}[d(S^n,g(f(S^n)))] &\le \mathbb{E}[d(S^n,\hat{S}^n)] + \mathbb{E}[d(S^n,\hat{s}_0^n)]e^{-M/\gamma} + \mathbb{E}[d(S^n,\hat{s}_0^n)\mathbf{1}_{\{i(S^n;\hat{S}^n)>\log\gamma\}}] \\
&\le D - \delta + \underbrace{D_{\max}e^{-\exp(n\delta)}}_{\to 0} + \underbrace{\mathbb{E}[d(S^n,\hat{s}_0^n)\mathbf{1}_{E_n}]}_{\to 0 \text{ (later)}}, \qquad (24.20)
\end{aligned}$$

where

$$E_n = \{i(S^n;\hat{S}^n) > \log\gamma\} = \left\{\frac{1}{n}\sum_{j=1}^n i(S_j;\hat{S}_j) > I(S;\hat{S}) + \delta\right\} \qquad \xRightarrow{\text{WLLN}} \mathbb{P}[E_n] \to 0$$

If we can show the expectation in (24.20) vanishes, then there exists an $(n,M,\overline{D})$-code with:

$$M = 2^{n(I(S;\hat{S})+2\delta)}, \quad \overline{D} = D - \delta + o(1) \le D.$$

To summarize, $\forall P_{\hat{S}|S}$ such that $\mathbb{E}[d(S,\hat{S})] \le D - \delta$ we have that:

$$\begin{aligned}
&R(D) \le I(S;\hat{S}) \\
\overset{\delta\downarrow 0}{\Longrightarrow}\ &R(D) \le \varphi_S(D-) = \varphi_S(D). \ \text{(continuity, since } D > D_0)
\end{aligned}$$

It remains to show the expectation in (24.20) vanishes. This is a simple consequence of the uniform integrability of the sequence $\{d(S^n, \hat{s}_0^n)\}$. (Indeed, any sequence $V_n \overset{L_1}{\to} V$ is uniformly integrable.) If you do not know what uniform integrability is, here is a self-contained proof.

Lemma 24.1. *For any positive random variable U, define $g(\delta) = \sup_{H:\mathbb{P}[H]\le\delta} \mathbb{E}[U\mathbf{1}_H]$. Then*[1] $\mathbb{E}U < \infty \Rightarrow g(\delta) \xrightarrow{\delta\to 0} 0$.

Proof. For any $b > 0$, $\mathbb{E}[U\mathbf{1}_H] \le \mathbb{E}[U\mathbf{1}_{\{U>b\}}] + b\delta$, where $\mathbb{E}[U\mathbf{1}_{\{U>b\}}] \xrightarrow{b\to\infty} 0$ by dominated convergence theorem. Then the proof is completed by setting $b = 1/\sqrt{\delta}$. □

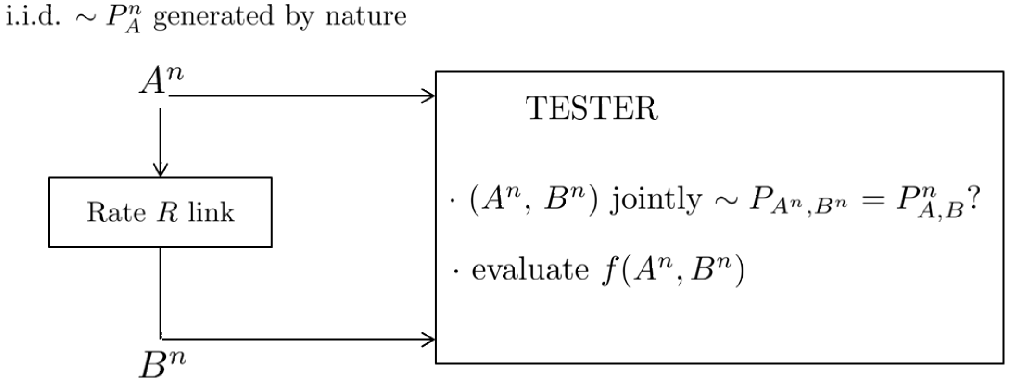
Now $d(S^n, \hat{s}_0^n) = \frac{1}{n}\sum U_j$, where U_j are iid copies of U. Since $\mathbb{E}[U] = D_{\max} < \infty$ by assumption, applying Lemma 24.1 yields $\mathbb{E}[d(S^n, \hat{s}_0^n)\mathbf{1}_{E_n}] = \frac{1}{n}\sum \mathbb{E}[U_j \mathbf{1}_{E_n}] \le g(\mathbb{P}[E_n]) \to 0$, since $\mathbb{P}[E_n] \to 0$. We are done proving the theorem. □

Note: It seems that in Section 24.2.1 and in Theorem 24.2 we applied different relaxations in showing the lower bound, how come they turn out to yield the same *tight* asymptotic result?

This is because the key to both proofs is to estimate the exponent (large deviations) of the underlined probabilities in (24.7) and (24.16), respectively. To get the right exponent, as we know, the key is to apply tilting (change of measure) to the distribution solving the information projection problem (24.9). In the case, when $P_{\overline{Y}} = (Q_{\hat{S}})^n = (P_{\hat{S}})^n$ is chosen as the solution to rate-distortion optimization $\inf I(S; \hat{S})$, the resulting tilting is precisely given by $2^{-i(X;Y)}$.

24.3* Covering lemma

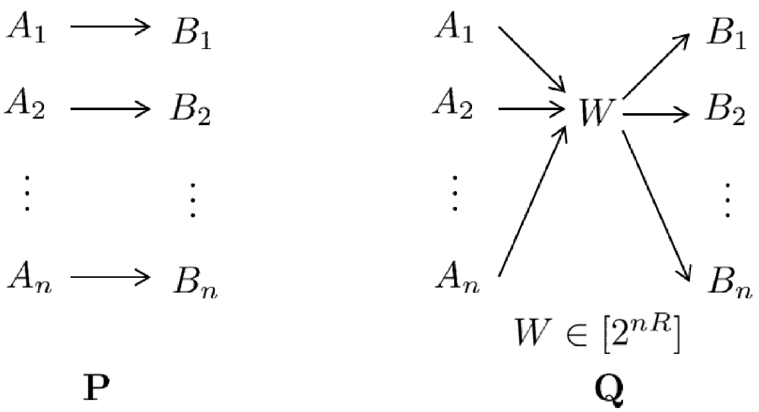
Goal:


i.i.d. $\sim P_A^n$ generated by nature
A^n
TESTER
Rate R link
$\cdot$ (A^n, B^n) jointly $\sim P_{A^n,B^n} = P_{A,B}^n$?
$\cdot$ evaluate $f(A^n, B^n)$
B^n

What's the minimum rate R needed to fool the tester?

In other words:

[1] In fact, $\Rightarrow$ is $\Leftrightarrow$.



Approximate P with Q such that for any function f, $\forall x$, we have:

$$\mathbb{P}[f(A^n, B^n) \le x] \approx \mathbb{Q}[f(A^n, B^n) \le x], \quad |W| \le 2^{nR}.$$

what is the minimum rate R to achieve this?

Some remarks:

1. The minimal rate will depend (although it is not obvious) on whether the encoder $A^n \to W$ knows about the test that the tester is running (or equivalently whether he knows the function $f(\cdot,\cdot)$).

2. If the function is known to be of the form $f(A^n, B^n) = \sum_{j=1}^n f_1(A_j, B_j)$, then evidently the job of the encoder is the following: For any realization of the sequence A^n, we need to generate a sequence B^n such that joint composition (empirical distribution) is very close to $P_{A,B}$.

3. If $R = H(A)$, we can compress A^n and send it to “B side”, who can reconstruct A^n perfectly and use that information to produce B^n through $P_{B^n|A^n}$.

4. If $R = H(B)$, “A side” can generate B^n according to $P_{A,B}^n$ and send that B^n sequence to the “B side”.

5. If $A \perp\!\!\!\perp B$, we know that $R = 0$, as “B side” can generate B^n independently.

Our previous argument turns out to give a sharp answer for the case when encoder is aware of the tester’s algorithm. Here is a precise result:

Theorem 24.5 (Covering Lemma). *$\forall P_{A,B}$ and $R > I(A;B)$, let $\mathcal{C} = \{c_1, \dots, c_M\}$ where each codeword c_j is i.i.d. drawn from distribution P_B^n. $\forall \epsilon > 0$, for $M \ge 2^{n(I(A;B)+\epsilon)}$ we have that:*

$$\mathbb{P}[\exists c \in \mathcal{C} \text{ such that } \hat{P}_{A^n,c} \approx P_{A,B}] \to 1$$

Stronger form: $\forall F$

$$\mathbb{P}[\exists c : (A^n, c) \in F] \ge \mathbb{P}[(A^n, B^n) \in F] + \underbrace{o(1)}_{\text{uniform in } F}$$

Proof. Following similar arguments of the proof for Theorem 24.3, we have

$$\begin{aligned}
\mathbb{P}[\forall c \in \mathcal{C} : (A^n, c) \notin F] &\le e^{-\gamma} + \mathbb{P}[\{(A^n, B^n) \notin F\} \cup \{i(A^n; B^n) > \log \gamma\}] \\
&= \mathbb{P}[(A^n, B^n) \notin F] + o(1) \\
\Rightarrow \mathbb{P}[\forall c \in \mathcal{C} : (A^n, c) \in F] &\ge \mathbb{P}[(A^n, B^n) \in F] + o(1)
\end{aligned}$$

□

Note: [Intuition] To generate B^n, there are around $2^{nH(B)}$ high probability sequences; for each A^n sequence, there are around $2^{nH(B|A)}$ B^n sequences that have the same joint distribution, therefore, it is sufficient to describe the class of B^n for each A^n sequence, and there are around $\frac{2^{nH(B)}}{2^{nH(B|A)}} = 2^{nI(A;B)}$ classes.

Although Covering Lemma is a powerful tool, it does not imply that the constructed joint distribution $Q_{A^nB^n}$ can fool any permutation invariant tester. In other words, it is not guaranteed that

$$\sup_{F\subset\mathcal{A}^n\times\mathcal{B}^n,\text{permut.invar.}} |Q_{A^n,B^n}(F) - P^n_{A,B}(F)| \to 0\,.$$

Indeed, a sufficient statistic for a permutation invariant tester is a joint type $\hat{P}_{A^n,c}$. Our code satisfies $\hat{P}_{A^n,c} \approx P_{A,B}$, but it might happen that $\hat{P}_{A^n,c}$ although close to $P_{A,B}$ still takes highly unlikely values (for example, if we restrict all c to have the same composition P_0, the tester can easily detect the problem since P^n_B-measure of all strings of composition P_0 cannot exceed $O(1/\sqrt{n})$). Formally, to fool permutation invariant tester we need to have small total variation between the distribution on the joint types under P and Q. (It is natural to conjecture that rate $R = I(A;B)$ should be sufficient to achieve this requirement, though).

A related question is about the minimal possible rate (i.e. cardinality of $W \in [2^{nR}]$) required to have small total variation:

$$\mathrm{TV}(Q_{A^n,B^n}, P^n_{AB}) \le \epsilon \tag{24.21}$$

Note that condition (24.21) guarantees that any tester (permutation invariant or not) is fooled to believe he sees the truly iid (A^n, B^n). The minimal required rate turns out to be (Cuff'2012):

$$R = \min_{A\to U\to B} I(A,B;U)$$

a quantity known as Wyner's common information $C(A;B)$. Showing that Wyner's common information is a lower-bound is not hard. Indeed, since $Q_{A^n,B^n} \approx P^n_{AB}$ (in TV) we have

$$I(Q_{A^{t-1},B^{t-1}}, Q_{A_tB_t|A^{t-1},B^{t-1}}) \approx I(P_{A^{t-1},B^{t-1}}, P_{A_tB_t|A^{t-1},B^{t-1}}) = 0$$

(Here one needs to use finiteness of the alphabet of A and B and the bounds relating $H(P) - H(Q)$ with $\mathrm{TV}(P,Q)$). We have (under Q!)

$$\begin{aligned} nR = H(W) &\ge I(A^n, B^n; W) &(24.22)\\ &\ge \sum_{t=1}^{T} I(A_t, B_t; W) - I(A_t, B_t; A^{t-1}B^{t-1}) &(24.23)\\ &\approx \sum_{t=1}^{T} I(A_t, B_t; W) &(24.24)\\ &\gtrsim nC(A;B) &(24.25)\end{aligned}$$

where in the last step we used the crucial observation that under Q there is a Markov chain

$$A_t \to W \to B_t$$

and that Wyner's common information $P_{A,B} \mapsto C(A;B)$ should be continuous in the total variation distance on $P_{A,B}$. Showing achievability is a little more involved.

Last time: For stationary memoryless (iid) sources and separable distortion, under the assumption that $D_{\max} < \infty$.

$$R(D) = R_i(D) = \inf_{P_{\hat{S}|S}:\mathbb{E}d(S,\hat{S})\le D} I(S;\hat{S}).$$

25.1 Evaluation of $R(D)$

So far we've proved some properties about the rate distortion function, now we'll compute its value for a few simple statistical sources. We'll do this in a somewhat unsatisfying way: guess the answer, then verify its correctness. At the end, we'll show that there is a pattern behind this method.

25.1.1 Bernoulli Source

Let $S \sim \text{Ber}(p)$, $p \le 1/2$, with Hamming distortion $d(S,\hat{S}) = \mathbf{1}\{S \ne \hat{S}\}$ and alphabets $\mathcal{A} = \hat{\mathcal{A}} = \{0,1\}$. Then $d(s^n,\hat{s}^n) = \frac{1}{n}\|s^n - \hat{s}^n\|_{\text{Hamming}}$ is the bit-error rate.

Claim: $R(D) = |h(p) - h(D)|^+$

Proof. Since $D_{\max} = p$, in the sequel we can assume $D < p$ for otherwise there is nothing to show.

(Achievability) We're free to choose any $P_{\hat{S}|S}$, so choose $S = \hat{S} + Z$, where $\hat{S} \sim \text{Ber}(p') \perp\!\!\!\perp Z \sim \text{Ber}(D)$, and p' is such that $p'(1-D) + (1-p')D = p$ so that $p' < p$. In other words, the backward channel $P_{S|\hat{S}}$ is a BSC(D). This induces some forward channel $P_{\hat{S}|S}$. Then,

$$I(S;\hat{S}) = H(S) - H(S|\hat{S}) = h(p) - h(D)$$

Since one such $P_{\hat{S}|S}$ exists, we have the upper bound $R(D) \le h(p) - h(D)$.

(Converse) First proof: For any $P_{\hat{S}|S}$ such that $P[S \ne \hat{S}] \le D \le p \le \frac{1}{2}$,

$$\begin{aligned} I(S;\hat{S}) &= H(S) - H(S|\hat{S}) \\ &= H(S) - H(S+\hat{S}|\hat{S}) \\ &\ge H(S) - H(S+\hat{S}) \\ &= h(p) - h(P[S \ne \hat{S}]) \\ &\ge h(p) - h(D) \end{aligned}$$

Second proof: Here is a more general strategy. Denote the random transformation from the achievability proof by $P^*_{\hat{S}|S}$. Now we need to show that there is no better $Q_{\hat{S}|S}$ with $\mathbb{E}_Q[d(S,\hat{S})] \le D$

and a smaller mutual information. Then consider the chain:

$$
\begin{aligned}
R(D) &\le I(P_S, Q_{\hat{S}|S}) = D(Q_{S|\hat{S}} \| P_S | Q_{\hat{S}}) \\
&= D(Q_{S|\hat{S}} \| P_{S|\hat{S}} | Q_{\hat{S}}) + \mathbb{E}_Q\left[\log \frac{P_{S|\hat{S}}}{P_S}\right] \\
(\text{Marginal } Q_{S\hat{S}} = P_S Q_{\hat{S}|S}) \quad &= D(Q_{S|\hat{S}} \| P_{S|\hat{S}} | Q_{\hat{S}}) + H(S) + \mathbb{E}_Q[\log D \mathbf{1}\{S \neq \hat{S}\} + \log \bar{D} \mathbf{1}\{S = \hat{S}\}]
\end{aligned}
$$

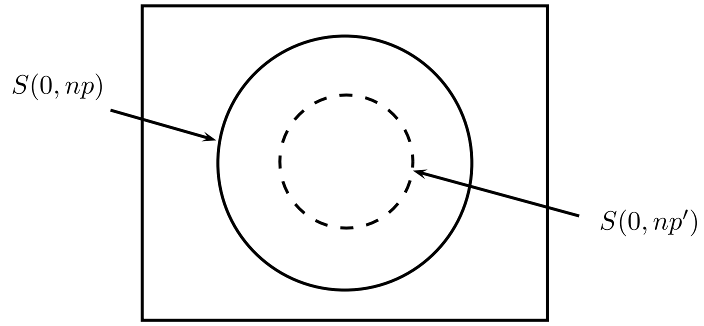
And we can minimize this expression by taking $Q_{S|\hat{S}} = P_{S|\hat{S}}$, giving

$$\ge 0 + H(S) + P[S = \hat{S}] \log(1 - D) + P[S \neq \hat{S}] \log D \ge h(p) - h(D) \quad (D \le 1/2) \tag{25.1}$$

Since the upper and lower bound agree, we have $R(D) = |h(p) - h(D)|^+$. □

For example, when $p = 1/2$, $D = .11$, then $R(D) = 1/2$ bit. In the Hamming game where we compressed 100 bits down to 50, we indeed can do this while achieving 11% average distortion, compared to the naive scheme of storing half the string and guessing on the other half, which achieves 25% average distortion.

Interpretation: By WLLN, the distribution $P_S^n = \text{Ber}(p)^n$ concentrates near the Hamming sphere of radius np as n grows large. The above result about Hamming sources tells us that the optimal reconstruction points are from $P_{\hat{S}}^n = \text{Ber}(p')^n$ where $p' < p$, which concentrates on a sphere of radius np' (note the reconstruction points are some exponentially small subset of this sphere).



Hamming Spheres

It is interesting to note that *none* of the reconstruction points are the same as any of the possible source values (with high probability).

25.1.2 Gaussian Source

The Gaussian source is defines as $\mathcal{A} = \hat{\mathcal{A}} = \mathbb{R}$, $S \sim \mathcal{N}(0, \sigma^2)$, $d(a, \hat{a}) = |a - \hat{a}|^2$ (MSE distortion).

Claim: $R(D) = \frac{1}{2} \log^+ \frac{\sigma^2}{D}$.

Proof. Since $D_{\max} = \sigma^2$, in the sequel we can assume $D < \sigma^2$ for otherwise there is nothing to show.

(Achievability) Choose $S = \hat{S} + Z$, where $\hat{S} \sim \mathcal{N}(0, \sigma^2 - D) \perp\!\!\!\perp Z \sim \mathcal{N}(0, D)$. In other words, the backward channel $P_{S|\hat{S}}$ is AWGN with noise power D. Since everything is jointly Gaussian, the forward channel can be easily found to be $P_{\hat{S}|S} = \mathcal{N}(\frac{\sigma^2 - D}{\sigma^2} S, \frac{\sigma^2 - D}{\sigma^2} D)$. Then

$$I(S; \hat{S}) = \frac{1}{2} \log \frac{\sigma^2}{D} \implies R(D) \le \frac{1}{2} \log \frac{\sigma^2}{D}$$

(Converse) Let $P_{\hat{S}|S}$ be any conditional distribution such that $\mathbb{E}_P|S-\hat{S}|^2 \le D$. Denote the forward channel in the achievability by $P^*_{\hat{S}|S}$. We use the same trick as before

$$\begin{aligned}
I(P_S, P_{\hat{S}|S}) &= D(P_{S|\hat{S}} \| P^*_{S|\hat{S}} | P_{\hat{S}}) + \mathbb{E}_P\left[\log \frac{P^*_{S|\hat{S}}}{P_S}\right] \\
&\ge \mathbb{E}_P\left[\log \frac{P^*_{S|\hat{S}}}{P_S}\right] \\
&= \mathbb{E}_P\left[\log \frac{\frac{1}{\sqrt{2\pi D}} e^{-\frac{(S-\hat{S})^2}{2D}}}{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{S^2}{2\sigma^2}}}\right] \\
&= \frac{1}{2}\log\frac{\sigma^2}{D} + \frac{\log e}{2}\mathbb{E}_P\left[\frac{S^2}{\sigma^2} - \frac{|S-\hat{S}|^2}{D}\right] \\
&\ge \frac{1}{2}\log\frac{\sigma^2}{D}.
\end{aligned}$$

Again, the upper and lower bounds agree. □

The interpretation in the Gaussian case is very similar to the case of the Hamming source. As n grows large, our source distribution concentrates on $S(0,\sqrt{n\sigma^2})$ (n-sphere in Euclidean space rather than Hamming), and our reconstruction points on $S(0,\sqrt{n(\sigma^2-D)})$. So again the picture is two nested sphere.

How sensitive is the rate-distortion formula to the Gaussianity assumption of the source?

Theorem 25.1. *Assume that $\mathbb{E}S = 0$ and $\operatorname{Var} S = \sigma^2$. Let the distortion metric be quadratic: $d(s,\hat{s}) = (s-\hat{s})^2$. Then*

$$\frac{1}{2}\log^+\frac{\sigma^2}{D} - D(P_S\|\mathcal{N}(0,\sigma^2)) \le R(D) = \inf_{P_{\hat{S}|S}:\mathbb{E}(\hat{S}-S)^2\le D} I(S;\hat{S}) \le \frac{1}{2}\log^+\frac{\sigma^2}{D}.$$

Note: This result is in exact parallel to what we proved in Theorem 17.6 for additive-noise channel capacity:

$$\frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) \le \sup_{P_X:\mathbb{E}X^2\le P} I(X;X+Z) \le \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) + D(P_Z\|\mathcal{N}(0,\sigma^2)).$$

where $\mathbb{E}Z = 0$ and $\operatorname{Var} Z = \sigma^2$.

Note: A simple consequence of Theorem 25.1 is that for source distributions with a density, the rate-distortion function grows according to $\frac{1}{2}\log\frac{1}{D}$ in the low-distortion regime as long as $D(P_S\|\mathcal{N}(0,\sigma^2))$ is finite. In fact, the first inequality, known as the *Shannon lower bound*, is asymptotically tight, i.e., $R(D) = \frac{1}{2}\log\frac{\sigma^2}{D} - D(P_S\|\mathcal{N}(0,\sigma^2)) + o(1)$ as $D \to 0$. Therefore in this regime performing uniform scalar quantization with accuracy $\frac{1}{\sqrt{D}}$ is in fact asymptotically optimal within an $o(1)$ term.

Proof. Again, assume $D < D_{\max} = \sigma^2$. Let $S_G \sim \mathcal{N}(0,\sigma^2)$.

(Achievability) Use the same $P^*_{\hat{S}|S} = \mathcal{N}(\frac{\sigma^2-D}{\sigma^2}S, \frac{\sigma^2-D}{\sigma^2}D)$ in the achievability proof of Gaussian rate-distortion function:

$$\begin{aligned}
R(D) &\le I(P_S, P^*_{\hat{S}|S}) \\
&= I(S; \frac{\sigma^2-D}{\sigma^2}S + W) && W \sim \mathcal{N}(0, \frac{\sigma^2-D}{\sigma^2}D) \\
&\le I(S_G; \frac{\sigma^2-D}{\sigma^2}S_G + W) && \text{by Gaussian saddle point (Theorem 4.6)} \\
&= \frac{1}{2}\log\frac{\sigma^2}{D}.
\end{aligned}$$

(Converse) For any $P_{\hat{S}|S}$ such that $\mathbb{E}(\hat{S}-S)^2 \le D$. Let $P^*_{S|\hat{S}} = \mathcal{N}(\hat{S}, D)$ denote AWGN with noise power D. Then

$$\begin{aligned}
I(S;\hat{S}) &= D(P_{S|\hat{S}} \| P_S | P_{\hat{S}}) \\
&= D(P_{S|\hat{S}} \| P^*_{S|\hat{S}} | P_{\hat{S}}) + \mathbb{E}_P\left[\log\frac{P^*_{S|\hat{S}}}{P_{S_G}}\right] - D(P_S \| P_{S_G}) \\
&\ge \mathbb{E}_P\left[\log\frac{\frac{1}{\sqrt{2\pi D}}e^{-\frac{(S-\hat{S})^2}{2D}}}{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{S^2}{2\sigma^2}}}\right] - D(P_S \| P_{S_G}) \\
&\ge \frac{1}{2}\log\frac{\sigma^2}{D} - D(P_S \| P_{S_G}).
\end{aligned}$$

□

Remark: The theory of quantization and the rate distortion theory at large have played a significant role in pure mathematics. For instance, Hilbert's thirteenth problem was partially solved by Arnold and Kolmogorov after they realized that they could classify spaces of functions looking at the optimal quantizer for such functions.

25.2* Analog of saddle-point property in rate-distortion

In the computation of $R(D)$ for the Hamming and Gaussian source, we guessed the correct form of the rate distortion function. In both of their converse arguments, we used the same trick to establish that any other $P_{\hat{S}|S}$ gave a larger value for $R(D)$. In this section, we formalize this trick, in an analogous manner to the saddle point property of the channel capacity. Note that typically we don't need any tricks to compute $R(D)$, since we can obtain a solution in parametric form to the unconstrained convex optimization

$$\min_{P_{\hat{S}|S}} I(S;\hat{S}) + \lambda\,\mathbb{E}[d(S,\hat{S})]$$

In fact there are also iterative algorithms (Blahut-Arimoto) that computes $R(D)$. However, for peace of mind it is good to know there are some general reasons why tricks like we used in Hamming/Gaussian actually are guaranteed to work.

Theorem 25.2. *1. Suppose P_{Y^*} and $P_{X|Y^*} \ll P_X$ are found with the property that $\mathbb{E}[d(X,Y^*)] \le D$ and for any P_{XY} with $\mathbb{E}[d(X,Y)] \le D$ we have*

$$\mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] \ge I(X;Y^*)\,. \tag{25.2}$$

Then $R(D) = I(X;Y^*)$.

2. Suppose that $I(X;Y^*) = R(D)$. Then for any regular branch of conditional probability $P_{X|Y^*}$ and for any P_{XY} satisfying

- *$\mathbb{E}[d(X,Y)] \le D$ and*
- *$P_Y \ll P_{Y^*}$ and*
- *$I(X;Y) < \infty$*

the inequality (25.2) holds.

Remarks:

1. The first part is a sufficient condition for optimality of a given P_{XY^*}. The second part gives a necessary condition that is convenient to narrow down the search. Indeed, typically the set of P_{XY} satisfying those conditions is rich enough to infer from (25.2):

$$\log \frac{dP_{X|Y^*}}{dP_X}(x|y) = R(D) - \theta[d(x,y) - D]\,,$$

for a positive $\theta > 0$.

2. Note that the second part is not valid without $P_Y \ll P_{Y^*}$ condition. The counter-example to this and various other erroneous (but frequently encountered) generalizations is the following: $\mathcal{A} = \{0,1\}$, $P_X = \text{Bern}(1/2)$, $\hat{\mathcal{A}} = \{0,1,0',1'\}$ and

$$d(0,0) = d(0,0') = 1 - d(0,1) = 1 - d(0,1') = 0\,.$$

The $R(D) = |1 - h(D)|^+$, but there are a bunch of non-equivalent optimal $P_{Y|X}$, $P_{X|Y}$ and P_Y's.

Proof. First part is just a repetition of the proofs above, so we focus on part 2. Suppose there exists a counter-example P_{XY} achieving

$$I_1 = \mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] < I^* = R(D)\,.$$

Notice that whenever $I(X;Y) < \infty$ we have

$$I_1 = I(X;Y) - D(P_{X|Y} \| P_{X|Y^*} | P_Y)\,,$$

and thus

$$D(P_{X|Y} \| P_{X|Y^*} | P_Y) < \infty\,. \tag{25.3}$$

Before going to the actual proof, we describe the principal idea. For every λ we can define a joint distribution

$$P_{X,Y_\lambda} = \lambda P_{X,Y} + (1-\lambda) P_{X,Y^*}\,.$$

Then, we can compute

$$I(X;Y_\lambda) = \mathbb{E}\left[\log \frac{P_{X|Y_\lambda}}{P_X}(X|Y_\lambda)\right] = \mathbb{E}\left[\log \frac{P_{X|Y_\lambda}}{P_{X|Y^*}}\frac{P_{X|Y^*}}{P_X}\right] \tag{25.4}$$

$$= D(P_{X|Y_\lambda}\|P_{X|Y^*}|P_{Y_\lambda}) + \mathbb{E}\left[\frac{P_{X|Y^*}(X|Y_\lambda)}{P_X}\right] \tag{25.5}$$

$$= D(P_{X|Y_\lambda}\|P_{X|Y^*}|P_{Y_\lambda}) + \lambda I_1 + (1-\lambda)I_*\,. \tag{25.6}$$

From here we will conclude, similar to Prop. 4.1, that the first term is $o(\lambda)$ and thus for sufficiently small λ we should have $I(X;Y_\lambda) < R(D)$, contradicting optimality of coupling P_{X,Y^*}.

We proceed to details. For every $\lambda \in [0,1]$ define

$$\rho_1(y) \triangleq \frac{dP_Y}{dP_{Y^*}}(y) \tag{25.7}$$

$$\lambda(y) \triangleq \frac{\lambda\rho_1(y)}{\lambda\rho_1(y)+\bar\lambda} \tag{25.8}$$

$$P^{(\lambda)}_{X|Y=y} = \lambda(y)P_{X|Y=y} + \bar\lambda(y)P_{X|Y^*=y} \tag{25.9}$$

$$dP_{Y_\lambda} = \lambda dP_Y + \bar\lambda dP_{Y^*} = (\lambda\rho_1(y)+\bar\lambda)dP_{Y^*} \tag{25.10}$$

$$D(y) = D(P_{X|Y=y}\|P_{X|Y^*=y}) \tag{25.11}$$

$$D_\lambda(y) = D(P^{(\lambda)}_{X|Y=y}\|P_{X|Y^*=y})\,. \tag{25.12}$$

Notice:

$$\text{On } \{\rho_1 = 0\}: \quad \lambda(y) = D(y) = D_\lambda(y) = 0$$

and otherwise $\lambda(y) > 0$. By convexity of divergence

$$D_\lambda(y) \le \lambda(y)D(y)$$

and therefore

$$\frac{1}{\lambda(y)}D_\lambda(y)1\{\rho_1(y)>0\} \le D(y)1\{\rho_1(y)>0\}\,.$$

Notice that by (25.3) the function $\rho_1(y)D(y)$ is non-negative and P_{Y^*}-integrable. Then, applying dominated convergence theorem we get

$$\lim_{\lambda\to0}\int_{\{\rho_1>0\}} dP_{Y^*}\frac{1}{\lambda(y)}D_\lambda(y)\rho_1(y) = \int_{\{\rho_1>0\}} dP_{Y^*}\rho_1(y)\lim_{\lambda\to0}\frac{1}{\lambda(y)}D_\lambda(y) = 0 \tag{25.13}$$

where in the last step we applied the result from Lecture 4

$$D(P\|Q) < \infty \qquad \Longrightarrow \qquad D(\lambda P + \bar\lambda Q\|Q) = o(\lambda)$$

since for each y on the set $\{\rho_1 > 0\}$ we have $\lambda(y) \to 0$ as $\lambda \to 0$.

On the other hand, notice that

$$\int_{\{\rho_1>0\}} dP_{Y^*}\frac{1}{\lambda(y)}D_\lambda(y)\rho_1(y)1\{\rho_1(y)>0\} = \frac{1}{\lambda}\int_{\{\rho_1>0\}} dP_{Y^*}(\lambda\rho_1(y)+\bar\lambda)D_\lambda(y) \tag{25.14}$$

$$= \frac{1}{\lambda}\int_{\{\rho_1>0\}} dP_{Y_\lambda}D_\lambda(y) \tag{25.15}$$

$$= \frac{1}{\lambda}\int_{\mathcal{Y}} dP_{Y_\lambda}D_\lambda(y) = \frac{1}{\lambda}D(P^{(\lambda)}_{X|Y}\|P_{X|Y^*}|P_{Y_\lambda})\,, \tag{25.16}$$

where in the penultimate step we used $D_\lambda(y) = 0$ on $\{\rho_1 = 0\}$. Hence, (25.13) shows

$$D(P^{(\lambda)}_{X|Y} \| P_{X|Y^*} | P_{Y_\lambda}) = o(\lambda)\,, \qquad \lambda \to 0\,.$$

Finally, since

$$P^{(\lambda)}_{X|Y} \circ P_{Y_\lambda} = P_X\,,$$

we have

$$I(X;Y_\lambda) = D(P^{(\lambda)}_{X|Y} \| P_{X|Y^*} | P_{Y_\lambda}) + \lambda\,\mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y)\right] + \bar{\lambda}\,\mathbb{E}\left[\log \frac{dP_{X|Y^*}}{dP_X}(X|Y^*)\right] \tag{25.17}$$

$$= I^* + \lambda(I_1 - I^*) + o(\lambda)\,, \tag{25.18}$$

contradicting the assumption

$$I(X;Y_\lambda) \ge I^* = R(D)\,.$$

□

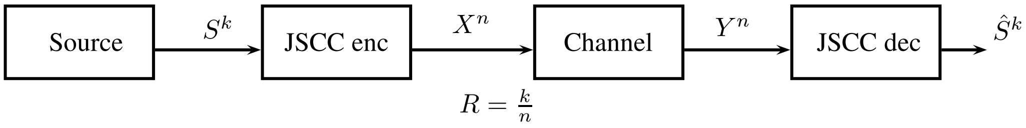
25.3 Lossy joint source-channel coding

The *lossy joint source channel coding problem* refers to the fundamental limits of lossy compression followed by transmission over a channel.

Problem Setup: For an $\mathcal{A}$-valued $(\{S_1, S_2, \dots\}$ and distortion metric $d : \mathcal{A}^k \times \hat{\mathcal{A}}^k \to \mathbb{R}$, a lossy JSCC is a pair (f, g) such that

$$S^k \xrightarrow{f} X^n \xrightarrow{\text{ch.}} Y^n \xrightarrow{g} \hat{S}^k$$

Definition 25.1. (f, g) is a (k, n, D)-JSCC if $\mathbb{E}[d(S^k, \hat{S}^k)] \le D$.



where ρ is the *bandwidth expansion factor*:

$$\rho = \frac{n}{k} \quad \text{channel uses/symbol.}$$

Our goal is to minimize ρ subject to a fidelity guarantee by designing the encoder/decoder pairs smartly. The asymptotic fundamental limit for a lossy JSCC is

$$\rho^*(D) = \limsup_{n\to\infty} \min\{\frac{n}{k} : \exists (k, n, D) - code\}$$

For simplicity in this lecture we will focus on JSCC for stationary memoryless sources with separable distortion + stationary memoryless channels.

25.3.1 Converse

The converse for the JSCC is quite simple. Note that since there is no ϵ under consideration, the strong converse is the same as the weak converse. The proof architecture is identical to the weak converse of lossless JSCC which uses Fano's inequality.

Theorem 25.3 (Converse)**.** *For any source such that*

$$R_i(D) = \lim_{k\to\infty} \frac{1}{k} \inf_{P_{\hat{S}^k|S^k}:\mathbb{E}[d(S^k,\hat{S}^k)]\le D} I(S^k;\hat{S}^k)$$

we have

$$\rho^*(D) \ge \frac{R_i(D)}{C_i}$$

Remark: The requirement of this theorem on the source isn't too stringent; the limit expression for $R_i(D)$ typically exists for stationary sources (like for the entropy rate)

Proof. Take a (k,n,D)-code $S^k \to X^n \to Y^n \to \hat{S}^k$. Then

$$\inf_{P_{\hat{S}^k|S^k}} I(S^k;\hat{S}^k) \le I(S^k;\hat{S}^k) \le I(X^k;Y^k) \le \sup_{P_{X^n}} I(X^n;Y^n)$$

Which follows from data processing and taking inf/sup. Normalizing by $1/k$ and taking the liminf as $n \to \infty$

$$\text{(LHS)}\ \liminf_{n\to\infty} \frac{1}{n} \sup_{P_{X^n}} I(X^n;Y^n) = C_i$$
$$\text{(RHS)}\ \liminf_{n\to\infty} \frac{1}{k_n} \inf_{P_{\hat{S}^{k_n}|S^{k_n}}} I(S^{k_n};\hat{S}^{k_n}) = R_i(D)$$

And therefore, any sequence of (k_n,n,D)-codes satisfies

$$\limsup_{n\to\infty} \frac{n}{k_n} \ge \frac{R_i(D)}{C_i}$$

□

Note: Clearly the assumptions in Theorem 25.3 are satisfied for memoryless sources. If the source S is iid Bern(1/2) with Hamming distortion, then Theorem 25.3 coincides with the weak converse for channel coding under bit error rate in Theorem 14.4:

$$k \le \frac{nC}{1-h(p_b)}$$

which we proved using ad hoc techniques. In the case of channel with cost constraints, e.g., the AWGN channel with $C(\text{SNR}) = \frac{1}{2}\log(1+\text{SNR})$, we have

$$p_b \ge h^{-1}\left(1 - \frac{C(\text{SNR})}{R}\right)$$

This is often referred to as the Shannon limit in plots comparing the bit-error rate of practical codes. See, e.g., Fig. 2 from [RSU01] for BIAWGN (binary-input) channel. *This is erroneous*, since the p_b above refers to the bit-error of data bits (or systematic bits), not all of the codeword bits. The latter quantity is what typically called BER in the coding-theoretic literature.

25.3.2 Achievability via separation

The proof strategy is similar to the lossless JSCC: We construct a separated lossy compression and channel coding scheme using our tools from those areas, i.e., let the JSCC encoder to be the concatenation of a loss compressor and a channel encoder, and the JSCC decoder to be the concatenation of a channel decoder followed by a loss compressor, then show that this separated construction is optimal.

Theorem 25.4. *For any stationary memoryless source* $(P_S, \mathcal{A}, \hat{\mathcal{A}}, d)$ *satisfying assumption A1 (below), and for any stationary memoryless channel* $P_{Y|X}$*,*

$$\rho^*(D) = \frac{R(D)}{C}$$

Note: The assumption on the source is to control the distortion incurred by the channel decoder making an error. Although we know that this is a low-probability event, without any assumption on the distortion metric, we cannot say much about its contribution to the end-to-end average distortion. This will not be a problem if the distortion metric is bounded (for which Assumption A1 is satisfied of course). Note that we do not have this nuisance in the lossless JSCC because we at most suffer the channel error probability (union bound).

The assumption is rather technical which can be skipped in the first reading. Note that it is trivially satisfied by bounded distortion (e.g., Hamming), and can be shown to hold for Gaussian source and MSE distortion.

Proof. The converse direction follows from the previous theorem. For the other direction, we constructed a separated compression / channel coding scheme. Take

$$\begin{aligned} &S^k \to W \to \hat{S}^k \ \text{ compressor to } W \in [2^{kR(D)+o(k)}] \text{ with } \mathbb{E}[d(S^k, \hat{S}^k)] \le D \\ &W \to X^n \to Y^n \to \hat{W} \ \text{ maximal probability of error channel code (assuming } kR(D) \le nC + o(n)) \\ &\qquad\qquad\qquad\qquad \text{with } \mathbb{P}[W \ne \hat{W}] \le \epsilon \ \ \forall P_W \end{aligned}$$

So that the overall system is

$$S^k \longrightarrow W \longrightarrow X^n \longrightarrow Y^n \longrightarrow \hat{W} \longrightarrow \hat{S}^k$$

Note that here we need a **maximum** probability of error code since when we concatenate these two schemes, W at the input of the channel is the output of the source compressor, which is not guaranteed to be uniform. Now that we have a scheme, we must analyze the average distortion to show that it meets the end-to-end distortion constraint. We start by splitting the expression into two cases

$$\mathbb{E}[d(S^k, \hat{S}^k)] = \mathbb{E}[d(S^k, \hat{S}^k(W))\mathbf{1}\{W = \hat{W}\}] + \mathbb{E}[d(S^k, \hat{S}^k(\hat{W}))\mathbf{1}\{W \ne \hat{W}\}]$$

By assumption on our lossy code, we know that the first term is $\le D$. In the second term, we know that the probability of the event $\{W \ne \hat{W}\}$ is small by assumption on our channel code, but we cannot say anything about $\mathbb{E}[d(S^k, \hat{S}^k(\hat{W}))]$ unless, for example, d is bounded. But by Lemma 25.1 (below), $\exists$ code $S^k \to W \to \hat{S}^k$ such that

(1) $\mathbb{E}[d(S^k, \hat{S}^k)] \le D$ holds

(2) $d(a_0^k, \hat{S}^k) \le L$ for all quantization outputs $\hat{S}^k$, where $a_0^k = (a_0, \ldots, a_0)$ is some fixed string of length k from the Assumption A1 below.

The second bullet says that all points in the reconstruction space are "close" to some fixed string. Now we can deal with the troublesome term

$$\begin{aligned}\mathbb{E}[d(S^k,\hat{S}^k(\hat{W}))\mathbf{1}\{W\neq\hat{W}\}] &\leq \mathbb{E}[\mathbf{1}\{W\neq\hat{W}\}\lambda(d(S^k,\hat{a}_0^k)+d(a_0^k,\hat{S}^k))]\\ \text{(by point (2) above)}\quad &\leq \lambda\mathbb{E}[\mathbf{1}\{W\neq\hat{W}\}d(S^k,\hat{a}_0^k)]+\lambda\mathbb{E}[\mathbf{1}\{W\neq\hat{W}\}L]\\ &\leq \lambda o(1)+\lambda L\epsilon\to 0 \text{ as } \epsilon\to 0\end{aligned}$$

where in the last step we applied the same uniform integrability argument that showed vanishing of the expectation in (24.20) before.

In all, our scheme meets the average distortion constraint. Hence we conclude that for $\forall\rho>\frac{R(D)}{C}$, $\exists$ sequence of $(k,n,D+o(1))$-codes. □

The following assumption is critical to the previous theorem:

Assumption A1: For a source $(P_S,\mathcal{A},\hat{\mathcal{A}},d)$, $\exists\lambda\geq 0, a_0\in\mathcal{A},\hat{a}_0\in\hat{\mathcal{A}}$ such that

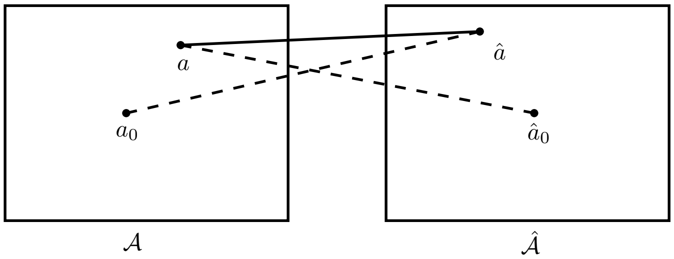
1. $d(a,\hat{a})\leq\lambda(d(a,\hat{a}_o)+d(a_0,\hat{a}))\ \ \forall a,\hat{a}$ (generalized triangle inequality)

2. $\mathbb{E}[d(S,\hat{a}_0)]<\infty$ (so that $D_{\max}<\infty$ too).

3. $\mathbb{E}[d(a_0,\hat{S})]<\infty$ for any output distribution $P_{\hat{S}}$ achieving the rate-distortion function $R(D)$ at some D.

4. $d(a_0,\hat{a}_0)<\infty$.

This assumption says that the spaces $\mathcal{A}$ and $\hat{\mathcal{A}}$ have "nice centers", in the sense that the distance between any two points is upper bounded by a constant times the distance from the centers to each point (see figure below).



But the assumption isn't easy to verify, or clear which sources satisfy the assumptions. Because of this, we now give a few sufficient conditions for Assumption A1 to hold.

Trivial Condition: If the distortion function is bounded, then the assumption A1 holds automatically. In other words, if we have a discrete source with finite alphabet $|\mathcal{A}|,|\hat{\mathcal{A}}|<\infty$ and a finite distortion function $d(a,\hat{a})<\infty$, then A1 holds. More generally, we have the following criterion.

Theorem 25.5 (Criterion for satisfying A1)**.** *If $\mathcal{A}=\hat{\mathcal{A}}$ and $d(a,\hat{a})=\rho^q(a,\hat{a})$ for some metric ρ with $q\geq 1$, and $D_{\max}\triangleq\inf_{\hat{a}_0}\mathbb{E}[d(S,\hat{a}_0)]<\infty$, then A1 holds.*

Proof. Take $a_0=\hat{a}_0$ that achieves finite D_p (in fact, any points can serve as centers in a metric space). Then

$$\begin{aligned}(\frac{1}{2}\rho(a,\hat{a}))^q &\leq \left(\frac{1}{2}\rho(a,a_0)+\frac{1}{2}\rho(a_0,\hat{a})\right)^q\\ \text{(Jensen's)}\quad &\leq \frac{1}{2}\rho^q(a,a_0)+\frac{1}{2}\rho^q(a_0,\hat{a})\end{aligned}$$

And thus $d(a,\hat{a}) \leq 2^{q-1}(d(a,a_0)+d(a_0,\hat{a}))$. Taking $\lambda = 2^{q-1}$ verifies (1) and (2) in A1. To verify (3), we can use this generalized triangle inequality for our source

$$d(a_0,\hat{S}) \leq 2^{q-1}(d(a_0,S)+d(S,\hat{S}))$$

Then taking the expectation of both sides gives

$$\begin{aligned}\mathbb{E}[d(a_0,\hat{S})] &\leq 2^{q-1}(\mathbb{E}[d(a_0,S)]+\mathbb{E}[d(S,\hat{S})]) \\ &\leq 2^{q-1}(D_{\max}+D) < \infty\end{aligned}$$

So that condition (3) in A1 holds. □

So we see that metrics raised to powers (e.g. squared Euclidean norm) satisfy the condition A1. The lemma used in Theorem 25.4 is now given.

Lemma 25.1. *Fix a source satisfying A1 and an arbitrary $P_{\hat{S}|S}$. Let $R > I(S;\hat{S})$, $L > \max\{\mathbb{E}[d(a_0,\hat{S})], d(a_0,\hat{a}_0)\}$ and $D > \mathbb{E}[d(S,\hat{S})]$. Then, there exists a $(k,2^{kR},D)$-code such that for every reconstruction point $\hat{x} \in \hat{\mathcal{A}}^k$ we have $d(a_0^k,\hat{x}) \leq L$.*

Proof. Let $\mathcal{X} = \mathcal{A}^k$, $\hat{\mathcal{X}} = \hat{\mathcal{A}}^k$ and $P_X = P_S^k, P_{Y|X} = P_{\hat{S}|S}^k$. Then apply the achievability bound for excess distortion from Theorem 24.4 with

$$d_1(x,\hat{x}) = \begin{cases} d(x,\hat{x}) & d(a_0^k,\hat{x}) \leq L \\ +\infty & \text{o/w}\end{cases}$$

Note that this is NOT a separable distortion metric. Also note that without any change in d_1-distortion we can remove all (if any) reconstruction points $\hat{x}$ with $d(a_0^k,\hat{x}) > L$. Furthermore, from the WLLN we have for any $D > D' > \mathbb{E}[d(S,\hat{S}')]$

$$\mathbb{P}[d_1(X,Y) > D'] \leq \mathbb{P}[d(S^k,\hat{S}^k) > D'] + \mathbb{P}[d(a_0^k,\hat{S}^k) > L] \to 0$$

as $k \to \infty$ (since $\mathbb{E}[d(S,\hat{S})] < D'$ and $\mathbb{E}[a_0,\hat{S}] < L$). Thus, overall we get $M = 2^{kR}$ reconstruction points $(c_1,\ldots,c_M)$ such that

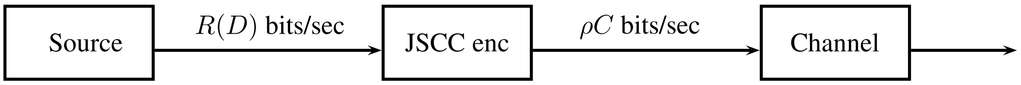
$$\mathbb{P}[\min_{j\in[M]} d(S^k,c_j) > D'] \to 0$$

and $d(a_0^k,c_j) \leq L$. By adding $c_{M+1} = (\hat{a}_0,\ldots,\hat{a}_0)$ we get

$$\mathbb{E}[\min_{j\in[M+1]} d(S^k,c_j)] \leq D' + \mathbb{E}[d(S^k,c_{M+1})1\{\min_{j\in[M]} d(S^k,c_j) > D'\}] = D' + o(1)\,,$$

where the last estimate follows from uniform integrability as in the vanishing of expectation in (24.20). Thus, for sufficiently large n the expected distortion is $\leq D$, as required. □

To summarize the results in this section, under stationarity and memorylessness assumptions on the source and the channel, we have shown that the following separately-designed scheme achieves the optimal rate for lossy JSCC: First compress the data, then encode it using your favorite channel code, then decompress at the receiver.



25.4 What is lacking in classical lossy compression?

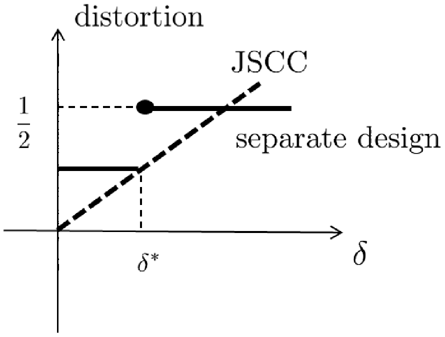
Examples of some issues with the classical compression theory:

- compression: we can apply the standard results in compression of a text file, but it is extremely difficult for image files due to the strong spatial correlation. For example, the first sentence and the last in Tolstoy's novel are pretty uncorrelated. But the regions in the upper-left and bottom-right corners of one image can be strongly correlated. Thus for practicing the lossy compression of videos and images the key problem is that of coming up with a good "whitening" basis.
- JSCC: Asymptotically the separation principle sounds good, but the separated systems can be very unstable - no graceful degradation. Consider the following example of JSCC.

  Example: Source = $Bern(\frac{1}{2})$, channel = $BSC(\delta)$.

  1. separate compressor and channel encoder designed for $\frac{R(D)}{C(\delta)} = 1$
  2. a simple JSCC:

  $$\rho = 1, X_i = S_i$$

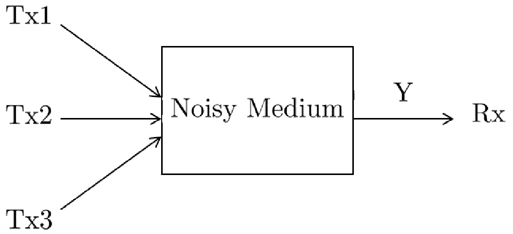


no graceful degradation of separately designed source channel code

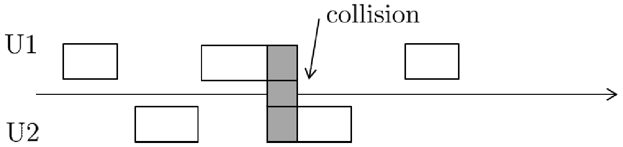
Part VI

Advanced topics

26.1 Problem motivation and main results



Note: In network community, people are mostly interested in channel access control mechanisms that help to detect or avoid data packet collisions so that the channel is shared among multiple users.



The famous ALOHA protocal achieves

$$\sum_i R_i \approx 0.37C$$

where C is the (single-user) capacity of the channel.[1]

In information theory community, the goal is to achieve

$$\sum_i R_i > C$$

The key to achieve this is to use coding so that collisions are resolvable.

In the following discussion we shall focus on the case with two users. This is without loss of much generality, as all the results can easily be extended to N users.

Definition 26.1.

- Multiple-access channel: $\{P_{Y^n|A^n,B^n} : \mathcal{A}^n \times \mathcal{B}^n \to \mathcal{Y}^n, n = 1, 2, \dots\}$.
- a (n, M_1, M_2, ϵ) code is specified by

$$f_1 : [M_1] \to \mathcal{A}^n, \quad f_2 : [M_2] \to \mathcal{B}^n$$
$$g : \mathcal{Y}^n \to [M_1] \times [M_2]$$

[1] Note that there is a lot of research about how to achieve even these 37%. Indeed, ALOHA in a nutshell simply postulates that everytime a user has a packet to transmit, he should attempt transmission in each time slot with probability p, independently. The optimal setting of p is the inverse of the number of actively trying users. Thus, it is non-trivial how to learn the dynamically changing number of active users without requiring a central authority. This is how ideas such as exponential backoff etc arise.

$$\begin{tikzcd}
W_1 \arrow[r, "f_1"] & A^n \arrow[dr] & & \hat{W}_1 \\
& & Y^n \arrow[ur, "g"] \arrow[dr] & \\
W_2 \arrow[r, "f_2"] & B^n \arrow[ur] & & \hat{W}_2
\end{tikzcd}$$

$$P$$

$W_1, W_2 \sim$ uniform, and the codes achieves

$$\mathbb{P}[\{W_1 \neq \hat{W}_1\} \bigcup \{W_2 \neq \hat{W}_2\}] \leq \epsilon$$

- Fundamental limit of capacity region

$$\mathcal{R}^*(n, \epsilon) = \{(R_1, R_2) : \exists \text{ a } (n, 2^{nR_1}, 2^{nR_2}, \epsilon) \text{ code}\}$$

- Asymptotics:

$$\mathcal{C}_\epsilon = \Big[\liminf_{n\to\infty} \mathcal{R}^*(n, \epsilon)\Big]$$

where $[\cdot]$ denotes the closure of a set.

Note: $\liminf$ and $\limsup$ of a sequence of sets $\{A_n\}$:

$$\liminf_n A_n = \{\omega : \omega \in A_n, \forall n \geq n_0\}$$

$$\limsup_n A_n = \{\omega : \omega \text{ infinitely occur}\}$$

- $$\mathcal{C} = \lim \mathcal{C}_\epsilon = \bigcap_{\epsilon > 0} \mathcal{C}_\epsilon$$

Theorem 26.1 (Capacity region)**.**

$$\mathcal{C}_\epsilon = \overline{co} \bigcup_{P_A, P_B} Penta(P_A, P_B) \tag{26.1}$$

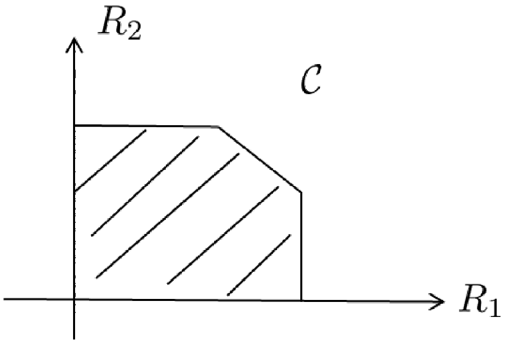
$$= \Big[\bigcup_{P_{U,A,B} = P_U P_{A|U} P_{B|U}} Penta(P_{A|U}, P_{B|U} | P_U)\Big] \tag{26.2}$$

where $\overline{co}$ is the set operator of constructing the convex hull followed by taking the closure, and $Penta(\cdot,\cdot)$ is defined as follows:

$$Penta(P_A, P_B) = \left\{ (R_1, R_2) : \begin{array}{r} 0 \leq R_1 \leq I(A; Y|B) \\ 0 \leq R_2 \leq I(B; Y|A) \\ R_1 + R_2 \leq I(A, B; Y) \end{array} \right\}$$

$$Penta(P_{A|U}, P_{B|U} | P_U) = \left\{ (R_1, R_2) : \begin{array}{r} 0 \leq R_1 \leq I(A; Y|B, U) \\ 0 \leq R_2 \leq I(B; Y|A, U) \\ R_1 + R_2 \leq I(A, B; Y|U) \end{array} \right\}$$

Note: the two forms in (26.1) and (26.2) are equivalent without cost constraints. In the case when constraints such as $\mathbb{E}\mathsf{c}_1(A) \leq P_1$ and $\mathbb{E}\mathsf{c}_2(B) \leq P_2$ are present, only the second expression yields the true capacity region.



Penta

26.2 MAC achievability bound

First, we introduce a lemma which will be used in the proof of Theorem 26.1.

Lemma 26.1. *$\forall P_A, P_B, P_{Y|A,B}$ such that $P_{A,B,Y} = P_A P_B P_{Y|A,B}$, and $\forall \gamma_1, \gamma_2, \gamma_{12} > 0$, $\forall M_1, M_2$, there exists a (M_1, M_2, ϵ) MAC code such that:*

$$\begin{aligned}\epsilon \leq &\mathbb{P}\Big[\{i_{12}(A,B;Y) \leq \log\gamma_{12}\} \bigcup \{i_1(A;Y|B) \leq \log\gamma_1\} \bigcup \{i_2(B;Y|A) \leq \log\gamma_2\}\Big] \\ &+ (M_1-1)(M_2-1)e^{-\gamma_{12}} + (M_1-1)e^{-\gamma_1} + (M_2-1)e^{-\gamma_2}\end{aligned} \tag{26.3}$$

Proof. We again use the idea of random coding.

Generate the codebooks

$$c_1, \dots, c_{M_1} \in \mathcal{A}, \quad d_1, \dots, d_{M_2} \in \mathcal{B}$$

where the codes are drawn i.i.d from distributions: $c_1, \dots, c_{M_1} \sim$ *i.i.d.* P_A, $d_1, \dots, d_{M_2} \sim$ *i.i.d.* P_B.

The decoder operates in the following way: report (m, m') if it is the unique pair that satisfies:

$$\begin{aligned}(P_{12}) &\quad i_{12}(c_m, d_{m'}; y) > \log\gamma_{12} \\ (P_1) &\quad i_1(c_m; y|d_{m'}) > \log\gamma_1 \\ (P_2) &\quad i_2(d_{m'}; y|c_m) > \log\gamma_2\end{aligned}$$

Evaluate the expected error probability:

$$\begin{aligned}\mathbb{E}P_e(c_1^{M_1}, d_1^{M_2}) = \mathbb{P}\Big[&\{(W_1, W_2) \text{ violate } (P_{12}) \text{ or } (P_1) \text{ or } (P_2)\} \\ &\bigcup \{\exists \text{ impostor } (W_1', W_2') \text{ that satisfy } (P_{12}) \text{ and } (P_1) \text{ and } (P_2)\}\Big]\end{aligned}$$

by symmetry of random codes, we have

$$\begin{aligned}P_e = \mathbb{E}[P_e|W_1 = m, W_2 = m'] = \mathbb{P}\Big[&\{(m, m') \text{ violate } (P_{12}) \text{ or } (P_1) \text{ or } (P_2)\} \\ &\bigcup \{\exists \text{ impostor } (i \neq m, j \neq m') \text{ that satisfy } (P_{12}) \text{ and } (P_1) \text{ and } (P_2)\}\Big]\end{aligned}$$

$$\Rightarrow P_e \leq \mathbb{P}\Big[\{i_{12}(A,B;Y) \leq \log\gamma_{12}\} \bigcup \{i_1(A;Y|B) \leq \log\gamma_1\} \bigcup \{i_2(B;Y|A) \leq \log\gamma_2\}\Big] + \mathbb{P}[E_{12}] + \mathbb{P}[E_1] + \mathbb{P}[E_2]$$

where

$$
\begin{aligned}
\mathbb{P}[E_{12}] &= \mathbb{P}[\{\exists(i \neq m, j \neq m') \text{ s.t. } i_{12}(c_m, d_{m'}; y) > \log\gamma_{12}\}] \\
&\leq (M_1 - 1)(M_2 - 1)\mathbb{P}[i_{12}(\overline{A}, \overline{B}; Y) > \log\gamma_{12}] \\
&= \mathbb{E}[e^{-i_{12}(A,B;Y)}\mathbf{1}\{i_{12}(A, B; Y) > \log\gamma_{12}\}] \\
&\leq e^{-\gamma_{12}} \\
\mathbb{P}[E_2] &= \mathbb{P}[\{\exists(j \neq m') \text{ s.t. } i_2(d_j; y|c_i) > \log\gamma_2\}] \\
&\leq (M_2 - 1)\mathbb{P}[i_2(\overline{B}; Y|A) > \log\gamma_2] \\
&= \mathbb{E}_A[e^{-i_2(B;Y|A)}\mathbf{1}\{i_2(B; Y|A) > \log\gamma_2\}|A] \\
&\leq \mathbb{E}_A[e^{-\gamma_2}|A] = e^{-\gamma_2} \\
\text{similarly } \ \mathbb{P}[E_1] &\leq e^{-\gamma_1}
\end{aligned}
$$

□

Note: [Intuition] Consider the decoding step when a random codebook is used. We observe Y and need to solve an M-ary hypothesis testing problem: Which of $\{P_{Y|A=c_m,B=d_{m'}}\}_{m,m'\in[M_1]\times[M_2]}$ produced the sample Y?

Recall that in P2P channel coding, we had a similar problem and the M-ary hypothesis testing problem was converted to M binary testing problems:

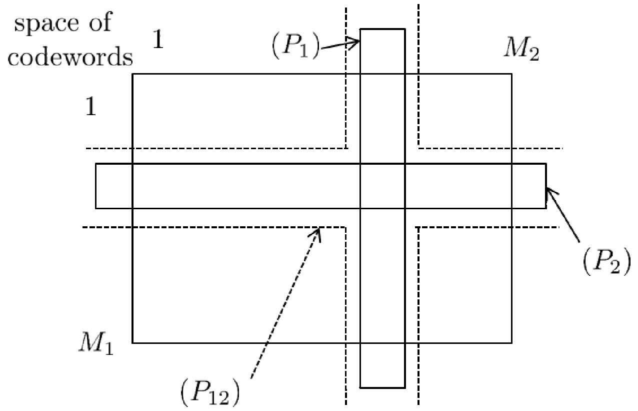
$$P_{Y|X=c_j} \quad \text{vs} \quad P_{Y_{-j}} \triangleq \sum_{i\neq j} \frac{1}{M-1} P_{Y|X=c_i} \approx P_Y$$

I.e. distinguish c_j (hypothesis H_0) against the average distribution induced by all other codewords (hypothesis H_1), which for a random coding ensemble $c_j \sim P_X$ is very well approximated by $P_Y = P_{Y|X} \circ P_X$. The optimal test for this problem is roughly

$$\frac{P_{Y|X=c}}{P_Y} \gtrsim \log(M-1) \quad \Longrightarrow \quad \text{decide } P_{Y|X=c_j} \tag{26.4}$$

since the prior for H_0 is $\frac{1}{M}$, while the prior for H_1 is $\frac{M-1}{M}$.

The proof above followed the same idea except that this time because of the two-dimensional grid structure:



there are in fact binary HT of three kinds

$$(P12)\quad \sim\quad \text{test } P_{Y|A=c_m,B=d_{m'}} \text{ vs } \frac{1}{(M_1-1)(M_2-1)}\sum_{i\neq m}\sum_{j\neq m'} P_{Y|A=c_i,B=d_j} \approx P_Y$$

$$(P1)\quad \sim\quad \text{test } P_{Y|A=c_m,B=d_{m'}} \text{ vs } \frac{1}{M_1-1}\sum_{i\neq m} P_{Y|A=c_i,B=d_{m'}} \approx P_{Y|B=d_{m'}}$$

$$(P2)\quad \sim\quad \text{test } P_{Y|A=c_m,B=d_{m'}} \text{ vs } \frac{1}{M_2-1}\sum_{j\neq m'} P_{Y|A=c_m,B=d_j} \approx P_{Y|A=c_m}$$

And analogously to (26.4) the optimal tests are given by comparing the respective information densities with $\log M_1 M_2$, $\log M_1$ and $\log M_2$.

Another observation following from the proof is that the following decoder would also achieve exactly the same performance:

- Step 1: rule out all cells (i,j) with $i_{12}(c_i, d_j; Y) \lesssim \log M_1 M_2$.
- Step 2: If the messages remaining are NOT all in one row or one column, then FAIL.
- Step 3a: If the messages remaining are all in one column m' then declare $\hat{W}_2 = m'$. Rule out all entries in that column with $i_1(c_i; Y|d_{m'}) \lesssim \log M_1$. If more than one entry remains, FAIL. Otherwise declare the unique remaining entry m as $\hat{W}_1 = m$.
- Step 3b: Similarly with column replaced by row, i_1 with i_2 and $\log M_1$ with $\log M_2$.

The importance of this observation is that in the regime when RHS of (26.3) is small, the decoder always finds it possible to basically decode one message, "subtract" its influence and then decode the other message. Which of the possibilities 3a/3b appears more often depends on the operating point (R_1, R_2) inside $\mathcal{C}$.

26.3 MAC capacity region proof

Proof. 1. Show $\mathcal{C}$ is convex.

Take $(R_1, R_2) \in \mathcal{C}_{\epsilon/2}$, and take $(R_1', R_2') \in \mathcal{C}_{\epsilon/2}$.

We merge the $(n, 2^{nR_1}, 2^{nR_2}, \epsilon/2)$ code and the $(n, 2^{nR_1}, 2^{nR_2}, \epsilon/2)$ code in the following time-sharing way: in the first n channels, use the first set of codes, and in the last n channels, use the second set of codes.

Thus we formed a new $(2n, 2^{R_1+R_1'}, 2^{R_2+R_2'}, \epsilon)$ code, we know that

$$\frac{1}{2}\mathcal{C}_{\epsilon/2} + \frac{1}{2}\mathcal{C}_{\epsilon/2} \subset \mathcal{C}_\epsilon$$

take limit at both sides

$$\frac{1}{2}\mathcal{C} + \frac{1}{2}\mathcal{C} \subset \mathcal{C}$$

also we know that $\mathcal{C} \subset \frac{1}{2}\mathcal{C} + \frac{1}{2}\mathcal{C}$, therefore $\mathcal{C} = \frac{1}{2}\mathcal{C} + \frac{1}{2}\mathcal{C}$ is convex.

Note: the set addition is defined in the following way:

$$\mathcal{A} + \mathcal{B} \triangleq \{(a+b) : a \in \mathcal{A}, b \in \mathcal{B}\}$$

2. Achievability

   STP: $\forall P_A, P_B, \forall (R_1, R_2) \in Penta(P_A, P_B)$, $\exists (n, 2^{nR_1}, 2^{nR_2}, \epsilon) code.$

   Apply Lemma 26.1 with:

$$P_A \to P_A^n, \quad P_B \to P_B^n, \quad P_{Y|A,B} \to P_{Y|A,B}^n$$
$$M_1 = 2^{nR_1}, \quad M_2 = 2^{nR_2},$$
$$\log \gamma_{12} = n(I(A,B;Y) - \delta), \quad \log \gamma_1 = n(I(A;Y|B) - \delta), \quad \log \gamma_2 = n(I(B;Y|A) - \delta).$$

   we have that there exists a (M_1, M_2, ϵ) code with

$$\epsilon \le \underbrace{\mathbb{P}\Big[\{\frac{1}{n}\sum_{k=1}^{n} i_{12}(A_k, B_k; Y_k) \le \log \gamma_{12} - \delta\} \bigcup \{\frac{1}{n}\sum_{k=1}^{n} i_1(A_k; Y_k|B_k) \le \log \gamma_1 - \delta\} \bigcup \{\frac{1}{n}\sum_{k=1}^{n} i_2(B_k; Y_k|A_k) \le \log \gamma_2 - \delta\}\Big]}_{①}$$
$$+ \underbrace{(2^{nR_1} - 1)(2^{nR_2} - 1)e^{-\gamma_{12}} + (2^{nR_1} - 1)e^{-\gamma_1} + (2^{nR_2} - 1)e^{-\gamma_2}}_{②}$$

   by WLLN, the first part goes to zero, and for any (R_1, R_2) such that $R_1 < I(A;Y|B) - \delta$ and $R_2 < I(B;Y|A) - \delta$ and $R_1 + R_2 < I(A,B;Y) - \delta$, the second part goes to zero as well. Therefore, if $(R_1, R_2) \in$ interior of the Penta, there exists a $(M_1, M_2, \epsilon = o(1))$ code.

3. Weak converse

$$\begin{tikzcd} W_1 \arrow[r, "f_1"] & A^n \arrow[dr] & & \hat{W}_1 \\ & & Y^n \arrow[ur, "g"] \arrow[dr] & \\ W_2 \arrow[r, "f_2"] & B^n \arrow[ur] & & \hat{W}_2 \end{tikzcd}$$

P

$$\begin{tikzcd} W_1 \arrow[r, "f_1"] & A^n & & \hat{W}_1 \\ & & Y^n \arrow[ur, "g"] \arrow[dr] & \\ W_2 \arrow[r, "f_2"] & B^n & & \hat{W}_2 \end{tikzcd}$$

$Q \in (*)$

$$\begin{tikzcd} W_1 \arrow[r, "f_1"] & A^n \arrow[dr] & & \hat{W}_1 \\ & & Y^n \arrow[ur, "g"] \arrow[dr] & \\ W_2 \arrow[r, "f_2"] & B^n & & \hat{W}_2 \end{tikzcd}$$

$Q_1 \in (*1)$

$$\begin{tikzcd} W_1 \arrow[r, "f_1"] & A^n & & \hat{W}_1 \\ & & Y^n \arrow[ur, "g"] \arrow[dr] & \\ W_2 \arrow[r, "f_2"] & B^n \arrow[ur] & & \hat{W}_2 \end{tikzcd}$$

$Q_2 \in (*2)$

$$\mathbb{Q}[W_1 = \hat{W}_1, W_2 = \hat{W}_2] = \frac{1}{M_1 M_2}, \quad \mathbb{P}[W_1 = \hat{W}_1, W_2 = \hat{W}_2] \ge 1 - \epsilon$$

   d-proc:

$$d(1 - \epsilon \| \frac{1}{M_1 M_2}) \le \inf_{\mathbb{Q} \in (*)} D(P \| Q) = I(A^n, B^n; Y^n)$$
$$\Rightarrow R_1 + R_2 \le \frac{1}{n} I(A^n, B^n; Y^n) + o(1)$$

To get separate bounds, we apply the same trick to evaluate the information flow from the link between $A \to Y$ and $B \to Y$ separately:

$$\mathbb{Q}_1[W_2 = \hat{W}_2] = \frac{1}{M_2}, \quad \mathbb{P}[W_2 = \hat{W}_2] \geq 1 - \epsilon$$

d-proc:

$$\begin{aligned} d(1-\epsilon \| \frac{1}{M_2}) \leq & \inf_{\mathbb{Q}_1 \in (*1)} D(P \| Q_1) = I(B^n; Y^n | A^n) \\ \Rightarrow R_2 \leq & \frac{1}{n} I(B^n; Y^n | A^n) + o(1) \end{aligned}$$

similarly we can show that

$$R_2 \leq \frac{1}{n} I(A^n; Y^n | B^n) + o(1)$$

For memoryless channels, we know that $\frac{1}{n} I(A^n, B^n; Y^n) \leq \frac{1}{n} \sum_k I(A_k, B_k; Y_k)$. Similarly, since given B^n the channel $A^n \to Y^n$ is still memoryless we have

$$I(A^n; Y^n | B^n) \leq \sum_{k=1}^n I(A_k; Y_k | B^n) = \sum_{k=1}^n I(A_k; Y_k | B_k)$$

Notice that each (A_i, B_i) pair corresponds to (P_{A_k}, P_{B_k}), $\forall k$ define

$$Penta_k(P_{A_k}, P_{B_k}) = \left\{ (R_{1,k}, R_{2,k}) : \begin{array}{r} 0 \leq R_{1,k} \leq I(A_k; Y_k | B_k) \\ 0 \leq R_{2,k} \leq I(B_k; Y_k | A_k) \\ R_{1,k} + R_{2,k} \leq I(A_k, B_k; Y_k) \end{array} \right\}$$

therefore

$$\begin{aligned} (R_1, R_2) \in & \left[\frac{1}{n} \sum_k Penta_k \right] \\ \Rightarrow C \in & \overline{co} \bigcup_{P_A, P_B} Penta \end{aligned}$$

□

§ 27. Examples of MACs. Maximal P_e and zero-error capacity.

27.1 Recap

Last time we defined the multiple access channel as the sequence of random transformations

$$\{P_{Y^n|A^nB^n} : \mathcal{A}^n \times \mathcal{B}^n \to \mathcal{Y}^n, n = 1, 2, \dots\}$$

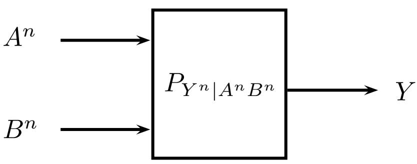
Furthermore, we showed that its capacity region is

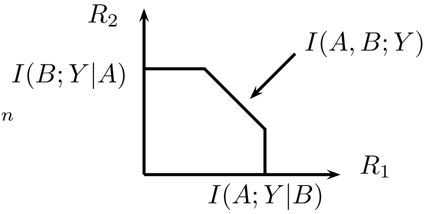
$$C = \{(R_1, R_2) : \exists (n, 2^{nR_1}, 2^{nR_2}, \epsilon) - \mathit{MAC\ code}\} = \overline{co} \bigcup_{P_A P_B} \text{Penta}(P_A, P_B)$$

were $\overline{co}$ denotes the convex hull of the sets Penta, and Penta is

$$\text{Penta}(P_A, P_B) = \begin{cases} R_1 \le I(A; Y|B) \\ R_2 \le I(B; Y|A) \\ R_1 + R_2 \le I(A, B; Y) \end{cases}$$

So a general MAC and one Penta region looks like





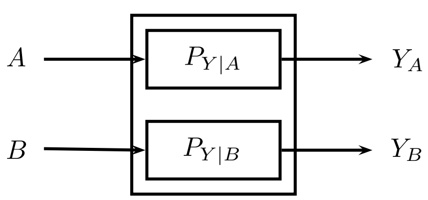
Note that the union of Pentas need not look like a Penta region itself, as we will see in a later example.

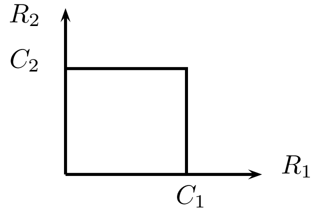
27.2 Orthogonal MAC

The trivial MAC is when each input sees its own independent channel: $P_{Y|AB} = P_{Y|A} P_{Y|B}$ where the receiver sees (Y_A, Y_B). In this situation, we expect that each transmitter can achieve it's own capacity, and no more than that. Indeed, our theorem above shows exactly this:

$$\text{Penta}(P_A, P_B) = \begin{cases} R_1 \le I(A; Y|B) = I(A; Y) \\ R_2 \le I(B; Y|A) = I(B; Y) \\ R_1 + R_2 \le I(A, B; Y) \end{cases}$$

Where in this case the last constraint is not applicable; it does not restrict the capacity region.





Hence our capacity region is a rectangle bounded by the individual capacities of each channel.

27.3 BSC MAC

Before introducing this channel, we need a definition and a theorem:

Definition 27.1 (Sum Capacity)**.** $C_{sum} \triangleq \max\{R_1 + R_2 : (R_1, R_2) \in C\}$

Theorem 27.1. $C_{sum} = \max_{A \perp\!\!\!\perp B} I(A, B; Y)$

Proof. Since the max above is achieved by an extreme point on one of the Penta regions, we can drop the convex closure operation to get

$$\max\{R_1 + R_2 : (R_1, R_2) \in \overline{co} \bigcup \text{Penta}(P_A, P_B)\} = \max\{R_1 + R_2 : (R_1, R_2) \in \bigcup \text{Penta}(P_A, P_B)\}$$
$$\max_{P_A, P_B} \{R_1 + R_2 : (R_1, R_2) \in \text{Penta}(P_A, P_B)\} \leq \max_{P_A, P_B} I(A, B; Y)$$

Where the last step follows from the definition of Penta. Now we need to show that the constraint on $R_1 + R_2$ in Penta is active at at least one point, so we need to show that $I(A, B; Y) \leq I(A; Y|B) + I(B; Y|A)$ when $A \perp\!\!\!\perp B$, which follows from applying Kolmogorov identities

$$\begin{aligned} I(A; Y, B) = 0 + I(A; Y|B) = I(A; Y) + I(A; B|Y) &\implies I(A; Y) \leq I(A; Y|B) \\ \implies I(A, B; Y) = I(A; Y) + I(B; Y|A) &\leq I(A; Y|B) + I(B; Y|A) \end{aligned}$$

Hence $\max_{P_A, P_B}\{R_1 + R_2 : (R_1, R_2) \in \text{Penta}(P_A, P_B)\} = \max_{P_A P_B} I(A, B; Y)$ □

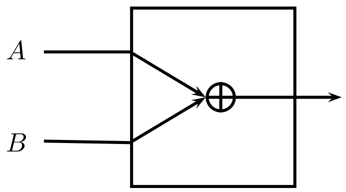
We now look at the BSC MAC, defined by

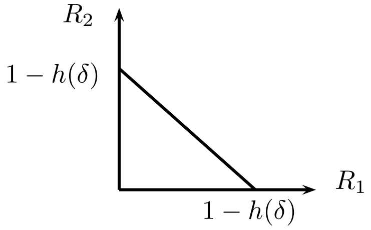
$$\begin{aligned} &Y = A + B + Z \mod 2 \\ &Z \sim \text{Ber}(\delta) \\ &A, B \in \{0, 1\} \end{aligned}$$

Since the output Y can only be 0 or 1, the capacity of this channel can be no larger than 1 bit. If B doesn't transmit at all, then A can achieve capacity $1 - h(\delta)$ (and B can achieve capacity when A doesn't transmit), so that $R_1, R_2 \leq 1 - h(\delta)$. By time sharing we can obtain any point between these two. This gives an inner bound on the capacity region. For an outer bound, we use Theorem 27.1, which gives

$$\begin{aligned} C_{sum} = \max_{P_A P_B} I(A, B; Y) &= \max_{P_A P_B} I(A, B; A + B + Z) \\ &= \max_{P_A P_B} H(A + B + Z) - H(Z) = 1 - h(\delta) \end{aligned}$$

Hence $R_1 + R_2 \leq 1 - h(\delta)$, so by this outer bound, we can do no better than time sharing between the two individual channel capacity points.





Remark: Even though this channel seems so simple, there are still hidden things about it, which we'll see later.

27.4 Adder MAC

Now we analyze the Adder MAC, which is a noiseless channel defined by:

$$\begin{aligned} &Y = A + B \quad (\text{over } \mathbb{Z}) \\ &A, B \in \{0, 1\} \end{aligned}$$

Intuitively, the game here is that when both A and B send either 0 or 1, we receiver 0 or 2 and can decode perfectly. However, when A sends 0 and B send 1, the situation is ambiguous. To analyze this channel, we start with an interesting fact

Interesting Fact 1: Any deterministic MAC ($Y = f(A, B)$) has $C_{sum} = \max H(Y)$. To see this, just expand $I(A, B; Y)$.

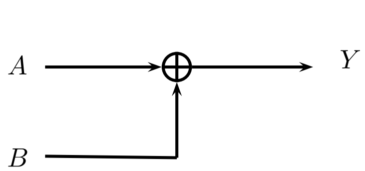
Therefore, the sum capacity of this MAC is

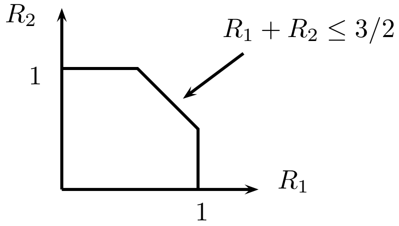
$$C_{sum} = \max_{A \perp\!\!\!\perp B} H(A + B) = H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) = \frac{3}{2} \text{ bits}$$

Which is achieved when both A and B are Ber(1/2). With this, our capacity region is

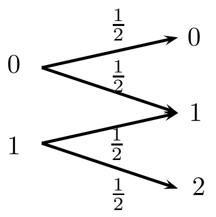
$$\text{Penta}(\text{Ber}(1/2), \text{Ber}(1/2)) = \begin{cases} R_1 \le I(A; Y|B) = H(A) = 1 \\ R_2 \le I(B; Y|A) = H(B) = 1 \\ R_1 + R_2 \le I(A, B; Y) = 3/2 \end{cases}$$

So the channel can be described by

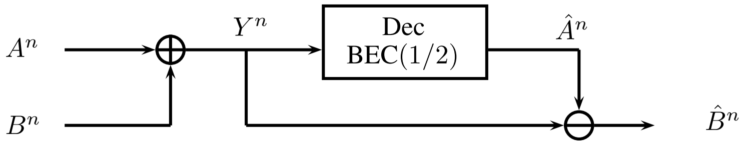




Now we can ask: how do we achieve the corner points of the region, e.g. $R_1 = 1/2$ and $R_2 = 1$? The answer gives insights into how to code for this channel. Take the greedy codebook $B = \{0, 1\}^n$ (the entire space), then the channel $A \to Y$ is a DMC:



Which we recognize as a BEC(1/2) (no preference to either −1 or 1), which has capacity 1/2. How do we decode? The idea is *successive cancellation*, where first we decode A, then remove $\hat{A}$ from Y, then decode B.



Using this strategy, we can use a single user code for the BEC (an object we understand well) to attain capacity.

27.5 Multiplier MAC

The Multiplier MAC is defined as

$$\begin{aligned}&Y = AB\\&A \in \{0,1\},\ B \in \{-1,1\}\end{aligned}$$

Note that $A = |Y|$ can always be resolved, and B can be resolved whenever $A = 1$. To find the capacity region of this channel, we'll use another interesting fact:

Interesting Fact 2: If $A = g(Y)$, then each $\text{Penta}(P_A, P_B)$ is a rectangle with

$$\text{Penta}(P_A, P_B) = \begin{cases} R_1 \leq H(A) \\ R_2 \leq I(A,B;Y) - H(A) \end{cases}$$

Proof. Using the assumption that $A = g(Y)$ and expanding the mutual information

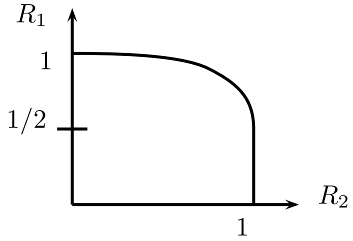
$$\begin{aligned} I(A;Y|B) + I(B;Y|A) &= H(A) - H(Y|A) - H(Y|A,B) = H(A,Y) - H(Y|A,B) \\ &= H(Y) - H(Y|A,B) = I(A,B;Y) \end{aligned}$$

Therefore the $R_1 + R_2$ constraint is not active, so our region is a rectangle. □

By symmetry, we take $P_B = \text{Ber}(1/2)$. When $P_A = \text{Ber}(p)$, the output has $H(Y) = p + h(p)$. Using the above fact, the capacity region for the Multiplier MAC is

$$C = \overline{co} \bigcup \begin{cases} R_1 \leq H(A) = h(p) \\ R_2 \leq H(Y) - H(A) = p \end{cases}$$

We can view this as the graph of the binary entropy function on its side, parametrized by p:



To achieve the extreme point $(1, 1/2)$ of this region, we can use the same scheme as for the Adder MAC: take the codebook of A to be $\{0,1\}^n$, then B sees a BEC(1/2). Again, successive cancellation decoding can be used.

For future reference we note:

Lemma 27.1. *The full capacity region of multiplier MAC is achieved with zero error.*

Proof. For a given codebook D of user B the number of messages that user A can send equals the total number of erasure patters that codebook D can tolerate with vanishing probability of error. Fix rate $R_2 < 1$ and let D be a row-span of a random linear $nR_2 \times n$ binary matrix. Then randomly erase each column with probability $1 - R_2 - \epsilon$. Since on average there will be $n(R_2 + \epsilon)$ columns left, the resulting matrix is still full-rank and the decoding is possible. In other words,

$$\mathbb{P}[D \text{ is decodable}, \# \text{ of erasures} \approx n(1 - R_2 - \epsilon)] \to 1\,.$$

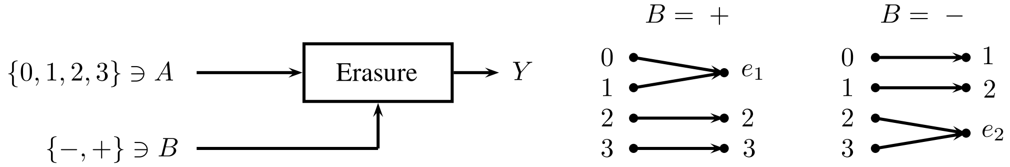
Hence, by counting the total number of erasures, for a random linear code we have

$$\mathbb{E}[\# \text{ of decodable erasure patterns for } D] \approx 2^{nh(1-R_2-\epsilon)+o(n)}\,.$$

And result follows by selecting a random element of the D-ensemble and then taking the codebook of user A to be the set of decodable erasure patterns for a selected D. □

27.6 Contraction MAC

The Contraction MAC is defined as



Here, B is received perfectly, We can use the fact above to see that the capacity region is

$$C = \begin{cases} R_1 \le \frac{3}{2} \\ R_2 \le 1 \end{cases}$$

For future reference we note the following:

Lemma 27.2. *The zero-error capacity of the contraction MAC satisfies*

$$R_1 \le h(1/3) + (2/3 - p)\log 2\,, \tag{27.1}$$

$$R_2 \le h(p) \tag{27.2}$$

for some $p \in [0, 1/2]$. *In particular, the point* $R_1 = \frac{3}{2}$, $R_2 = 1$ *is not achievable with zero error.*

Proof. Let C and D denote the zero-error codebooks of two users. Then for each string $b^n \in \{+,-\}^n$ denote

$$U_{b^n} = \{a^n : a_j \in \{0,1\} \text{ if } b_j = +, a_j \in \{2,3\} \text{ if } b_j = -\}\,.$$

Then clearly for each b^n we have

$$|U_{b^n}| \le 2^{d(b^n,D)}\,,$$

where $d(b^n, D)$ denotes the minimum Hamming distance from string b^n to the set D. Then,

$$|C| \le \sum_{b^n} 2^{d(b^n,D)} \tag{27.3}$$

$$= \sum_{j=0}^{n} 2^j |\{b^n : d(b^n, D) = j\}| \tag{27.4}$$

For a given cardinality $|D|$ the set that maximizes the above sum is the Hamming ball. Hence, $R_2 = h(p) + O(1)$ implies

$$R_2 \le \max_{q\in[p,1]} h(q) + (q-p)\log 2 = h(1/3) + (2/3-p)\log 2\,.$$

□

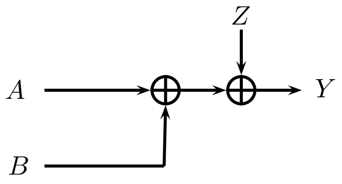
27.7 Gaussian MAC

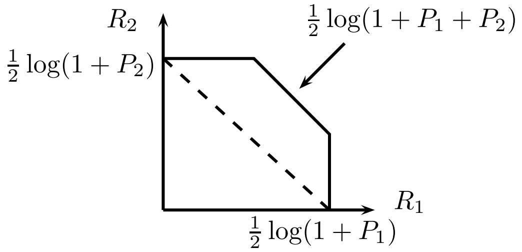
Perhaps the most important MAC is the Gaussian MAC. This is defined as

$$\begin{aligned} &Y = A + B + Z \\ &Z \sim \mathcal{N}(0,1) \\ &\mathbb{E}[A^2] \le P_1,\ \mathbb{E}[B^2] \le P_2 \end{aligned}$$

Evaluating the mutual information, we see that the capacity region is

$$\begin{aligned} I(A;Y|B) &= I(A;A+Z) \le \frac{1}{2}\log(1+P_1) \\ I(B;Y|A) &= I(B;B+Z) \le \frac{1}{2} \le (1+P_2) \\ I(A,B;Y) &= h(Y) - h(Z) \le \frac{1}{2}\log(1+P_1+P_2) \end{aligned}$$





Where the region is Penta$(\mathcal{N}(0,P_1),\mathcal{N}(0,P_2))$. How do we achieve the rates in this region? We'll look at a few schemes.

1. TDMA: A and B switch off between transmitting at full rate and not transmitting at all. This achieves any rate pair in the form

   $$R_1 = \lambda\frac{1}{2}\log(1+P_1), \quad R_2 = \bar{\lambda}\frac{1}{2}\log(1+P_2)$$

   Which is the dotted line on the plot above. Clearly, there are much better rates to be gained by smarter schemes.

2. FDMA (OFDM): Dividing users into different frequency bands rather than time windows gives an enormous advantage. Using frequency division, we can attain rates

   $$R_1 = \lambda\frac{1}{2}\log\left(1+\frac{P_1}{\lambda}\right), \quad R_2 = \bar{\lambda}\frac{1}{2}\log\left(1+\frac{P_2}{\bar{\lambda}}\right)$$

In fact, these rates touch the boundary of the capacity region at its intersection with the $R_1 = R_2$ line. The optimal rate occurs when the power at each transmitter makes the noise look white:

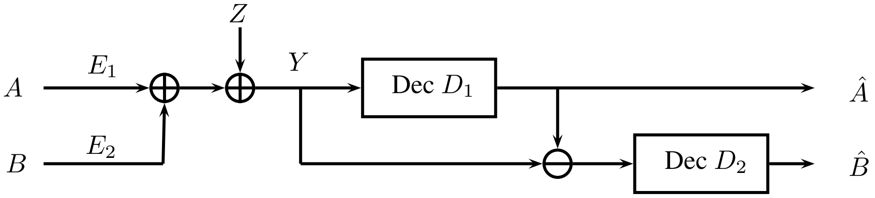
$$\frac{P_1}{\lambda} = \frac{P_2}{\bar{\lambda}} \implies \lambda^* = \frac{P_1}{P_1 + P_2}$$

While this touches the capacity region at one point, it doesn't quite reach the corner points. Note, however, that practical systems (e.g. cellular networks) typically employ *power control* that ensures received powers P_i of all users are roughly equal. In this case (i.e. when $P_1 = P_2$) the point where FDMA touches the capacity boundary is at a very desirable location of *symmetric rate* $R_1 = R_2$. This is one of the reasons why modern standards (e.g. LTE 4G) do not employ any specialized MAC-codes and use OFDM together with good single-user codes.

3. Rate Splitting/Successive Cancellation: To reach the corner points, we can use successive cancellation, similar to the decoding schemes in the Adder and Multiplier MACs. We can use rates:

$$\begin{aligned} R_2 &= \frac{1}{2}\log(1 + P_2) \\ R_1 &= \frac{1}{2}(\log(1 + P_1 + P_2) - \log(1 + P_2)) = \frac{1}{2}\log\left(1 + \frac{P_1}{1 + P_2}\right) \end{aligned}$$

The second expression suggests that A transmits at a rate for an AWGN channel that has power constraint P_1 and noise $1 + P_2$, i.e. the power used by B looks like noise to A.



Theorem 27.2. *There exists a successive-cancellation code (i.e.* (E_1, E_2, D_1, D_2)*) that achieves the corner points of the Gaussian MAC capacity region.*

Proof. Random coding: $B^n \sim \mathcal{N}(0, P_2)^n$. Since A^n now sees noise $1 + P_2$, there exists a code for A with rate $R_1 = \frac{1}{2}\log(1 + P_1/(1 + P_2))$. □

This scheme (unlike the above two) can tolerate frame un-synchronization between the two transmitters. This is because any chunk of length n has distribution $\mathcal{N}(0, P_2)^n$. It has generalizations to non-corner points and to arbitrary number of users. See [RU96] for details.

27.8 MAC Peculiarities

Now that we've seen some nice properties and examples of MACs, we'll look at cases where MACs differ from the point to point channels we've seen so far.

1. Max probability of error ≠ average probability of error.

Theorem 27.3. $C^{(max)} \neq C$

Proof. The key observation for deterministic MAC is that $C^{(max)} = C_0$ (zero error capacity) when $\epsilon \leq 1/2$. This is because when any two strings can be confused, the maximum probability of error

$$\max_{m,m'} \mathbb{P}[\hat{W}_1 \neq m \cup \hat{W}_2 \neq m' | W_1 = m, W_2 = m']$$

Must be larger than 1/2. □

For some of the channels we've seen

- Contraction MAC: $C_0 \neq C$
- Multiplier MAC: $C_0 = C$
- Adder MAC: $C_0 \neq C$. For this channel, no one yet can show that $C_{0,sum} < 3/2$. The idea is combinatorial in nature: produce two sets (Sidon sets) such that all pairwise sums between the two do not overlap.

2. Separation does not hold: In the point to point channel, through joint source channel coding we saw that an optimal architecture is to do source coding then channel coding separately. This doesn't hold for the MAC. Take as a simple example the Adder MAC with a correlated source and bandwidth expansion factor $\rho = 1$. Let the source (S, T) have joint distribution

$$P_{ST} = \begin{bmatrix} 1/3 & 1/3 \\ 0 & 1/3 \end{bmatrix}$$

We encode S^n to channel input A^n and T^n to channel input B^n. The simplest possible scheme is to not encoder at all; simply take $S_j = A_j$ and $T_j = B_j$. Take the decoder

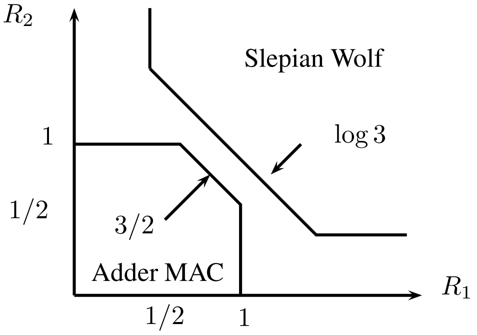
$$\begin{array}{rcc} & \hat{S} & \hat{T} \\ Y_j = 0 \implies & 0 & 0 \\ Y_j = 1 \implies & 0 & 1 \\ Y_j = 2 \implies & 1 & 1 \end{array}$$

Which gives $\mathbb{P}[\hat{S}^n = S^n, \hat{T}^n = T^n] = 1$, since we are able to take advantage of the zero entry in joint distribution of our correlated source.

Can we achieve this with a separated source? Amazingly, even though the above scheme is so simple, we can't! The compressors in the separated architecture operate in the Slepian Wolf region

$$\begin{cases} R_1 \geq H(S|T) \\ R_2 \geq H(T|S) \\ R_1 + R_2 \geq H(S,T) = \log 3 \end{cases}$$

Hence the sum rate for compression must be $\geq \log 3$, while the sum rate for the Adder MAC must be $\leq 3/2$, so these two regions do not overlap, hence we can not operate at a bandwidth expansion factor of 1 for this source and channel.



3. Linear codes beat generic ones: Consider a BSC-MAC and suppose that two users A and B have independet k-bit messages $W_1, W_2 \in \mathbb{F}_2^k$. Suppose the receiver is only interested in estimating $W_1 + W_2$. What is the largest ratio k/n? Clearly, separation can achieve

$$k/n \approx \frac{1}{2}(\log 2 - h(\delta))$$

by simply creating a scheme in which both W_1 and W_2 are estimated and then their sum is computed.

A more clever solution is however to encode

$$\begin{aligned} A^n &= G \cdot W_1\,, \\ B^n &= G \cdot W_2\,, \\ Y^n &= A^n + B^n + Z^n = G(W_1 + W_2) + Z^n\,. \end{aligned}$$

where G is a generating matrix of a good k-to-n linear code. Then, provided that

$$k < n(\log 2 - h(\delta)) + o(n)$$

the sum $W_1 + W_2$ is decodable (see Theorem 16.2). Hence even for a simple BSC-MAC there exist clever ways to exceed MAC capacity for certain scenarios. Note that this "distributed computation" can also be viewed as lossy source coding with a distortion metric that is only sensitive to discrepancy between $W_1 + W_2$ and $\hat{W}_1 + \hat{W}_2$.

4. Dispersion is unknown: We have seen that for the point-to-point channel, not only we know the capacity, but the next-order terms (see Theorem 20.2). For the MAC-channel only the capacity is known. In fact, let us define

$$R^*_{sum}(n,\epsilon) \triangleq \sup\{R_1 + R_2 : (R_1, R_2) \in \mathcal{R}^*(n,\epsilon)\}\,.$$

Now, take Adder-MAC as an example. A simple exercise in random-coding with $P_A = P_B = \text{Ber}(1/2)$ shows

$$R^*_{sum}(n,\epsilon) \geq \frac{3}{2}\log 2 - \sqrt{\frac{1}{4n}}Q^{-1}(\epsilon)\log 2 + O(\frac{\log n}{n})\,.$$

In the converse direction the situation is rather sad. In fact the best bound we have is only slightly better than the Fano's inequality [?]. Namely for each $\epsilon > 0$ there is a constant $K_\epsilon > 0$ such that

$$R^*_{sum}(n,\epsilon) \leq \frac{3}{2}\log 2 + K_\epsilon \frac{\log n}{\sqrt{n}}\,.$$

So it is not even known if sum-rate approaches sum-capacity from above or from below as $n \to \infty$! What is even more surprising, is that the dependence of the residual term on ϵ is not clear at all. In fact, despite the decades of attempts, even for $\epsilon = 0$ the best known bound to date is just the Fano's inequality(!)

$$R^*_{sum}(n, 0) \le \frac{3}{2}\,.$$

Let's play the following game: Given a stream of Bern(p) bits, with *unknown* p, we want to turn them into pure random bits, i.e., independent fair coin flips Bern(1/2). Our goal is to find a universal way to extract the most number of bits.

In 1951 von Neumann proposed the following scheme: Divide the stream into pairs of bits, output 0 if 10, output 1 if 01, otherwise do nothing and move to the next pair. Since both 01 and 10 occur with probability pq (where $q = 1 - p$), regardless of the value of p, we obtain fair coin flips at the output. To measure the efficiency of von Neumann's scheme, note that, on average, we have $2n$ bits in and $2pqn$ bits out. So the efficiency (rate) is pq. The question is: Can we do better?

Several variations:

1. Universal v.s. non-universal: know the source distribution or not.
2. Exact v.s. approximately fair coin flips: in terms of total variation or Kullback-Leibler divergence

We only focus on the universal generation of exactly fair coins.

28.1 Setup

Recall from Lecture 6 that $\{0,1\}^* = \cup_{k\geq 0}\{0,1\}^k = \{\varnothing, 0, 1, 00, 01, \dots\}$ denotes the set of all finite-length binary strings, where $\varnothing$ denotes the empty string. For any $x \in \{0,1\}^*$, let $l(x)$ denote the length of x.

Let's first introduce the definition of random number generator formally. If the input vector is X^n, denote the output (variable-length) vector by $Y \in \{0,1\}^*$. Then the desired property of Y is the following: Conditioned on the length of Y being k, Y is uniformly distributed on $\{0,1\}^k$.

Definition 28.1 (Extractor)**.** We say $\Psi : \{0,1\}^* \to \{0,1\}^*$ an *extractor* if

1. $\Psi(x)$ is a prefix of $\Psi(y)$ if x is a prefix of y.
2. For any n and any $p \in (0,1)$, if $X^n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$, then $\Psi(X^n) \sim \text{Bern}(1/2)^k$ conditioned on $l(\Psi(X^n)) = k$.

The *rate* of Ψ is

$$r_\Psi(p) = \limsup_{n\to\infty} \frac{\mathbb{E}[l(\Psi(X^n))]}{n}, \quad X^n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p).$$

Note that the von Neumann scheme above defines a valid extractor Ψ_{vN} (with $\Psi_{\text{vN}}(x^{2n+1}) = \Psi_{\text{vN}}(x^{2n})$), whose rate is $r_{\text{vN}}(p) = pq$.

28.2 Converse

No extractor has a rate higher than the binary entropy function. The proof is simply data processing inequality for entropy and the converse holds even if the extractor is allowed to be non-universal (depending on p).

Theorem 28.1. *For any extractor Ψ and any $p \in (0,1)$,*

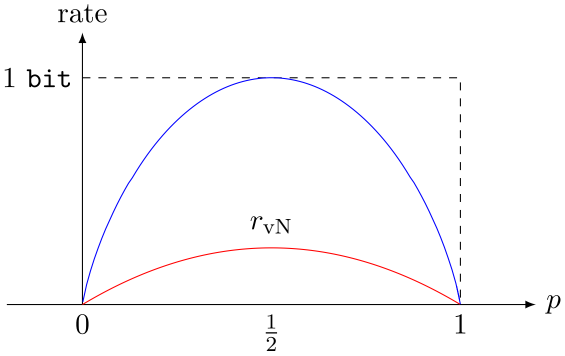
$$r_\Psi(p) \geq h(p) = p \log_2 \frac{1}{p} + q \log_2 \frac{1}{q}.$$

Proof. Let $L = \Psi(X^n)$. Then

$$nh(p) = H(X^n) \geq H(\Psi(X^n)) = H(\Psi(X^n)|L) + H(L) \geq H(\Psi(X^n)|L) = \mathbb{E}[L] \quad \texttt{bits},$$

where the step follows from the assumption on Ψ that $\Psi(X^n)$ is uniform over $\{0,1\}^k$ conditioned on $L = k$. □

The rate of von Neumann extractor and the entropy bound are plotted below. Next we present two extractors, due to Elias [Eli72] and Peres [Per92] respectively, that attain the binary entropy function. (More precisely, both ideas construct a sequence of extractors whose rate approaches the entropy bound).



28.3 Elias' construction of RNG from lossless compressors

The intuition behind Elias' scheme is the following:

1. For iid X^n, the probability of each string only depends its *type*, i.e., the number of 1's. Therefore conditioned on the number of ones, X^n is uniformly distributed (over the type class). This observation holds universally for any p.

2. Given a uniformly distributed random variable on some finite set, we can easily turn it into *variable-length* fair coin flips. For example, if U is uniform over $\{1,2,3\}$, we can map $1 \mapsto \varnothing, 2 \mapsto 0$ and $3 \mapsto 1$.

Lemma 28.1. *Given U uniformly distributed on $[M]$, there exists $f : [M] \to \{0,1\}^*$ such that conditioned on $l(f(U)) = k$, $f(U)$ is uniformly over $\{0,1\}^k$. Moreover,*

$$\log_2 M - 4 \leq \mathbb{E}[l(f(U))] \leq \log_2 M \quad \textit{bits}.$$

Proof. We defined f by partitioning $[M]$ into subsets whose cardinalities are powers of two, and assign elements in each subset to binary strings of that length. Formally, denote the binary expansion of M by $M = \sum_{i=0}^{n} m_i 2^i$, where the most significant bit $m_n = 1$ and $n = \lfloor \log_2 M \rfloor + 1$. Those non-zero m_i's defines a partition $[M] = \cup_{j=0}^{t} M_j$, where $|M_i| = 2^{i_j}$. Map the elements of M_j to $\{0,1\}^{i_j}$.

To prove the bound on the expected length, the upper bound follows from the same entropy argument $\log_2 M = H(U) \geq H(f(U)) \geq H(f(U)|l(f(U))) = \mathbb{E}[l(f(U))]$, and the lower bound follows from

$$\mathbb{E}[l(f(U))] = \frac{1}{M}\sum_{i=0}^{n} m_i 2^i \cdot i = n - \frac{1}{M}\sum_{i=0}^{n} m_i 2^i (n-i) \geq n - \frac{2^n}{M}\sum_{i=0}^{n} 2^{i-n}(n-i) \geq n - \frac{2^{n+1}}{M} \geq n-4,$$

where the last step follows from $n \leq \log_2 M + 1$. □

Elias' extractor. Let $w(x^n)$ define the Hamming weight (number of ones) of a binary string. Let $T_k = \{x^n \in \{0,1\}^n : w(x^n) = k\}$ define the Hamming sphere of radius k. For each $0 \leq k \leq n$, we apply the function f from Lemma 28.1 to each T_k. This defines a mapping $\Psi_{\mathrm{E}} : \{0,1\}^n \to \{0,1\}^*$ and then we extend it to $\Psi_{\mathrm{E}} : \{0,1\}^n \to \{0,1\}^*$ by applying the mapping per n-bit block and discard the last incomplete block. Then it is clear that the rate is given by $\frac{1}{n}\mathbb{E}[l(\Psi_{\mathrm{E}})(X^n)]$. By Lemma 28.1, we have

$$\mathbb{E}\log\binom{n}{w(X^n)} - 4 \leq \mathbb{E}[l(\Psi_{\mathrm{E}})(X^n)] \leq \mathbb{E}\log\binom{n}{w(X^n)}$$

Using Stirling's expansion (see, e.g., [Ash65, Lemma 4.7.1]), we have $\frac{2^{nh(p)}}{\sqrt{8npq}} \leq \binom{n}{k} \leq \frac{2^{nh(p)}}{\sqrt{2\pi npq}}$ where $p = 1-q = k/n \in (0,1)$ and hence $\mathbb{E}[l(\Psi_{\mathrm{E}})(X^n)] = nh(p) + O(\log n)$. Therefore the extraction rate approaches $h(p)$ as $n \to \infty$.

28.4 Peres' iterated von Neumann's scheme

Main idea: Recycle the bits thrown away in von Neumann's scheme and iterate. What did von Neumann's extractor discard: (a) bits from equal pairs. (b) location of the distinct pairs. To achieve the entropy bound, we need to extract the randomness out of these two parts as well.

First some notations: Given x^{2n}, let $k = l(\Psi_{\mathrm{vN}}(x^{2n}))$ denote the number of consecutive distinct bit-pairs.

- Let $1 \leq m_1 < \ldots < m_k \leq n$ denote the locations such that $x_{2m_j} \neq x_{2m_j - 1}$.
- Let $1 \leq i_1 < \ldots < i_{n-k} \leq n$ denote the locations such that $x_{2i_j} = x_{2i_j - 1}$.
- $y_j = x_{2m_j}$, $v_j = x_{2i_j}$, $u_j = x_{2j} \oplus x_{2j+1}$.

Here y^k are the bits that von Neumann's scheme outputs and both v^{n-k} and u^n are discarded. Note that u^n is important because it encodes the location of the y^k and contains a lot of information. Therefore von Neumann's scheme can be improved if we can extract the randomness out of both v^{n-k} and u^n.

Peres' extractor: For each $t \in \mathbb{N}$, recursively define an extractor Ψ_t as follows:

- Set Ψ_1 to be von Neumann's extractor Ψ_{vN}, i.e., $\Psi_1(x^{2n+1}) = \Psi_1(x^{2n}) = y^k$.
- Define Ψ_t by $\Psi_t(x^{2n}) = \Psi_t(x^{2n+1}) = (\Psi_1(x^{2n}), \Psi_{t-1}(u^n), \Psi_{t-1}(v^{n-k}))$.

Example: Input $x = 100111010011$ of length $2n = 12$. Output recursively:

$$\begin{array}{l}(\underline{011})(110100)(101)\\ \qquad(\underline{1})(010)(10)(\underline{0})\\ \qquad(\underline{1})(\underline{0})\end{array}$$

Note that the bits that enter into the iteration are longer iid. To compute the rate of Ψ_t, it is convenient to introduce the notion of exchangeability. We say X^n are *exchangeable* if the joint distribution is invariant under permutation, that is, $P_{X_1,\dots,X_n} = P_{X_{\pi(1)},\dots,X_{\pi(n)}}$ for any permutation π on $[n]$. In particular, if X_i's are binary, then X^n are exchangeable if and only if the joint distribution only depends on the *Hamming weight*, i.e., $P_{X^n=x^n} = p(w(x^n))$. Examples: X^n is iid Bern(p); X^n is uniform over the Hamming sphere T_k.

Lemma 28.2 (Ψ_t preserves exchangebility)**.** *Let X^{2n} be exchangeable and $L = \Psi_1(X^{2n})$. Then conditioned on $L = k$, Y^k, U^n and V^{n-k} are independent and exchangeable. Furthermore, $Y^k \overset{i.i.d.}{\sim}$ Bern$(\frac{1}{2})$ and U^n is uniform over T_k.*

Proof. If suffices to show that $\forall y, y' \in \{0,1\}^k, u, u' \in T_k$ and $v, v' \in \{0,1\}^{n-k}$ such that $w(v) = w(v')$, $\mathbb{P}[Y^k = y, U^n = u, V^{n-k} = v | L = k] = \mathbb{P}[Y^k = y', U^n = u', V^{n-k} = v' | L = k]$. Note that X^{2n} and the triple (Y^k, U^n, V^{n-k}) are in one-to-one correspondence of each other (to reconstruct X^{2n}, simply read the k distinct pairs from Y and fill them according to the locations ones in U and fill the remaining equal pairs from V). Finally, note that u, y, v and y', u', v' correspond to two input strings x and x' of identical Hamming weight ($w(x) = 2k + 2w(v)$) and hence of identical probability due to the exchangeability of X^{2n}. [Examples: $(y,u,v) = (01, 1100, 01) \Rightarrow x = (10010011)$, $(y,u,v) = (11, 1010, 10) \Rightarrow x' = (01110100)$.]

Computing the marginals, we conclude that both Y^k and U^n are uniform over their respective support set.[1] □

Lemma 28.3 (Ψ_t is an extractor)**.** *Let X^{2n} be exchangeable. Then $\Psi_t(X^{2n}) \overset{i.i.d.}{\sim}$ Bern$(1/2)$ conditioned on $l(\Psi_t(X^{2n})) = m$.*

Proof. Note that $\Psi_t(X^{2n}) \in \{0,1\}^*$. It is equivalent to show that for all $s^m \in \{0,1\}^m$,

$$\mathbb{P}[\Psi_t(X^{2n}) = s^m] = 2^{-m}\mathbb{P}[l(\Psi_t(X^{2n})) = m].$$

Proceed by induction on t. The base case of $t = 1$ follows from Lemma 28.2 (the distribution of the Y part). Assume Ψ_{t-1} is an extractor. Recall that $\Psi_t(X^{2n}) = (\Psi_1(X^{2n}), \Psi_{t-1}(U^n), \Psi_{t-1}(V^{n-k}))$ and write the length as $L = L_1 + L_2 + L_3$, where $L_2 \perp\!\!\!\perp L_3 | L_1$ by Lemma 28.2. Then

$$\begin{aligned}
&\mathbb{P}[\Psi_t(X^{2n}) = s^m]\\
&= \sum_{k=0}^{m} \mathbb{P}[\Psi_t(X^{2n}) = s^m | L_1 = k]\mathbb{P}[L_1 = k]\\
&\overset{\text{Lemma 28.2}}{=} \sum_{k=0}^{m}\sum_{r=0}^{m-k} \mathbb{P}[L_1 = k]\mathbb{P}[Y^k = s^k | L_1 = k]\mathbb{P}[\Psi_{t-1}(U^n) = s_{k+1}^{k+r} | L_1 = k]\mathbb{P}[\Psi_{t-1}(V^{n-k}) = s_{k+r+1}^{m} | L_1 = k]\\
&\overset{\text{induction}}{=} \sum_{k=0}^{m}\sum_{r=0}^{m-k} \mathbb{P}[L_1 = k]2^{-k}2^{-r}\mathbb{P}[L_2 = r | L_1 = k]2^{-(m-k-r)}\mathbb{P}[L_3 = m-k-r | L_1 = k]\\
&= 2^{-m}\mathbb{P}[L = m].
\end{aligned}$$

□

[1] If X^{2n} is iid Bern(p), then V^{n-k} is iid Bern$(p^2/(p^2+q^2))$, since $L \sim$ Binom$(n, 2pq)$ and $\mathbb{P}[Y^k = y, U^n = u, V^{n-k} = v | L = k] = 2^{-k} \cdot \binom{n}{k}^{-1} \cdot (\frac{p^2}{p^2+q^2})^m (\frac{q^2}{p^2+q^2})^{n-k-m}$, where $m = w(v)$. In general we cannot say much more than the fact that V^{n-k} is exchangeable.

Next we compute the rate of Ψ_t. Let $X^{2n} \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$. Then by SLLN, $\frac{1}{2n} l(\Psi_1(X^{2n})) \triangleq \frac{L_n}{2n}$ converges a.s. to pq. Assume that $\frac{1}{2n} l(\Psi_{t-1}(X^{2n})) \xrightarrow{\text{a.s.}} r_{t-1}(p)$. Then

$$\frac{1}{2n} l(\Psi_{t-1}(X^{2n})) = \frac{L_n}{2n} + \frac{1}{2n} l(\Psi_{t-1}(U^n)) + \frac{1}{2n} l(\Psi_{t-1}(V^{n-L_n})).$$

Note that $U^n \overset{\text{i.i.d.}}{\sim} \text{Bern}(2pq)$, $V^{n-L_n}|L_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(p^2/(p^2+q^2))$ and $L_n \xrightarrow{\text{a.s.}} \infty$. Then the induction hypothesis implies that $\frac{1}{2n} l(\Psi_{t-1}(U^n)) \xrightarrow{\text{a.s.}} r_{t-1}(2pq)$ and $\frac{1}{2(n-L_n)} l(\Psi_{t-1}(V^{n-L_n})) \xrightarrow{\text{a.s.}} r_{t-1}(p^2/(p^2+q^2))$. We obtain the recursion:

$$r_t(p) = pq + \frac{1}{2} r_{t-1}(2pq) + \frac{p^2+q^2}{2} r_{t-1}\left(\frac{p^2}{p^2+q^2}\right) \triangleq (Tr_{t-1})(p), \tag{28.1}$$

where the operator T maps a continuous function on $[0,1]$ to another. Note that $f \le g$ pointwise then $Tf \le Tg$. Then it can be shown that r_t converges monotonically from below to the fixed-point of T, which turns out to be exactly the binary entropy function h. Instead of directly verifying $Th = h$, next we give a simple proof: Consider $X_1, X_2 \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$. Then $2h(p) = H(X_1, X_2) = H(X_1 \oplus X_2, X_1) = H(X_1 \oplus X_2) + H(X_1 | X_1 \oplus X_2) = h(p^2+q^2) + 2pqh(\frac{1}{2}) + (p^2+q^2)h(\frac{p^2}{p^2+q^2})$.

The convergence of r_t to h are shown in Fig. 28.1.

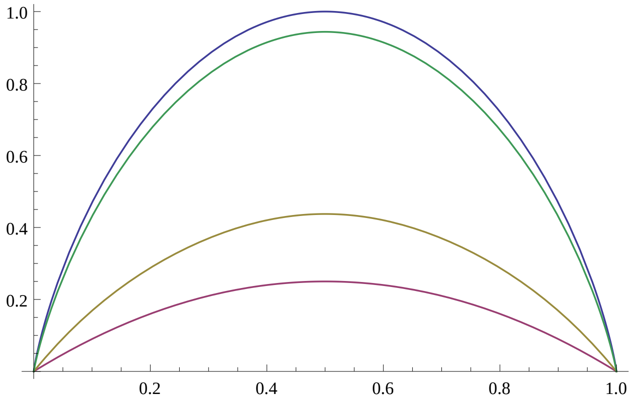


Figure 28.1: Rate function r_t for $t = 1, 4, 10$ versus the binary entropy function.

28.5 Bernoulli factory

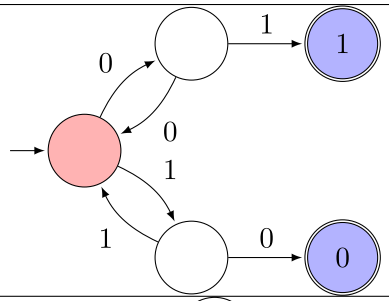
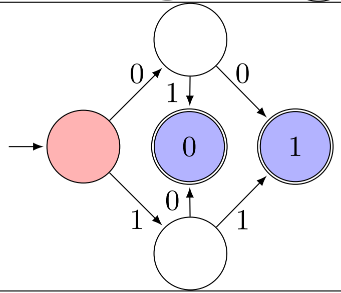
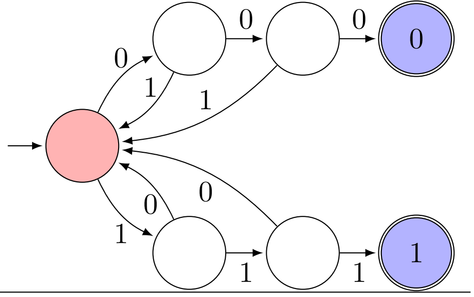
Given a stream of Bern(p) bits with unknown p, for what kind of function $f : [0,1] \to [0,1]$ can we simulate iid bits from Bern($f(p)$). Our discussion above deals with $f(p) \equiv \frac{1}{2}$. The most famous example is whether we can simulate Bern($2p$) from Bern(p), i.e., $f(p) = 2p \wedge 1$. Keane and O'Brien [KO94] showed that all f that can be simulated are either constants or "polynomially bounded away from 0 or 1": for all $0 < p < 1$, $\min\{f(p), 1 - f(p)\} \ge \{p, 1-p\}^n$ for some $n \ge 1$. In particular, doubling the bias is impossible.

The above result deals with what $f(p)$ can be simulated in principle. What type of computational devices are needed for such as task? Note that since $r_1(p)$ is quadratic in p, all rate functions r_t that arise from the iteration (28.1) are rational functions (ratios of polynomials), converging to the binary entropy function as Fig. 28.1 shows. It turns out that for any rational function f that satisfies $0 < f < 1$ on $(0,1)$, we can generate independent Bern($f(p)$) from Bern(p) using either of the following schemes [MP05]:

1. *Finite-state machine* (FSM): initial state (red), intermediate states (white) and final states (blue, output 0 or 1 then reset to initial state).

2. *Block simulation*: let A_0, A_1 be disjoint subsets of $\{0,1\}^k$. For each k-bit segment, output 0 if falling in A_0 or 1 if falling in A_1. If neither, discard and move to the next segment. The block size is at most the degree of the denominator polynomial of f.

The next table gives examples of these two realizations:

Goal	Block simulation	FSM
$f(p) = 1/2$	$A_0 = 10; A_1 = 01$	
$f(p) = 2pq$	$A_0 = 00, 11; A_1 = 01, 10$	
$f(p) = \frac{p^3}{p^3+q^3}$	$A_0 = 000; A_1 = 111$	

Exercise: How to generate $f(p) = 1/3$?

It turns out that the only type of f that can be simulated using either FSM or block simulation is rational function. For $f(p) = \sqrt{p}$, which satisfies Keane-O'Brien's characterization, it cannot be simulated by FSM or block simulation, but it can be simulated by pushdown automata (PDA), which are FSM operating with a stack [MP05].

What is the optimal Bernoulli factory with the best rate is unclear. Clearly, a converse is the entropy bound $\frac{h(p)}{h(f(p))}$, which can be trivial (bigger than one).

28.6 Related problems

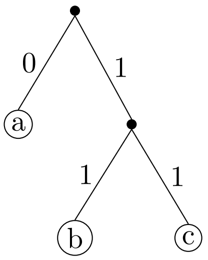
28.6.1 Generate samples from a given distribution

The problem of how to turn pure bits into samples of a given distribution P is in a way the opposite direction of what we have been considering so far. This can be done via Knuth-Yao's tree algorithm:

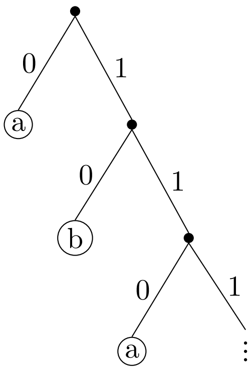
Starting at the root, flip a fair coin for each edge and move down the tree until reaching a leaf node and outputting the symbol. Let L denote the number of flips, which is a random variable. Then $H(P) \le \mathbb{E}[L] \le H(P) + 2\texttt{bits}$.

Examples:

- To generate $P = [1/2, 1/4, 1/4]$ on $\{a, b, c\}$, use the finite tree: $\mathbb{E}[L] = 1.5$.



- To generate $P = [1/3, 2/3]$ on $\{a, b\}$ (note that $2/3 = 0.1010\ldots, 1/3 = 0.0101\ldots$), use the infinite tree: $\mathbb{E}[L] = 2$ (geometric distribution)



28.6.2 Approximate random number generator

The goal is to design $f : \mathcal{X}^n \to \{0,1\}^k$ s.t. $f(X^n)$ is *close to* fair coin flips in distribution in certain distances (TV or KL). One formulation is that $D(P_{f(X^n)} \| \text{Uniform}) = o(k)$.

Intuitions: The connection to lossless data compression is as follows: A good compressor squeezes out all the redundancy of the source. Therefore its output should be close to pure bits, otherwise we can compress it furthermore. So good lossless compressors should act like good approximate random number generators.

Bibliography

[Ahl82] Rudolf Ahlswede. An elementary proof of the strong converse theorem for the multiple-access channel. *J. Combinatorics, Information and System Sciences*, 7(3), 1982.

[AN07] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[AS08] Noga Alon and Joel H. Spencer. *The Probabilistic Method.* John Wiley & Sons, 3rd edition, 2008.

[Ash65] Robert B. Ash. *Information Theory.* Dover Publications Inc., New York, NY, 1965.

[BNO03] Dimitri P Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar. *Convex analysis and optimization.* Athena Scientific, Belmont, MA, USA, 2003.

[Boh38] H. F. Bohnenblust. Convex regions and projections in Minkowski spaces. *Ann. Math.*, 39(2):301–308, 1938.

[Bro86] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. In S. S. Gupta, editor, *Lecture Notes-Monograph Series*, volume 9. Institute of Mathematical Statistics, Hayward, CA, 1986.

[CB90] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of bayes methods. *IEEE Trans. Inf. Theory*, 36(3):453–471, 1990.

[CB94] Bertrand S Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

[Ç11] Erhan Çinlar. *Probability and Stochastics.* Springer, New York, 2011.

[Cho56] Noam Chomsky. Three models for the description of language. *IRE Trans. Inform. Th.*, 2(3):113–124, 1956.

[CK81] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic, New York, 1981.

[CS83] J. Conway and N. Sloane. A fast encoding method for lattice codes and quantizers. *IEEE Transactions on Information Theory*, 29(6):820–824, Nov 1983.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.

[Doo53] Joseph L. Doob. *Stochastic Processes.* New York Wiley, 1953.

[Eli55] Peter Elias. Coding for noisy channels. *IRE Convention Record*, 3:37–46, 1955.

[Eli72] P. Elias. The Efficient Construction of an Unbiased Random Sequence. *Annals of Mathematical Statistics*, 43(3):865–870, 1972.

[ELZ05] Uri Erez, Simon Litsyn, and Ram Zamir. Lattices which are good for (almost) everything. *IEEE Transactions on Information Theory*, 51(10):3401–3416, Oct. 2005.

[EZ04] U. Erez and R. Zamir. Achieving $\frac{1}{2}\log(1+\mathrm{SNR})$ on the AWGN channel with lattice encoding and decoding. *IEEE Trans. Information Theory*, IT-50:2293–2314, Oct. 2004.

[For89] Jr. Forney, G.D. Multidimensional constellations. II. voronoi constellations. *IEEE Journal on Selected Areas in Communications*, 7(6):941–958, Aug 1989.

[Har] Sergiu Hart. Overweight puzzle. http://www.ma.huji.ac.il/~hart/puzzle/overweight.html.

[HV11] P. Harremoës and I. Vajda. On pairs of f-divergences and their joint range. *IEEE Trans. Inf. Theory*, 57(6):3230–3235, Jun. 2011.

[KO94] M.S. Keane and G.L. O'Brien. A Bernoulli factory. *ACM Transactions on Modeling and Computer Simulation*, 4(2):213–219, 1994.

[Kos63] VN Koshelev. Quantization with minimal entropy. *Probl. Pered. Inform*, (14):151–156, 1963.

[Loe97] Hans-Andrea Loeliger. Averaging bounds for lattices and linear codes. *IEEE Transactions on Information Theory*, 43(6):1767–1773, Nov. 1997.

[MP05] Elchanan Mossel and Yuval Peres. New coins from old: computing with unknown bias. *Combinatorica*, 25(6):707–724, 2005.

[OE15] O. Ordentlich and U. Erez. A simple proof for the existence of "good" pairs of nested lattices. *IEEE Transactions on Information Theory*, Submitted Aug. 2015.

[OPS48] BM Oliver, JR Pierce, and CE Shannon. The philosophy of pcm. *Proceedings of the IRE*, 36(11):1324–1331, 1948.

[Per92] Yuval Peres. Iterating von Neumann's procedure for extracting random bits. *Annals of Statistics*, 20(1):590–597, 1992.

[PPV10] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, May 2010.

[PPV11] Y. Polyanskiy, H. V. Poor, and S. Verdú. Minimum energy to send k bits with and without feedback. *IEEE Trans. Inf. Theory*, 57(8):4880–4902, August 2011.

[PW14] Y. Polyanskiy and Y. Wu. Peak-to-average power ratio of good codes for Gaussian channel. *IEEE Trans. Inf. Theory*, 60(12):7655–7660, December 2014.

[PW15] Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and Bayesian networks. *arXiv preprint arXiv:1508.06025*, 2015.

[Ree65] Alec H Reeves. The past present and future of PCM. *IEEE Spectrum*, 2(5):58–62, 1965.

[RSU01] Thomas J. Richardson, Mohammad Amin Shokrollahi, and Rüdiger L. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *IEEE Transactions on Information Theory*, 47(2):619–637, 2001.

[RU96] Bixio Rimoldi and Rüdiger Urbanke. A rate-splitting approach to the gaussian multiple-access channel. *Information Theory, IEEE Transactions on*, 42(2):364–375, 1996.

[RZ86] Ryabko B. Reznikova Zh. Analysis of the language of ants by information-theoretical methods. *Problemi Peredachi Informatsii*, 22(3):103–108, 1986. English translation: http://reznikova.net/R-R-entropy-09.pdf.

[SF11] Ofer Shayevitz and Meir Feder. Optimal feedback communication via posterior matching. *IEEE Trans. Inf. Theory*, 57(3):1186–1222, 2011.

[Sio58] Maurice Sion. On general minimax theorems. *Pacific J. Math*, 8(1):171–176, 1958.

[Smi71] J. G. Smith. The information capacity of amplitude and variance-constrained scalar Gaussian channels. *Information and Control*, 18:203 – 219, 1971.

[Spe15] Spectre. SPECTRE: Short packet communication toolbox, 2015. GitHub repository.

[Spi96] Daniel A. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Transactions on Information Theory*, 42(6):1723–1731, 1996.

[Spi97] Daniel A. Spielman. The complexity of error-correcting codes. In *Fundamentals of Computation Theory*, pages 67–84. Springer, 1997.

[SV11] Wojciech Szpankowski and Sergio Verdú. Minimum expected length of fixed-to-variable lossless compression without prefix constraints. *IEEE Trans. Inf. Theory*, 57(7):4017–4025, 2011.

[TV05] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[UR98] R. Urbanke and B. Rimoldi. Lattice codes can achieve capacity on the AWGN channel. *IEEE Transactions on Information Theory*, 44(1):273–278, 1998.

[Ver07] S. Verdú. *EE528–Information Theory, Lecture Notes*. Princeton Univ., Princeton, NJ, 2007.

[Yek04] Sergey Yekhanin. Improved upper bound for the redundancy of fix-free codes. *IEEE Trans. Inf. Theory*, 50(11):2815–2818, 2004.

[Yos03] Nobuyuki Yoshigahara. *Puzzles 101: A Puzzlemaster's Challenge*. A K Peters, Natick, MA, USA, 2003.

[Yu97] Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435, 1997.

[Zam14] Ram Zamir. *Lattice Coding for Signals and Networks*. Cambridge University Press, Cambridge, 2014.

[ZY97] Zhen Zhang and Raymond W Yeung. A non-Shannon-type conditional inequality of information quantities. *IEEE Trans. Inf. Theory*, 43(6):1982–1986, 1997.

[ZY98] Zhen Zhang and Raymond W Yeung. On characterization of entropy function via information inequalities. *IEEE Trans. Inf. Theory*, 44(4):1440–1452, 1998.