§ 16. Linear codes. Channel capacity

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

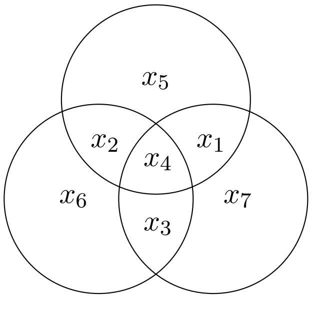
Definition 16.1 (Linear code)**.** Let $\mathcal{X} = \mathcal{Y} = \mathbb{F}_q^n$, $M = q^k$. Denote the codebook by $\mathcal{C} \triangleq \{c_u : u \in \mathbb{F}_q^k\}$. A code $f : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is a **linear code** if $\forall u \in \mathbb{F}_q^k$, $c_u = uG$ (row-vector 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 $g_{\text{synd}}(s) = \operatorname{argmax}_{z:Hx^T=s} P_Z(z)$, is called the "syndrome decoder", which decodes the most likely realization of the noise.

Proof. 1. The maximum likelihood decoder for linear code is

$$g(y) = \operatorname*{argmax}_{c\in\mathcal{C}} P_{Y|X}(y|c) = \operatorname*{argmax}_{c:Hc^T=0} P_Z(y-c) = y - \underbrace{\operatorname*{argmax}_{z:Hz^T=Hy^T} P_Z(z)}_{\triangleq g_{\text{synd}}(Hy^T)},$$

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:

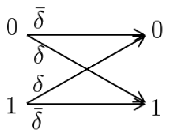
- A channel is *discrete* if $\mathcal{A}$ and $\mathcal{B}$ are finite.
- A channel is *additive-noise* if $\mathcal{A} = \mathcal{B}$ are abelian group, and

$$P_{y^n|x^n} = P_{Z^n}(y^n - x^n) \Leftrightarrow Y^n = X^n + Z^n.$$

- 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.

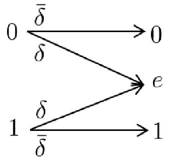
Example:

BSC



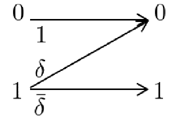
$$\begin{bmatrix} \bar{\delta} & \delta \\ \delta & \bar{\delta} \end{bmatrix}$$

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

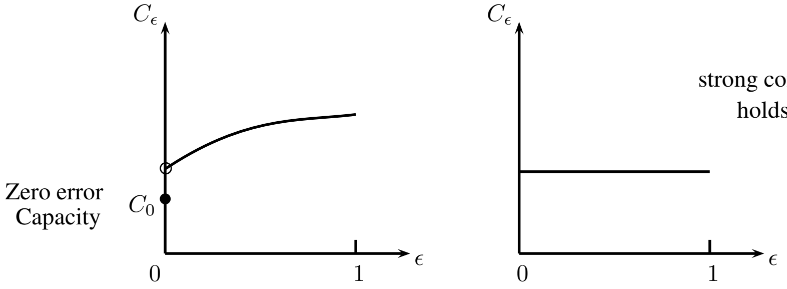
$$M^*(n,\epsilon) = \max\{M : \exists\ (n, M, \epsilon)\text{-code}\} \tag{16.3}$$
$$M^*_{\max}(n,\epsilon) = \max\{M : \exists\ (n, M, \epsilon)_{\max}\text{-code}\} \tag{16.4}$$

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)**.**

$$C_\epsilon = \sup\{R : \forall\delta > 0, \exists n_0(\delta), \forall n \ge n_0(\delta), \exists (n, 2^{n(R-\delta)}, \epsilon) \text{ code}\}$$
$$C = \sup\{R : \forall\epsilon > 0, \forall\delta > 0, \exists n_0(\delta,\epsilon), \forall n \ge n_0(\delta,\epsilon), \exists (n, 2^{n(R-\delta)}, \epsilon) \text{ code}\}$$

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. $\forall\tau \in (0,1)$,

$$\tau M^*(n, \epsilon(1-\tau)) \le M^*_{\max}(n,\epsilon) \le M^*(n,\epsilon)$$

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 $V = \operatorname{Var}[i(X^*;Y^*)] < \infty$, 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

$$\begin{aligned}
&\log M \geq nC + \sqrt{nV}t + \log\left(\epsilon - \mathbb{P}\left[\sum i(X_k;Y_k) \leq nC + \sqrt{nV}t\right]\right)\\
\Longrightarrow\ &\log M \geq nC + \sqrt{nV}t + \log\left(\epsilon - \Phi(t)\right) \quad \forall\Phi(t) < \epsilon\\
\Longrightarrow\ &\frac{\log M - nC}{\sqrt{nV}} \geq t + \frac{\log(\epsilon - \Phi(t))}{\sqrt{nV}}
\end{aligned}$$

Where $\Phi(t)$ is the standard normal CDF. Taking the liminf of both sides

$$\liminf_{n\to\infty} \frac{\log M^*(n,\epsilon) - nC}{\sqrt{nV}} \geq t \quad \forall t\ s.t.\ \Phi(t) < \epsilon$$

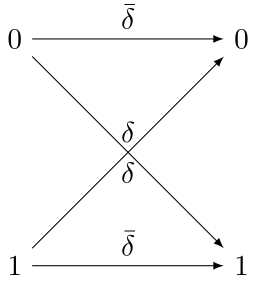
Taking $t \nearrow \Phi^{-1}(\epsilon)$, and writing the liminf in little o form completes the proof

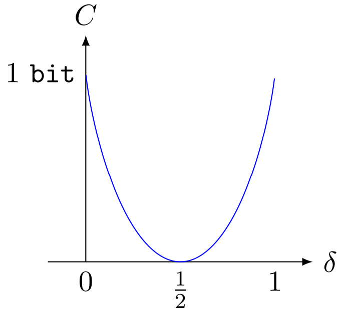
$$\log M^*(n,\epsilon) \geq nC - \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n})$$

□

16.4 Examples of DMC

Binary symmetric channels





$$Y = X + Z, \quad Z \sim \text{Bern}(\delta) \perp\!\!\!\perp X$$

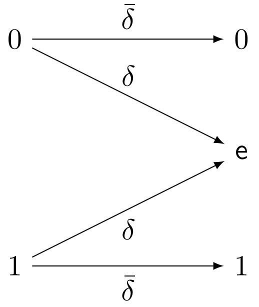
<u>Capacity of BSC:</u>

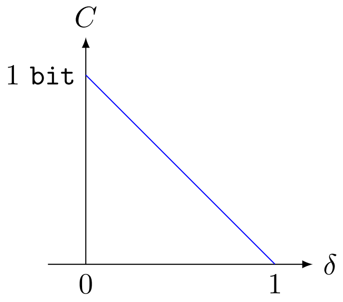
$$C = \sup_{P_X} I(X;Y) = 1 - h(\delta)$$

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 $\{P_{X^n}, n = 1, 2, \ldots\}$ 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:

$$g(n) \triangleq \sup_{a_j \geq 0: \sum_{j=1}^n a_j = 1} \sum_{j=1}^n a_j \log^2 a_j \,. \tag{16.16}$$

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,

$$2g(|\mathcal{B}|) \leq 2\log^2 |\mathcal{B}| \,.$$

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

$$\sum_{k=1}^n b_k x_k \leq b_m \sum_{k=1}^m x_k + \sum_{k=m+1}^n b_k x_k$$

Now throw in the rest of the x_k's in the summation

$$\Longrightarrow \frac{1}{b_n} \sum_{k=1}^n b_k x_k \leq \frac{b_m}{b_n} \sum_{k=1}^\infty x_k + \sum_{k=m+1}^n \frac{b_k}{b_n} x_k \leq \frac{b_m}{b_n} \sum_{k=1}^\infty x_k + \sum_{k=m+1}^\infty x_k$$

$$\Longrightarrow \lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^n b_k x_k \leq \sum_{k=m+1}^\infty x_k \to 0$$

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:

$$\operatorname{Var}[i(X;Y)] = \rho^2(X,Y)\log^2 e \leq \log^2 e \,, \qquad \rho(X,Y) = \frac{\operatorname{cov}[X,Y]}{\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}} \,. \tag{16.17}$$

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 $\operatorname{Var}[\cdot] = \operatorname{Var}[\mathbb{E}[\cdot|\tilde{X}]] + \operatorname{Var}[\cdot|\tilde{X}]$ 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

$$\operatorname{Var}[\mathbb{E}[i(\tilde{X};Y)|\tilde{X}]] = \frac{2\log^2 e}{4\sigma_Y^4}\sigma_{\tilde{X}}^4 \,.$$

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.

6.441 Information Theory
Spring 2016