CS3236 Lecture Notes #4:
Channel Coding

Jonathan Scarlett

March 22, 2023

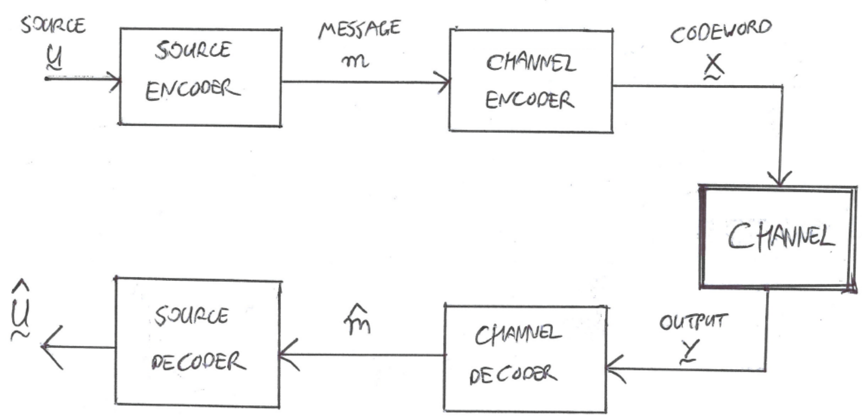
Useful references:

- Cover/Thomas Chapter 7
- MacKay Chapters 8–10
- Shannon's 1948 paper "A Mathematical Introduction to Communication"

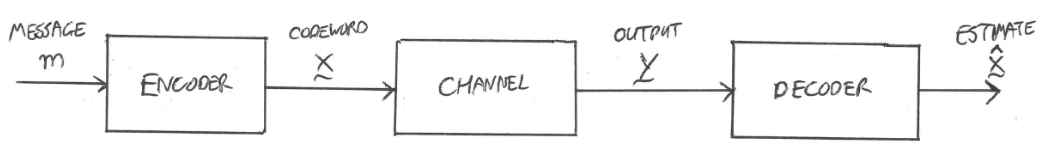
1 Setup

Overview.

- Full communication setup (source and channel coding):

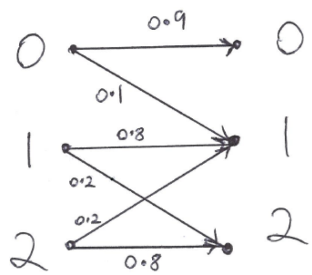


- Channel coding setup:



Channel model.

- The *channel* is the medium over which we transmit information
- We denote the input by x and the output by y (or X and Y when we want to highlight that they are random)
- We assume (for now) that the channel input and output only take finitely many possible values (e.g., binary, $x \in \{0,1\}$ and $y \in \{0,1\}$). These sets of possible inputs/outputs are denoted by $\mathcal{X}$ and $\mathcal{Y}$. We call these the *input alphabet* and *output alphabet*.
- We adopt a *probabilistic modeling* approach: When the input is $x \in \mathcal{X}$, a given output $y \in \mathcal{Y}$ is produced with probability $P_{Y|X}(y|x)$.
- The channel transition probabilities are typically depicted graphically. A simple example:



Problem description.

- We generically view the communication problem as seeking to transmit a message $m \in \{1, \dots, M\}$. In particular, if a fixed-length source code outputs a length-k sequence of bits, then we can set $M = 2^k$ and map each such sequence to a unique index m.
- The *encoder* takes as input the message m, and outputs a sequence of channel inputs $x_1, \dots, x_n$. To make the dependence on the message explicit, we define the *codeword* $\mathbf{x}^{(m)} = (x_1^{(m)}, \dots, x_n^{(m)})$, which is the sequence produced when the message is m.
  - The collection of codewords $\mathcal{C} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$ is referred to as the *codebook*. It is known at both the encoder and decoder, but only the encoder knows m.

  The codeword $\mathbf{x}^{(m)}$ is transmitted over the channel in n uses, and the resulting output sequence is denoted by $\mathbf{y} = (y_1, \dots, y_n)$.
- We focus (for now) on *discrete memoryless channels*:
  - Discrete: The input/output alphabets $\mathcal{X}$ and $\mathcal{Y}$ are finite, as stated above;
  - Memoryless: When we transmit several symbols (say, n of them) over the channel in successive uses, the outputs are (conditionally) independent:

$$P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i).$$

- Given the output sequence $\mathbf{y}$ (and knowledge of the codebook $\mathcal{C}$), the *decoder* forms an estimate $\hat{m}$ of the message m.

A fundamental trade-off.

- Clearly we would like $\hat{m} = m$; if not then an *error* has occurred. Accordingly, we define the *error probability*

$$P_{\mathrm{e}} = \mathbb{P}[\hat{m} \neq m]. \tag{1}$$

  We will henceforth consider this probability as being averaged over m uniform on $\{1, \dots, M\}$ (along with the randomness in the channel), though without much extra effort we can actually get similar results for the *maximal* error probability $\max_{m=1,\dots,M} \mathbb{P}[\hat{m} \neq m \,|\, m \text{ chosen}]$.

- We would like to transmit as much data as possible (i.e., high M); instead of considering M directly, we usually measure this via the *rate* (measured in bits per channel use):

$$R = \frac{1}{n} \log_2 M.$$

  That is, the number of messages is $M = 2^{nR}$.

  - For instance, if $M = 2^n$ then $R = 1$, which makes sense because n bits (each a 0 or 1) corresponds to 2^n possible combinations (of 0s and 1s).

- The quantity n also plays a fundamental role; it is referred to as the *block length*.

- Key question: What is the fundamental trade-off between error probability P_{e}, rate R, and block length n? In particular, how high can the rate be while keeping the error probability small?

2 Channel Capacity

Definition.

- **Definition.** The channel capacity C is defined to be the maximum[1] of all rates R such that, for any target error probability $\epsilon > 0$, there exists a block length n and codebook $\mathcal{C} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$ with $M = 2^{nR}$ codewords such that $P_{\mathrm{e}} \leq \epsilon$.

  - In simpler terms: This is the highest rate such that the error probability can be made arbitrarily small at *some* (possibly large) block length.

- **Channel Coding Theorem.** The capacity of a discrete memoryless channel $P_{Y|X}$ is

$$C = \max_{P_X} I(X;Y).$$

  The proof is split into two parts (given in later sections):

  - Achievability part: For any $R < C$, there exists a code of rate at least R with arbitrarily small error probability.

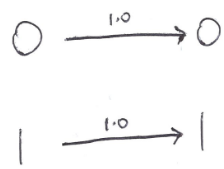
[1] More mathematically precisely, the supremum.

- Converse part: For any $R > C$, any code of rate at least R cannot have arbitrarily small error probability.

- **Definition.** For a given channel $P_{Y|X}$, any input distribution P_X maximizing the mutual information above is called a *capacity-achieving input distribution*.

Examples.

- Noiseless channel:
  - Consider a noiseless channel with $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ in which the output deterministically equals the input (i.e., $Y = X$):
  - An illustration:



  - Since $Y = X$, we have $H(X|Y) = 0$ (there is no uncertainty in X once we know Y), and hence

$$I(X;Y) = H(X) - H(X|Y) = H(X).$$

Therefore, the capacity is

$$C = \max_{P_X} I(X;Y) = \max_{P_X} H(X) = 1$$

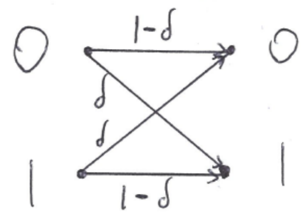
since the entropy of a binary random variable is at most one (achieved when $P_X(0) = P_X(1) = \frac{1}{2}$).

  - This result should not be surprising – if there is no noise, we can reliably transmit one bit per channel use without even doing any coding!

- Binary symmetric channel:
  - Again consider $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, but now each input is flipped with some probability $\delta \in (0, 1)$:

$$P_{Y|X}(y|x) = \begin{cases} 1 - \delta & y = x \\ \delta & y = 1 - x. \end{cases}$$

  - An illustration:



  - In this case, it is more convenient to use the expansion $I(X;Y) = H(Y) - H(Y|X)$.

– In general we have $H(Y|X) = \sum_x P_X(x) H(Y|X=x)$, but due to the symmetry things simplify. Specifically, regardless of whether we condition on $X=0$ or $X=1$, the conditional probabilities of Y are still δ and $1-\delta$, and so $H(Y|X=x) = H_2(\delta)$, where $H_2(\delta) = \delta \log_2 \frac{1}{\delta} + (1-\delta)\log_2 \frac{1}{1-\delta}$ is the binary entropy function.

– This gives $H(Y|X) = H_2(\delta)$ and hence

$$C = \max_{P_X} I(X;Y) = \max_{P_X} \big(H(Y) - H_2(\delta)\big).$$

If we were maximizing over P_Y directly, we could get $\max H(Y) = 1$ by the same argument as the noiseless case by letting P_Y be uniform. But in this case, even though we can only control P_X, we can still produce uniform P_Y – just let P_X be uniform!

  * Indeed, if $P_X(0) = P_X(1) = \frac{1}{2}$, then we have

$$P_Y(0) = \frac{1}{2}(1-\delta) + \frac{1}{2}\delta = \frac{1}{2},$$

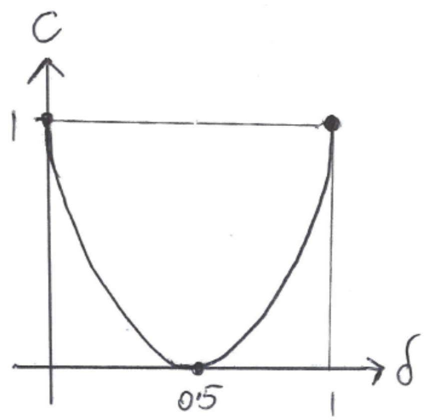
  and similarly $P_Y(1) = \frac{1}{2}$.

– Therefore, the capacity is

$$C = 1 - H_2(\delta)$$

and the capacity-achieving input distribution is $P_X(0) = P_X(1) = \frac{1}{2}$.

– An illustration:



– As expected, setting $\delta = 0$ recovers the noiseless capacity $C = 1$. Notice also that $\delta = \frac{1}{2}$ gives capacity zero, because in this case we have $P_{Y|X}(y|x) = \frac{1}{2}$ regardless of the input x, so the output carries no information about the input.

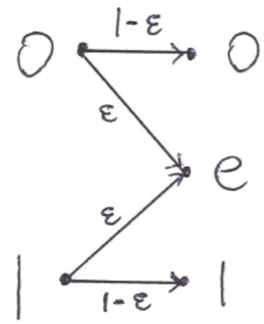
- Binary erasure channel:

  – Consider $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{0,1,e\}$, and transition probabilities

$$P_{Y|X}(y|x) = \begin{cases} 1-\epsilon & y = x \\ \epsilon & y = e \\ 0 & y = 1-x \end{cases}$$

for some *erasure probability* ϵ. In words, the output equals the input with probability $1-\epsilon$, but is "erased" (corresponding to output e) with probability ϵ.

– An illustration:



– This time it turns out easier to use the expansion $I(X;Y) = H(X) - H(X|Y)$, though the $I(X;Y) = H(Y) - H(Y|X)$ approach is also possible (see the tutorial).

– $H(X|Y)$ is fairly easy to characterize, because $H(X|Y=0) = H(X|Y=1) = 0$ (there is no uncertainty in X when $Y \neq e$). Hence,

$$H(X|Y) = \sum_y P_Y(y)H(X|Y=y) = P_Y(e)H(X|Y=e).$$

Then, given $Y = e$, we have

$$P_{X|Y}(0|e) = \frac{P_{XY}(0,e)}{P_Y(e)} = \frac{P_X(0)\epsilon}{\epsilon} = P_X(0),$$

and similarly $P_{X|Y}(1|e) = P_X(1)$. Hence, $H(X|Y=e) = H(X)$.

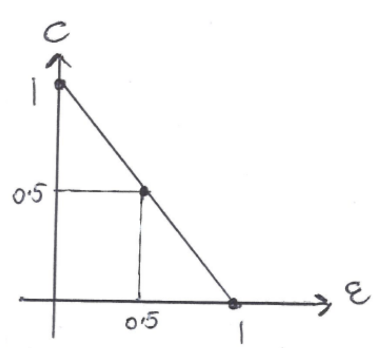
– Combining the above findings gives

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= (1-\epsilon)H(X). \end{aligned}$$

– Upon maximizing over P_X, we can get the maximal value $H(X) = 1$ with $P_X(0) = P_X(1) = \frac{1}{2}$. Therefore, the capacity is

$$C = 1 - \epsilon.$$

An illustration:



- In all of these examples, the capacity-achieving input distribution is uniform.
  - In fact, much more general classes of symmetric channels (not necessarily binary) have a uniform capacity-achieving input distribution. See Cover/Thomas Section 7.2 for details.
  - For non-symmetric channels, the capacity-achieving P_X may be non-uniform. Moreover, we often can't find the optimal choice analytically, so instead need to do so numerically (efficient algorithms for doing this are known; see Cover/Thomas Section 10.8).

3 Jointly Typical Sequences

The following definition and properties will be crucial in proving the achievability part mentioned above.

- **Definition**: A pair $(\mathbf{x}, \mathbf{y})$ of length-n input and output sequences is said to be *jointly typical* with respect to a joint distribution P_{XY} if the following conditions hold:

$$2^{-n(H(X)+\epsilon)} \le P_{\mathbf{X}}(\mathbf{x}) \le 2^{-n(H(X)-\epsilon)}$$
$$2^{-n(H(Y)+\epsilon)} \le P_{\mathbf{Y}}(\mathbf{y}) \le 2^{-n(H(Y)-\epsilon)}$$
$$2^{-n(H(X,Y)+\epsilon)} \le P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \le 2^{-n(H(X,Y)-\epsilon)}.$$

  The set of all such sequences is denoted by $\mathcal{T}_n(\epsilon)$, and is called the *jointly typical set*.
  - In simpler terms: The X sequence, Y sequence, and joint (X, Y) sequence are all typical according to the previous lecture's definition.

- **Key properties**:[2]
  1. (Equivalent definition) We have $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}_n(\epsilon)$ if and only if the following conditions hold:

$$H(X) - \epsilon \le \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{P_X(x_i)} \le H(X) + \epsilon$$
$$H(Y) - \epsilon \le \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{P_Y(y_i)} \le H(Y) + \epsilon$$
$$H(X,Y) - \epsilon \le \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{P_{XY}(x_i, y_i)} \le H(X,Y) + \epsilon.$$

  2. (High probability) $\mathbb{P}[(\mathbf{X}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$.
  3. (Cardinality upper bound) $|\mathcal{T}_n(\epsilon)| \le 2^{n(H(X,Y)+\epsilon)}$.
  4. (Probability for independent sequences) If $(\mathbf{X}', \mathbf{Y}') \sim P_{\mathbf{X}}(\mathbf{x}')P_{\mathbf{Y}}(\mathbf{y}')$ are independent copies of $(\mathbf{X}, \mathbf{Y})$, then the probability of joint typicality is

$$\mathbb{P}[(\mathbf{X}', \mathbf{Y}') \in \mathcal{T}_n(\epsilon)] \le 2^{-n(I(X;Y)-3\epsilon)}.$$

  - The first three properties have similar intuition to the "X-only" setting of the previous lecture.

[2] Near-matching lower bounds can also be shown for the final two properties, but these are omitted here.

- The final one is distinct from that setting. Intuitively, if $\mathbf{X}'$ and $\mathbf{Y}'$ are generated independently, then the "further" P_{XY} is from being independent, the less likely it is for those independent sequences to be jointly typical with respect to P_{XY}. Mutual information naturally arises because it measures "how far" (X, Y) are from being independent: $I(X;Y) = D(P_{XY} \| P_X \times P_Y)$.
- In fact, the fourth property is a special case of a more general result: If a sequence $\mathbf{Z} = (Z_1, \dots, Z_n)$ is drawn i.i.d. from some distribution Q_Z, then the probability that it is typical with respect to some other distribution P_Z is roughly $2^{-nD(P_Z \| Q_Z)}$.

- **Proofs**:

1. Simple re-arranging like in the previous lecture.
2. Law of large numbers applied (separately) to the 3 conditions in the first property.
3. Same as the previous lecture via $P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \ge 2^{-n(H(X,Y)+\epsilon)}$ and $\sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{T}_n(\epsilon)} P_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) \le 1$.
4. We have

$$\begin{aligned}\mathbb{P}[(\mathbf{X}', \mathbf{Y}') \in \mathcal{T}_n(\epsilon)] &= \sum_{(\mathbf{x}', \mathbf{y}') \in \mathcal{T}_n(\epsilon)} P_{\mathbf{X}}(\mathbf{x}') P_{\mathbf{Y}}(\mathbf{y}') \\ &\overset{(a)}{\le} \sum_{(\mathbf{x}', \mathbf{y}') \in \mathcal{T}_n(\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\ &\overset{(b)}{\le} 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\ &\overset{(c)}{=} 2^{-n(I(X;Y)-3\epsilon)},\end{aligned}$$

where (a) uses the fact that $P_{\mathbf{X}}(\mathbf{x}') \le 2^{-n(H(X)-\epsilon)}$ and $P_{\mathbf{Y}}(\mathbf{y}') \le 2^{-n(H(Y)-\epsilon)}$ within $\mathcal{T}_n(\epsilon)$, (b) uses the upper bound in property 3, and (c) uses $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

4 Achievability via Random Coding

Overview.

- Challenge: Devising explicit/specific codes and studying their performance is very difficult.
- Key idea (the probabilistic method): Show that **randomly chosen** codes perform well on average. Obviously, the best possible code must perform at least as well as the average.
- Note: The good code whose existence we prove may have very high computation/storage requirements. This approach merely shows that reliable communication is *mathematically* possible for rates below capacity, but not how to get there with a *practical* design.

Codebook generation.

- Recall that the encoding is done via a codebook $\mathcal{C} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$, where message m is encoded into the length-n sequence $\mathbf{x}^{(m)} = (x_1^{(m)}, \dots, x_n^{(m)})$.
- We consider the following *random coding* approach:

Generate each symbol $X_i^{(m)}$ of each codeword randomly and

independently according to some distribution P_X (to be specified)

Note that we use capital letters for $\mathbf{X}$, $X_i^{(m)}$, etc. when we want to highlight that they are random.

- For example, if $\mathcal{X} = \{0, 1\}$ and $P_X(1) = P_X(0) = \frac{1}{2}$, then we are just setting every bit of every codeword according to a fair "coin flip".

Encoding and decoding.

- As mentioned above, the encoder simply maps m to $\mathbf{X}^{(m)} = (X_1^{(m)}, \dots, X_n^{(m)}) \in \mathcal{X}^n$, which is transmitted via n uses of the channel.
- The decoder receives the output sequence $\mathbf{Y} = (Y_1, \dots, Y_n)$, and also knows the codebook. For each $\tilde{m} = 1, \dots, M$, it checks whether the pair $(\mathbf{X}^{(\tilde{m})}, \mathbf{Y})$ is jointly typical, and does the following:
  - If there exists a unique $\tilde{m}$ that joint typicality holds, then the decoder estimates $\hat{m} = \tilde{m}$.
  - If there exists no such $\tilde{m}$, or multiple such $\tilde{m}$, an error is declared (or alternatively, $\hat{m}$ is simply chosen at random).

  Note that "joint typicality" is defined with respect to $P_{XY} = P_X \times P_{Y|X}$. The channel $P_{Y|X}$ was fixed as part of the problem, whereas P_X is something we chose ourselves (during the codebook generation).
- Note that the joint distributions between the codewords and the output are exactly those we need to apply properties 2 and 4 of joint typicality:
  - For the correct m (i.e., $\mathbf{X}^{(m)}$ is transmitted), $P_{\mathbf{Y}|\mathbf{X}}$ is i.i.d. according to $P_{Y|X}$, and $\mathbf{X}^{(m)}$ itself is i.i.d. according to P_X by construction, so overall $(\mathbf{X}^{(m)}, \mathbf{Y})$ is i.i.d. on $P_{XY} = P_X \times P_{Y|X}$.
  - For any incorrect $\tilde{m}$ (i.e., $\mathbf{X}^{(\tilde{m})}$ is a non-transmitted codeword), we have that $\mathbf{X}^{(\tilde{m})}$ and $\mathbf{Y}$ are independent, since $\mathbf{Y}$ only depends on the transmitted codeword, not the other ones. Therefore, the joint distribution of $(\mathbf{X}^{(\tilde{m})}, \mathbf{Y})$ takes the form $P_{\mathbf{X}}(\mathbf{x})P_{\mathbf{Y}}(\mathbf{y})$.

Analysis of the error probability.

- In order to have $\hat{m} = m$, it is clearly sufficient that the following two events occur:
  1. $(\mathbf{X}^{(m)}, \mathbf{Y})$ is jointly typical;
  2. None of the other $(\mathbf{X}^{(\tilde{m})}, \mathbf{Y})$ are jointly typical (with $\tilde{m} \neq m$).
- Let $\overline{P}_{\mathrm{e}}^{(m)}$ denote the error probability given that the message is m, averaged over both the randomness in the channel *and* the random codebook (previously we only averaged over the former). This is called the *random-coding error probability.*
- We have just argued that the success probability $1 - \overline{P}_{\mathrm{e}}^{(m)}$ satisfies

$$1 - \overline{P}_{\mathrm{e}}^{(m)} \geq \mathbb{P}\bigg[(\mathbf{X}^{(m)}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon) \cap \bigcap_{\tilde{m} \neq m} \Big\{(\mathbf{X}^{(\tilde{m})}, \mathbf{Y}) \notin \mathcal{T}_n(\epsilon)\Big\}\bigg],$$

  which, by de Morgan's laws, is equivalent to

$$\overline{P}_{\mathrm{e}}^{(m)} \leq \mathbb{P}\bigg[(\mathbf{X}^{(m)}, \mathbf{Y}) \notin \mathcal{T}_n(\epsilon) \cup \bigcup_{\tilde{m} \neq m} \Big\{(\mathbf{X}^{(\tilde{m})}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)\Big\}\bigg].$$

- Using the union bound $\mathbb{P}[A_i \cup \ldots \cup A_N] \leq \sum_{i=1}^{N} \mathbb{P}[A_i]$, we obtain

$$\overline{P}_{\mathrm{e}}^{(m)} \leq \mathbb{P}\big[(\mathbf{X}^{(m)}, \mathbf{Y}) \notin \mathcal{T}_n(\epsilon)\big] + \sum_{\tilde{m} \neq m} \mathbb{P}\big[(\mathbf{X}^{(\tilde{m})}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)\big].$$

- By the i.i.d. random coding method and the memoryless property of the channel, $(\mathbf{X}^{(m)}, \mathbf{Y})$ is i.i.d. on P_{XY}. Moreover, since $\mathbf{X}^{(m)}$ is the only codeword that $\mathbf{Y}$ depends on, we also have that $(\mathbf{X}^{(\tilde{m})}, \mathbf{Y})$ is an independent pair with the same $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$ marginals as $(\mathbf{X}^{(m)}, \mathbf{Y})$.

- Therefore, the joint typicality properties in the previous section give $(\mathbf{X}^{(m)}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)$ with probability approaching one (as n increases), and that the probability of $(\mathbf{X}^{(\tilde{m})}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)$ is at most $2^{-n(I(X;Y)-3\epsilon)}$, which gives

$$\begin{aligned}\overline{P}_{\mathrm{e}}^{(m)} &\leq \mathbb{P}\big[(\mathbf{X}^{(m)}, \mathbf{Y}) \notin \mathcal{T}_n(\epsilon)\big] + \sum_{\tilde{m} \neq m} \mathbb{P}\big[(\mathbf{X}^{(\tilde{m})}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)\big] \\ &\overset{(a)}{\leq} \delta_n + \sum_{\tilde{m} \neq m} 2^{-n(I(X;Y)-3\epsilon)} \\ &\overset{(b)}{\leq} \delta_n + M \times 2^{-n(I(X;Y)-3\epsilon)},\end{aligned}$$

  where in (a) δ_n denotes a sequences that tends to 0 as $n \to \infty$, and in (b) we used the fact that the number of terms in the summation is $M - 1 \leq M$.

- Since $M = 2^{nR}$, we find that for $R < I(X;Y) - 3\epsilon$ the overall upper bound on $\overline{P}_{\mathrm{e}}^{(m)}$ tends to zero as $n \to \infty$. Since ϵ may be arbitrarily small, this means $\overline{P}_{\mathrm{e}}^{(m)}$ can be made arbitrarily small for any rate R arbitrarily close to $I(X;Y)$.

- Since this holds for any m, it also holds for the random-coding error probability $\frac{1}{M}\sum_{m=1}^{M} \overline{P}_{\mathrm{e}}^{(m)}$ averaged over the message m. (In fact, due to the symmetry of random coding, $\overline{P}_{\mathrm{e}}^{(m)}$ is the same for all m.)

- Finally, by choosing P_X to achieve the maximum in the definition $C = \max_{P_X} I(X;Y)$, we deduce that we can get vanishing error probability for rates arbitrarily close to the capacity C.

(Optional) Alternative proof.

- In an interesting alternative proof, instead of the notion of joint typicality we considered in the discrete setting, the decoder looks for a codeword $\mathbf{x}$ such that

$$\sum_{i=1}^{n} \log_2 \frac{P_{Y|X}(y_i|x_i)}{P_Y(y_i)} \geq \gamma$$

  for some threshold γ. This can be viewed as a form of *one-sided* typicality.

- Using a simple change of measure argument, one can show that a given *incorrect* codeword passes this threshold test with probability at most $2^{-\gamma}$. By the union bound, the probability of this occurring for *any* incorrect codeword is at most $M2^{-\gamma}$, which tends to zero if we set γ to be slightly above $\log_2 M$.

- By the law of large numbers, for the *correct* codeword, $\sum_{i=1}^{n} \log_2 \frac{P_{Y|X}(y_i|x_i)}{P_Y(y_i)}$ is close to $nI(X;Y)$ with high probability. Therefore, to exceed the threshold $\gamma \approx \log_2 M = nR$, we just need $R < I(X;Y)$.

- This proof is rooted in two early works: "Certain results in coding theory for noisy channels" (Shannon, 1957) and "A new basic theorem of information theory" (Feinstein, 1954).

5 Converse via Fano's Inequality

- Let m denote a transmitted message uniform on $\{1, \dots, M\}$, and let $\hat{m}$ be its estimate (in a slight shift from our usual convention, these are random variables even though they are written in lower-case).
- The error probability is $P_{\mathrm{e}} = \mathbb{P}[\hat{m} \neq m]$. Fano's inequality from the previous lecture[3] states that

$$\begin{aligned} H(m|\hat{m}) &\leq H_2(P_{\mathrm{e}}) + P_{\mathrm{e}} \log_2(M-1) \\ &\leq 1 + P_{\mathrm{e}} \log_2 M. \end{aligned}$$

- Since m is uniform on $\{1, \dots, M\}$, we have $H(m) = \log_2 M$, which gives

$$\begin{aligned} I(m;\hat{m}) &= H(m) - H(m|\hat{m}) \\ &\geq \log_2 M - P_{\mathrm{e}} \log_2 M - 1 \\ &= (1 - P_{\mathrm{e}}) \log_2 M - 1, \end{aligned}$$

  where the inequality uses the previous display equation. Simple re-arranging gives

$$P_{\mathrm{e}} \geq 1 - \frac{I(m;\hat{m}) + 1}{\log_2 M}.$$

  Intuitively, this says that to achieve a small error probability, we need the amount of information that $\hat{m}$ reveals about m to be close to the prior uncertainty in m (which is $\log_2 M$).

- The key step is to bound the mutual information. We have:

$$\begin{aligned} I(m;\hat{m}) &\overset{(a)}{\leq} I(\mathbf{X};\mathbf{Y}) \\ &\overset{(b)}{=} H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) \\ &\overset{(c)}{\leq} \sum_{i=1}^{n} H(Y_i) - H(\mathbf{Y}|\mathbf{X}) \\ &\overset{(d)}{=} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|\mathbf{X}) \\ &\overset{(e)}{=} \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\overset{(f)}{=} \sum_{i=1}^{n} I(X_i;Y_i) \\ &\overset{(g)}{\leq} nC, \end{aligned}$$

  where:

[3] Now with $(m, \hat{m})$ in place of the generic symbols $(X, \hat{X})$ used in that lecture.

- (a) uses the data processing inequality (note that $m \to \mathbf{X} \to \mathbf{Y} \to \hat{m}$ forms a Markov chain);
- (b) and (f) use the definition of mutual information;
- (c) uses the sub-additivity of entropy;
- (d) uses the fact that the Y_i are conditionally independent given $\mathbf{X}$ (and entropy is additive for independent random variables), i.e., the "memoryless" assumption;
- (e) uses the fact that Y_i depends on $\mathbf{X}$ only through X_i;
- (g) uses the definition of capacity (C is the maximum mutual information between X and Y).

- Combining the previous two dot points with $\log_2 M = \log_2 2^{nR} = nR$ gives

$$P_\mathrm{e} \geq 1 - \frac{C + 1/n}{R},$$

which means that P_e is bounded away from 0 as $n \to \infty$ whenever $R > C$.

  - **A minor technical detail:** We originally stated the channel coding theorem for arbitrary n, not only $n \to \infty$. However, the result for $n \to \infty$ implies the result for arbitrary n. Indeed, the only way to get arbitrarily small error probability at finite n is to have $P_\mathrm{e} = 0$. But if we can achieve $P_\mathrm{e} = 0$ at some rate with finite block length, we can also achieve it as $n \to \infty$ by simply using that codebook many times in succession.

6 (Optional) Joint Source-Channel Coding

- If we can successfully perform both source coding and channel coding, then we can form the overall communication system as shown in the first figure of this document (Page 1).

- Denoting the source block length by k and the channel block length by n, and taking both to be sufficiently large, we obtain the following condition for overall reliable communication:

$$\underbrace{n \times C}_{\text{Total Capacity}} > \underbrace{k \times H}_{\text{Total Entropy}}$$

or equivalently

$$\frac{k}{n} < \frac{C}{H}.$$

Indeed, this result follows from a simple combination of the source coding and channel coding theorems. We first compress the source and represent it using roughly $M \approx 2^{kH}$ bits, and then we send the corresponding index $m \in \{1, \dots, M\}$ across the channel in n uses.

- It may seem strange that we are removing first redundancy (source coding) only to then add redundancy (channel coding) – could a joint approach be better? This is known as *joint source-channel coding.*

- **Separation theorem.** Even with joint source-channel coding, reliable communication is impossible if $\frac{k}{n} > \frac{C}{H}$. Therefore, separate source-channel coding is asymptotically optimal at large block lengths.
  - Proof: Mostly similar to that above based on Fano's inequality. See Section 7.13 of Cover/Thomas.
  - Note: The gains can be significant at *finite* block lengths (beyond the scope of this course).