Outline

This lecture will cover

- Fano’s inequality.
- channel capacity and some channel models.
- a preview of the channel coding theorem.
- the tools that are needed to establish the channel coding theorem.

All illustrations are borrowed from the book.
Fano’s inequality

Estimate $X$ from $Y$. Relate error in guessing $X$ to $H(X|Y)$.

We know that $H(X|Y) = 0$ if $X = g(Y)$ (Problem 2.5) $\rightarrow$ can estimate $X$ with zero error probability. Extension: $H(X|Y)$ “small” $\rightarrow$ can estimate $X$ with low error probability.

Formally: $X$ has $p(x)$, $Y$ related via $p(y|x)$, estimate $\hat{X} = g(Y)$ with alphabet $\hat{X}$, error probability $P_e = \Pr \{ \hat{X} \neq X \}$.

Fano’s inequality: For $X \rightarrow Y \rightarrow \hat{X}$

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y).$$

Weaker: $1 + P_e \log |\mathcal{X}| \geq H(X|Y)$ or

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$
Motivation and preview

A communicates with B: A induces a state in B. Physical process gives rise to noise.

Mathematical analog: source $W$, transmitted sequence $X^n$, etc.

Two $X^n$ may give the same $Y^n$ — inputs confusable.

Idea: use only a subset of all possible $X^n$ such that there is, with high probability, only one likely $X^n$ to result in each $Y^n$.

Map $W$ into “widely spaced” $X^n$. Then $\hat{W} = W$ with high probability.

Channel capacity: maximum rate (source bits/channel use) at which we can carry out the above steps.
Channel capacity

Discrete channel: input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, probability transition matrix $p(y|x)$.

Memoryless channel: current output depends only on the current input, conditionally independent of previous inputs or outputs.

“Information” channel capacity of a discrete memoryless channel is

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

Shannon’s channel coding theorem: $C$ highest rate (bits per channel use) at which information can be sent with arbitrary low probability of error.
Some channels I

Noiseless binary channel
- \( I(X; Y) = H(X) - H(X|Y) = H(X) \).
- \( C = 1 \), achieved for uniform \( X \).

Noisy channel with nonoverlapping outputs
- output random, but input uniquely determined.
- \( C = 1 \), achieved for uniform \( X \).
Some channels II

Noisy typewriter

- input either unchanged or shifted (both w.p. $\frac{1}{2}$).
- use of every second input: log 13 bits per transmission without error.
- $I(X; Y) = H(Y) - H(Y|X) = H(Y) - H(\frac{1}{2}, \frac{1}{2}) = H(Y) - 1$.
- $C = \max I(X; Y) = \log 26 - 1 = \log 13$. 

Noiseless subset of inputs
Some channels III

Binary symmetric channel
- simplest channel with errors.
- probability of switched input is $p$.
- “all received bits unreliable”.
- $C = 1 - H(p)$ achieved for uniform $X$.

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

$$= H(Y) - \sum p(x)H(Y|X = x)$$

$$= H(Y) - \sum p(x)H(p)$$

$$= H(Y) - H(p)$$

$$\leq 1 - H(p).$$

Reminder: $H(p) = -p \log p - (1 - p) \log(1 - p)$. 
Some channels IV

Binary erasure channel
- bits are lost rather than corrupted.
- fraction $\alpha$ are erased.
- $e$: receiver knows that it does not know.
- $I(X; Y) = H(Y) - H(Y|X) = H(Y) - H(\alpha)$.
- $C = 1 - \alpha$.
- feedback discussion and surprising fact.

Introduce $E$ with $E = 1$ if $Y = e$. Let $\pi = Pr \{X = 1\}$. Then

\[
H(Y) = H(Y, E) = H(E) + H(Y|E) = H(\alpha) + (1 - \alpha)H(\pi)
\]

and $I(X; Y) = (1 - \alpha)H(\pi)$ yields $C = (1 - \alpha)$ for $\pi = \frac{1}{2}$. 
Symmetric channels I

Transmission matrix. Example for $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$:

$$
p(y|x) = \begin{bmatrix}
0.3 & 0.2 & 0.5 \\
0.5 & 0.3 & 0.2 \\
0.2 & 0.5 & 0.3
\end{bmatrix}
$$

$\Pr\{Y = 1|X = 0\} = 0.2$. Rows must add up to 1.

This is a symmetric channel: row 1 is a permutation of row 2. Other rows and columns are permutations too.

Let $r$ be one row in $p(y|x)$. Then

$$
I(X; Y) = H(Y) - H(Y|X) = H(Y) - H(r) \leq \log |\mathcal{Y}| - H(r).
$$
Symmetric channels II

$I(X; Y)$ maximized for uniform $Y$. Achieved by uniform $X$:

$$p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum p(y|x) = c \frac{1}{|\mathcal{X}|}$$

with $c$ sum over one column.

Generalization: each row is a permutation of every other row, and all column sums are equal. Example:

$$p(y|x) = \begin{bmatrix} 1/3 & 1/6 & 1/2 \\ 1/3 & 1/2 & 1/6 \end{bmatrix}.$$

Channel capacity for weakly symmetric channels is

$$C = \log |\mathcal{Y}| - H(r).$$
Properties of channel capacity

Properties:

- $C \geq 0$, since $I(X; Y) \geq 0$.
- $C \leq \log |\mathcal{X}|$ and $C \leq \log |\mathcal{Y}|$.
- $I(X; Y)$ continuous function of $p(x)$.
- $I(X; Y)$ concave in $p(x)$.

Consequences:

- maximum exists and is finite.
- convex optimization tools can be employed.
Intuitive idea:
- for large block lengths every channel looks like the noisy typewriter.
- one (typical) input sequence gives \( \approx 2^{nH(Y|X)} \) output sequences.
- total number of (typical) output sequences \( \approx 2^{nH(Y)} \) must be divided into sets of size \( 2^{nH(Y|X)} \).
- total number of disjoint sets \( \leq 2^{n(H(Y) - H(Y|X))} = 2^{nI(X;Y)} \).
- can send at most \( 2^{nI(X;Y)} \) distinguishable sequences of length \( n \).
- channel capacity as log of the maximum number of distinguishable sequences.
Definitions I

- **discrete channel:** $(\mathcal{X}, p(y|x), \mathcal{Y})$.

- **$n$th extension of the discrete memoryless channel:** $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$ with $p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$.

- **no feedback:** $p(y^n|x^n) = \prod_{i=1}^n p(y_i|x_i)$. (default case in the book.)

- **$(M, n)$ code for $(\mathcal{X}, p(y|x), \mathcal{Y})$:**
  1. index set $\{1, 2, \ldots, M\}$.
  2. encoding function $X^n: \{1, 2, \ldots, M\} \rightarrow \mathcal{X}^n$ with codewords $x^n(1), \ldots, x^n(M)$. all codewords form the codebook.
  3. decoding function: $g : \mathcal{Y}^n \rightarrow \{1, 2, \ldots, M\}$. 
Definitions II

- conditional prob. of error: \( \lambda_i = \Pr \{ g(Y^n) \neq i | X^n = x^n(i) \} \).
- maximal prob. of error: \( \lambda^{(n)} = \max_{i \in \{1, \ldots, M\}} \lambda_i \).
- average prob. of error for an \((M, n)\) code: \( P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i \).
- rate of an \((M, n)\) code: \( R = \log(M)/n \) bits per transmission.
- rate \( R \) achievable if there exists a sequence of \((\lceil 2^{nR} \rceil, n)\) codes such that \( \lambda^{(n)} \to 0 \) as \( n \to 0 \).
- capacity is the supremum of all achievable rates.
Jointly typical sequences I

Idea: decode $Y^n$ as index $i$ if $X^n(i)$ is jointly typical with $Y^n$.

The set $A^{(n)}_\epsilon$ of jointly typical sequences $\{(x^n, y^n)\}$ w.r.t. $p(x, y)$ is given by

$$A^{(n)}_\epsilon = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \right. $$

$$\left. \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \right.$$  

$$\left. \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \right.$$  

$$\left. \left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \right\},$$

where $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$. 

Jointly typical sequences II

Joint AEP: Let \((X^n, Y^n)\) have lengths \(n\), drawn i.i.d. from \(p(x^n, y^n)\). Then:

1. \(\Pr \left\{ (X^n, Y^n) \in A^{(n)}_{\epsilon} \right\} \to 1 \text{ as } n \to \infty.\)

2. \(|A^{(n)}_{\epsilon}| \leq 2^{n(H(X,Y) + \epsilon)}\).

3. \(\Pr \left\{ (\tilde{X}^n, \tilde{Y}^n) \in A^{(n)}_{\epsilon} \right\} \leq 2^{-n(I(X; Y) - 3\epsilon)}\) for \((\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)\).

- \(2^{nH(X)}\) typical \(X\) sequences.
- \(2^{nH(Y)}\) typical \(Y\) sequences.
- only \(2^{nH(X,Y)}\) jointly typical sequences.
- one in \(2^{nI(X; Y)}\) pairs is jointly typical.