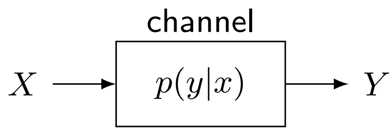
Information Theory

Lecture 4

- Discrete channels, codes and capacity: CT7
  - Channels: CT7.1–2
  - Capacity and the coding theorem: CT7.3–7 and CT7.9
  - Combining source and channel coding: CT7.13

Discrete Channels



- Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets.
- A *discrete channel* is a random mapping $p(y|x)$: $\mathcal{X} \longmapsto \mathcal{Y}$.
- The *nth extension of the discrete channel* is a random mapping $p(y_1^n|x_1^n)$: $\mathcal{X}^n \longmapsto \mathcal{Y}^n$, defined for all $n \geq 1$, $x_1^n \in \mathcal{X}^n$ and $y_1^n \in \mathcal{Y}^n$.
  - A pmf $p(x_1^n)$ induces a pmf $p(y_1^n)$ via the channel,

$$p(y_1^n) = \sum_{x_1^n} p(y_1^n|x_1^n)p(x_1^n)$$

- The channel is *stationary* if for any n

$$p(y_1^n|x_1^n) = p(y_{1+k}^{n+k}|x_{1+k}^{n+k}), \quad k = 1, 2, \ldots$$

- A stationary channel is *memoryless* if

$$p(y_m|x_1^m y_1^{m-1}) = p(y_m|x_m), \quad m = 2, 3, \ldots$$

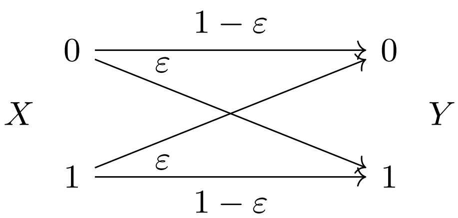
That is, *the channel output at time m does not depend on past inputs or outputs*.

- Furthermore, if the channel is used *without feedback*

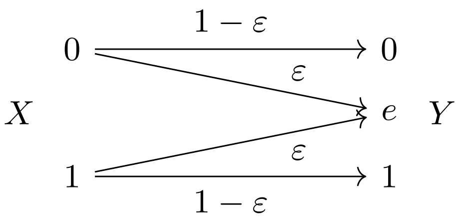
$$p(y_1^n|x_1^n) = \prod_{m=1}^{n} p(y_m|x_m), \quad n = 2, 3, \ldots$$

That is, *each time the channel is used its effect on the output is independent of previous and future uses*.

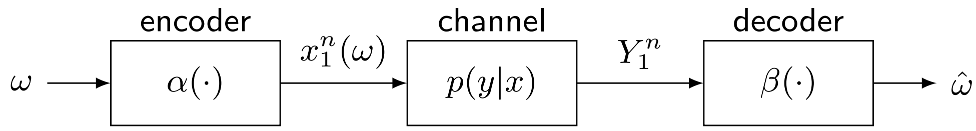
- A *discrete memoryless channel* (DMC) is completely described by the triple $(\mathcal{X}, p(y|x), \mathcal{Y})$.
- The *binary symmetric channel* (BSC) with *crossover probability* ε,
  - a DMC with $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $p(1|0) = p(0|1) = \varepsilon$



- The *binary erasure channel* (BEC) with *erasure probability* ε,
  - a DMC with $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1, e\}$ and $p(e|0) = p(e|1) = \varepsilon$



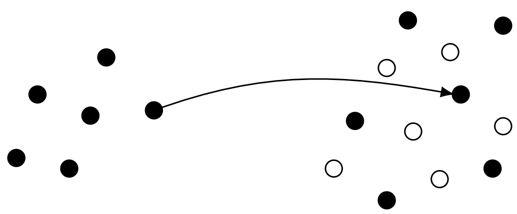
A Block Channel Code



- An (M, n) *block channel code* for a DMC $(\mathcal{X}, p(y|x), \mathcal{Y})$ is defined by:
  1. An *index set* $\mathcal{I}_M \triangleq \{1, \dots, M\}$.
  2. An *encoder mapping* $\alpha : \mathcal{I}_M \longmapsto \mathcal{X}^n$. The set
  $$\mathcal{C}_n \triangleq \left\{ x_1^n : x_1^n = \alpha(i),\ \forall\, i \in \mathcal{I}_M \right\}$$
  of *codewords* is called the *codebook*.
  3. A *decoder mapping* $\beta : \mathcal{Y}^n \longmapsto \mathcal{I}_M$.
- The *rate* of the code is
$$R \triangleq \frac{\log M}{n} \quad \text{[bits per channel use]}$$

Why?

- M different codewords $\{x_1^n(1), \dots, x_1^n(M)\}$ can convey $\log M$ bits of *information* per codeword, or R bits per channel use.
- Consider $M = 2^k$, $|\mathcal{X}| = 2$, and assume that $k < n$. Then k "information bits" are mapped into $n > k$ "coded bits." Introduces *redundancy*; can be employed by the decoder to *correct channel errors*.



Error Probabilities

- Information symbol $\omega \in \mathcal{I}_M$, with $p(i) = \Pr(\omega = i)$. Then, for a given DMC and a given code

$$\omega \longrightarrow X_1^n = \alpha(\omega) \longrightarrow Y_1^n \longrightarrow \hat{\omega} = \beta(Y_1^n)$$

- Define:

1. The *conditional* error probability: $\lambda_i = \Pr(\hat{\omega} \neq i | \omega = i)$
2. The *maximal* error probability: $\lambda^{(n)} = \max\{\lambda_1, \dots, \lambda_M\}$
3. The *average* error probability:

$$P_e^{(n)} = \Pr(\hat{\omega} \neq \omega) = \sum_{i=1}^{M} \lambda_i \, p(i)$$

Jointly Typical Sequences

The set $A_\varepsilon^{(n)}$ of *jointly typical sequences* with respect to a pmf $p(x,y)$ is the set $\{(x_1^n, y_1^n)\}$ of sequences for which

$$\left| -\frac{1}{n} \log p(x_1^n) - H(X) \right| < \varepsilon$$

$$\left| -\frac{1}{n} \log p(y_1^n) - H(Y) \right| < \varepsilon$$

$$\left| -\frac{1}{n} \log p(x_1^n, y_1^n) - H(X,Y) \right| < \varepsilon$$

where

$$p(x_1^n, y_1^n) = \prod_{m=1}^{n} p(x_m, y_m)$$

$$p(x_1^n) = \sum_{y_1^n} p(x_1^n, y_1^n), \qquad p(y_1^n) = \sum_{x_1^n} p(x_1^n, y_1^n)$$

and where the entropies are computed based on $p(x,y)$.

The joint AEP

Let (X_1^n, Y_1^n) drawn according to $p(x_1^n, y_1^n) = \prod_{m=1}^n p(x_m, y_m)$

- $\Pr\big((X_1^n, Y_1^n) \in A_\varepsilon^{(n)}\big) > 1 - \varepsilon$ for sufficiently large n.
- $|A_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)}$.
- $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)2^{n(H(X,Y)-\varepsilon)}$ for sufficiently large n.
- If $\tilde{X}_1^n$ and $\tilde{Y}_1^n$ are drawn independently according to $p(x_1^n) = \sum_{y_1^n} p(x_1^n, y_1^n)$ and $p(y_1^n) = \sum_{x_1^n} p(x_1^n, y_1^n)$, then

$$\Pr\big((\tilde{X}_1^n, \tilde{Y}_1^n) \in A_\varepsilon^{(n)}\big) \leq 2^{-n(I(X;Y)-3\varepsilon)}$$

  and for sufficiently large n

$$\Pr\big((\tilde{X}_1^n, \tilde{Y}_1^n) \in A_\varepsilon^{(n)}\big) \geq (1-\varepsilon)2^{-n(I(X;Y)+3\varepsilon)}$$

  with $I(X;Y)$ computed for the pmf $p(x,y)$.

Channel Capacity

- For a fixed n, a code can convey more information for large M $\implies$ we would like to *maximize the rate* $R = \frac{1}{n}\log M$ without sacrificing performance
  - Which is the largest R that allows for a (very) low $P_e^{(n)}$??
- For a given channel we say that the rate R is *achievable* if there exists a sequence of (M,n) codes, with $M = \lceil 2^{nR} \rceil$, such that the maximal probability of error $\lambda^{(n)} \to 0$ as $n \to \infty$.

The *capacity* C of a channel is the *supremum of all rates that are achievable over the channel*.

Random Code Design

- Choose a joint pmf $p(x_1^n)$ on $\mathcal{X}^n$.
- *Random code design*: Draw M codewords $x_1^n(i)$, $i = 1, \ldots,$ M, i.i.d. according to $p(x_1^n)$ and let these define a codebook

$$\mathcal{C}_n = \left\{x_1^n(1), \ldots, x_1^n(M)\right\}.$$

- *Note*: The interpretation here is that the codebook is *"designed" in a random fashion*. When the resulting code then is *used*, the codebook must, of course, be fixed and known. . .

A Lower Bound for C of a DMC

- A DMC $(\mathcal{X}, p(y|x), \mathcal{Y})$.
- Fix a pmf $p(x)$ for $x \in \mathcal{X}$.
  Generate $\mathcal{C}_n = \{x_1^n(1), \ldots, x_1^n(M)\}$ using $p(x_1^n) = \prod p(x_m)$.
- A data symbol ω is generated according to a *uniform distribution* on $\mathcal{I}_M$, and $x_1^n(\omega)$ is transmitted.
- The channel produces a corresponding output sequence Y_1^n.
- Let $A_\varepsilon^{(n)}$ be the typical set w.r.t. $p(x, y) = p(y|x)p(x)$.
  At the receiver, the decoder then uses the following decision rule. Index $\hat{\omega}$ was sent if:
  - $\left(x_1^n(\hat{\omega}), Y_1^n\right) \in A_\varepsilon^{(n)}$ for some small ε;
  - no other ω corresponds to a jointly typical $\left(x_1^n(\omega), Y_1^n\right)$.

Now study

$$\pi_n = \Pr(\hat{\omega} \neq \omega)$$

where "Pr" is over the random codebook selection, the data variable ω and the channel.

- Symmetry $\implies \pi_n = \Pr(\hat{\omega} \neq 1 | \omega = 1)$
- Let

$$E_i = \{(x_1^n(i), Y_1^n) \in A_\varepsilon^{(n)}\}$$

then for a sufficiently large n,

$$\begin{aligned} \pi_n &= P(E_1^c \cup E_2 \cup \cdots \cup E_M) \le P(E_1^c) + \sum_{i=2}^{M} P(E_i) \\ &\le \varepsilon + (M-1)2^{-n(I(X;Y)-3\varepsilon)} \le \varepsilon + 2^{-n(I(X;Y)-R-3\varepsilon)} \end{aligned}$$

because of the union bound and the joint AEP.

- Note that

$$I(X;Y) = \sum_{x,y} p(y|x)p(x) \log \frac{p(y|x)}{p(y)}$$

with $p(y) = \sum_x p(y|x)p(x)$, where $p(x)$ generated the random codebook and $p(y|x)$ is given by the channel.

- Let $\mathcal{C}_{\text{tot}}$ be the set of all possible codebooks that can be generated by $p(x_1^n) = \prod p(x_m)$, then *at least one* $\mathcal{C}_n \in \mathcal{C}_{\text{tot}}$ must give

$$P_e^{(n)} \le \pi_n \le \varepsilon + 2^{-n(I(X;Y)-R-3\varepsilon)}$$

$\implies$ as long as $R < I(X;Y) - 3\varepsilon$ there exists at least one $\mathcal{C}_n \in \mathcal{C}_{\text{tot}}$, say $\mathcal{C}_n^*$, that can give $P_e^{(n)} \to 0$ as $n \to \infty$.

- Order the codewords in $\mathcal{C}_n^*$ according to the corresponding λ_i's and throw away the worst half $\implies$
  - new rate $R' = R - n^{-1}$
  - for the remaining codewords

$$\frac{\lambda^{(n)}}{2} \leq \varepsilon + 2^{-n(I(X;Y)-R-3\varepsilon)}$$

$\implies$ for any $p(x)$, all rates $R < I(X;Y) - 3\varepsilon$ achievable
$\implies$ all rates $R < \max_{p(x)} I(X;Y) - 3\varepsilon$ achievable $\implies$

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

An Upper Bound for C of a DMC

- Let $\mathcal{C}_n = \{x_1^n(1), \ldots, x_1^n(M)\}$ be any sequence of codes that can achieve $\lambda^{(n)} \to 0$ at a fixed rate $R = \frac{1}{n} \log M$.
- Note that $\lambda^{(n)} \to 0 \implies P_e^{(n)} \to 0$ for any $p(\omega)$.
  We can assume $\mathcal{C}_n$ encodes equally probable $\omega \in \mathcal{I}_M$.
- Fano's inequality $\implies$

$$R \leq \frac{1}{n} + P_e^{(n)} R + \frac{1}{n} I(x_1^n(\omega); Y_1^n) \leq \frac{1}{n} + P_e^{(n)} R + \max_{p(x)} I(X;Y)$$

That is, for any fixed achievable R

$$\lambda^{(n)} \to 0 \implies R \leq \max_{p(x)} I(X;Y) \implies C \leq \max_{p(x)} I(X;Y)$$

The Channel Coding Theorem for DMC's

Theorem (the channel coding theorem)

For a given DMC $(\mathcal{X}, p(y|x), \mathcal{Y})$*, let* $p(x)$ *be a pmf on* $\mathcal{X}$ *and let*

$$C = \max_{p(x)} I(X;Y)$$
$$= \max_{p(x)} \left\{ \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(y|x)p(x) \log \frac{p(y|x)}{\sum_{x \in \mathcal{X}} p(y|x)p(x)} \right\}$$

Then C *is the capacity of the channel. That is, all rates* $R < C$ *and no rates* $R > C$ *are achievable.*

The Joint Source–Channel Coding Theorem

- A (stationary and ergodic) discrete source $\mathcal{S}$ with entropy rate $H(\mathcal{S})$ [bits/source symbol].
  - A length-L block of source symbols can be coded into k bits, and then reconstructed without errors as long as $k/L > H(\mathcal{S})$ and as $L \to \infty$.
- A DMC $(\mathcal{X}, p(y|x), \mathcal{Y})$ with capacity C [bits/channel use].
  - If $k/n < C$ a channel code exists that can convey k bits of information per n channel uses without errors as $n \to \infty$.
- L source symbols $\to$ k information bits $\to$ n channel symbols; will convey the source symbols without errors as long as

$$H(\mathcal{S}) < \frac{k}{L} < \frac{n}{L} \cdot C$$

- Hence, as long as $H(\mathcal{S}) < C$ [bits/source symbol] the source can be transmitted without errors, as both $L \to \infty$ and $n \to \infty$ (assuming $n/L = 1$).
- If $H(\mathcal{S}) > C$ there is *no way* of constructing a system with an error probability that is not bounded away from zero. (*Fano's inequality, etc.*)
- *No system* exists that can communicate a source without errors for $H(\mathcal{S}) > C$. *One way* of achieving error-free performance, for $H(\mathcal{S}) < C$, is to use separate source and channel coding.