Information Theory

Lecture 11

- When is channel capacity maximum mutual information, and when it's not what is it then?
- Based on: S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. on IT*, July 1994

Motivating Example

- Binary channel: $Y_m = X_m + Z_m$, memoryless noise Z_m
- $W \in \{0, 1\}$, $p = \Pr(W = 1)$. A value $W = w$ drawn once:
  - $w = 1 \Rightarrow \Pr(Z_m = 1) = \alpha \leq 1/2, \ \forall m$
  - $w = 0 \Rightarrow \Pr(Z_m = 1) = \beta \leq \alpha, \ \forall m$
- Let

$$
\begin{aligned}
C(w) &= \max_{p(x)} I(X; Y|W = w) \\
&= 1 - w\, h(\alpha) - (1 - w)\, h(\beta)
\end{aligned}
$$

with $h(t) = -t \log t - (1 - t) \log(1 - t)$

- Is any of these entities the true capacity of the channel?
  - $C_1 = E[C(W)] = 1 - ph(\alpha) - (1-p)h(\beta)$
  - $C_2 = \max_w C(w) = C(0) = 1 - h(\beta)$
  - $C_3 = \min_w C(w) = C(1) = 1 - h(\alpha)$
- E.g., $\alpha = 1/2$, $\beta = 0$, $p = 1/2 \Rightarrow$
  - $C_1 = 1/2$
  - $C_2 = \max_w C(w) = C(0) = 1$
  - $C_3 = \min_w C(w) = C(1) = 0$

  (in bits per channel use)
- *Is there a general formula that always holds?*

Definitions

- A sequence $\{X_n\}$ of discrete random variables;
  $x_1^N = (x_1, \ldots, x_N)$, $p(x_1^N) = \Pr(X_1^N = x_1^N)$;
  $p_X = \{p(x_1^n)\}_{n=1}^{\infty}$
- *Entropy*
$$H(X_1^N) = E[-\log p(X_1^N)]$$
- *Entropy rate*
$$\bar{H}(X) = \lim_{N\to\infty} \frac{1}{N} H(X_1^N)$$

- *Limit in probability* of $\{X_n\}$

$$\tilde{x} = \operatorname*{l.p}_{n\to\infty} X_n$$

if for any $\varepsilon > 0$ it holds that $\Pr(|\tilde{x} - X_n| > \varepsilon) \to 0$ as $n \to \infty$

- *Liminf in probability* of $\{X_n\}$

$$\tilde{x} = \operatorname*{l.inf.p}_{n\to\infty} X_n$$

if $\tilde{x} =$ *supremum of all* x *for which* $\Pr(X_n \leq x) \to 0$ *as* $n \to \infty$

- *Limsup in probability* of $\{X_n\}$

$$\tilde{x} = \operatorname*{l.sup.p}_{n\to\infty} X_n$$

if $\tilde{x} =$ *infimum of all* x *for which* $\Pr(X_n \geq x) \to 0$ *as* $n \to \infty$

- $\tilde{x} = \mathrm{l.p}\, X_n < \infty$ exists $\implies$ $\tilde{x} = \mathrm{l.inf.p}\, X_n = \mathrm{l.sup.p}\, X_n$

- A two-dimensional sequence $\{(X_n, Y_n)\}$.
  Component-sequences $\{X_n\}$ and $\{Y_n\}$,
  $p(x_1^N, y_1^N) = \Pr(X_1^N = x_1^N, Y_1^N = y_1^N)$
- *Information density*:

$$I_N = I_N(X_1^N, Y_1^N) = \log \frac{p(X_1^N, Y_1^N)}{p(X_1^N)p(Y_1^N)}$$

- *Mutual information*:

$$I(X_1^N; Y_1^N) = E\left[\log \frac{p(X_1^N, Y_1^N)}{p(X_1^N)p(Y_1^N)}\right] = E[I_N]$$

- *Information rate*:

$$\bar{I} = \lim_{N\to\infty} \frac{1}{N} I(X_1^N; Y_1^N)$$

- *Ergodicity* (c.f., CT15.7): Let $\{X_n\}$ be a process described by p_X. Let $\mathbf{X} = \cdots, X_{-1}, X_0, X_1, \cdots$ be an infinite sequence drawn from $\{X_n\}$ and let $\mathbf{X}^{(t)}$ denote $\mathbf{X}$ shifted t positions in time, that is, $X_n^{(t)} = X_{n+t}$. The process $\{X_n\}$ is *ergodic* if for any $\mathbf{X}$ and any t, it holds that $\Pr(\mathbf{X} = \mathbf{X}^{(t)}) = 0$ or 1.
- Let $g_n(\mathbf{x})$ be a function of the components x_1^n in $\mathbf{x}$. A discrete and *stationary* process $\{X_n\}$ is ergodic iff for all $n \geq 1$ and all $g_n(\mathbf{X})$ with $E[|g_n(\mathbf{X})|] < \infty$ it holds that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} g_n(\mathbf{X}^{(t)}) = E[g_n(\mathbf{X})] \quad \text{w.p.1}$$

- For a discrete stationary and ergodic $\{X_n\}$,

$$\lim_{N\to\infty} \frac{1}{N} \log p(X_1^N) = \bar{H}(X) \quad \text{w.p.1}$$

- A *channel* with discrete input X_n and output Y_n is described, for all N, by

$$p(y_1^N|x_1^N) = \Pr(Y_1^N = y_1^N \,|\, X_1^N = x_1^N)$$

- Transmitting a uniform random variable $\omega \in \{1, \dots, M\}$ over a channel $p(y_1^N|x_1^N)$ using an (N, M) *code*:
  - *Encoding* $x_1^N = \alpha(\omega)$; *decoding* $\hat{\omega} = \beta(y_1^N) \in \{1, \dots, M\}$; *rate* $R = N^{-1} \log M$; *probability of error* $P_e^{(N)} = P(\hat{\omega} \neq \omega)$
  - A rate R is *achievable*[1] if there exists a sequence of (N, M) codes with rate R such that $P_e^{(N)} \to 0$ as $N \to \infty$
- The *capacity* C of a channel $p(y_1^N|x_1^N)$ is *the supremum of all achievable rates* for that channel

[1] In terms of $P_e^{(N)} \to 0$; can be strengthened to $P_{\max} \to 0$

Results

- *Feinstein's Lemma* (1954): Fix N and a channel $p(y_1^N|x_1^N)$. For any $p(x_1^N)$ and $\gamma > 0$, there exists an (N, M) code with rate R for which

$$P_e^{(N)} \leq \Pr\left(N^{-1} I_N \leq R + \gamma\right) + e^{-\gamma N}$$

where

$$I_N = \log \frac{p(Y_1^N, X_1^N)}{p(X_1^N) p(Y_1^N)}$$

- *Corollary* [Verdú and Han, 94]:

$$C \geq \sup_{p_X} \left\{ \operatorname{l.inf.p}_{n \to \infty} \frac{1}{n} I_n \right\}$$

- *Theorem* [Verdú and Han, 94]: For every $\gamma > 0$, using any (N, M) code with rate R in coding a uniform $\omega \in \{1, \ldots, M\}$ over a channel $p(y_1^N|x_1^N)$ results in

$$P_e^{(N)} \geq \Pr\left(N^{-1} I_N \leq R - \gamma\right) - e^{-\gamma N}$$

- *Corollary*:

$$C \leq \sup_{p_X} \left\{ \operatorname{l.inf.p}_{n \to \infty} \frac{1}{n} I_n \right\}$$

- **A general formula for channel capacity**:

$$\boxed{C = \sup_{p_X} \left\{ \operatorname{l.inf.p}_{n \to \infty} \frac{1}{n} I_n \right\}}$$

An Alternative Lower Bound for C

- Fix $p(x_1^n)$. Let $p(x_1^n, y_1^n)$ describe X_1^n and Y_1^n when the channel is driven by X_1^n, and let

$$\mathcal{I} = \operatorname*{l.inf.p}_{n\to\infty} \frac{1}{n} I_n$$

- Let $T_\varepsilon^{(n)}$, $\varepsilon > 0$, be the set of sequences (x_1^n, y_1^n) for which

$$\frac{1}{n} \log \frac{p(y_1^n|x_1^n)}{p(y_1^n)} > \mathcal{I} - \varepsilon$$

  - C.f. the definition of typical sequences
  - Notice that $(x_1^n, y_1^n) \in T_\varepsilon^{(n)} \Rightarrow p(y_1^N) < p(y_1^N|x_1^N)2^{-n(\mathcal{I}-\varepsilon)}$

- Generate $\mathcal{C}_n = \{x_1^n(1), \dots, x_1^n(M)\}$ using $p(x_1^n)$.
- A data symbol ω is generated according to a uniform distribution on $\{1, \dots, M\}$, and $x_1^n(\omega)$ is transmitted.
- The channel produces a corresponding output sequence Y_1^n
- The decoder uses the following decision rule:
  - Index $\hat\omega$ was sent if: $\left(x_1^n(\hat\omega), Y_1^n\right) \in T_\varepsilon^{(n)}$ for some small ε, and no other $\hat\omega$
- 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 $E_i = \{(X_1^n(i), Y_1^n) \in T_\varepsilon^{(n)}\}$, then

$$\pi_n = P(E_1^c \cup E_2 \cup \dots \cup E_M) \leq P(E_1^c) + \sum_{i=2}^{M} P(E_i)$$

- It holds that $P(E_1^c) \to 0$ because of the definition of $\mathcal{I}$
- Also, for $i > 1$

$$P(E_i) = \sum_{(x,y)\in T} p(x)p(y) < \sum_{(x,y)\in T} p(x)p(y|x)2^{-n(\mathcal{I}-\varepsilon)} < 2^{-n(\mathcal{I}-\varepsilon)}$$

where $x = x_1^n$, $y = y_1^n$ and $T = T_\varepsilon^{(n)}$, and consequently

$$\sum_{i=2}^{M} P(E_i) < (M-1)2^{-n(\mathcal{I}-\varepsilon)} \leq 2^{-n(\mathcal{I}-R-\varepsilon)}$$

$$\boxed{R = \mathcal{I} - 2\varepsilon \text{ is achievable!}}$$

Discrete Memoryless Channels

- For a discrete (stationary and) *memoryless* channel (DMC),

$$p(y_1^N|x_1^N) = p(y_1|x_1)\cdots p(y_N|x_N)$$

- In [Verdú and Han, 94] it is shown that the $p(x_1^N)$ that achieves the supremum in the formula for C is of the form

$$p(x_1^N) = p(x_1)\cdots p(x_N)$$

Hence,

$$\text{l.inf.p}\ \frac{1}{n} I_N(X_1^n; Y_1^n) = I(X;Y)$$

evaluated for $p(x) = p(x_1)$ and $p(y|x) = p(y_1|x_1)$, since information density converges to mutual information.

- Thus, we get Shannon's formula

$$C = \max_{p(x_1)} I(X_1;Y_1)$$

When is Mutual Information Relevant to Capacity?

- If at least one of $\mathcal{X}$ and $\mathcal{Y}$ is finite, then

$$\operatorname*{l.inf.p}_{n\to\infty} \frac{1}{n} I_n \leq \liminf_{n\to\infty} \frac{1}{n} I(X_1^n; Y_1^n)$$
$$\leq \limsup_{n\to\infty} \frac{1}{n} I(X_1^n; Y_1^n) \leq \operatorname*{l.sup.p}_{n\to\infty} \frac{1}{n} I_n$$

- *Corollary*: If at least one of $\mathcal{X}$ and $\mathcal{Y}$ is finite, and if

$$\sup_{p_X} \operatorname*{l.inf.p}_{n\to\infty} \frac{1}{n} I_n = \sup_{p_X} \operatorname*{l.sup.p}_{n\to\infty} \frac{1}{n} I_n$$

  then

$$C = \sup_{p_X} \bar{I} = \lim_{n\to\infty} \frac{1}{n} \sup_{p(x_1^n)} I(X_1^n; Y_1^n)$$

- A channel is *stationary and ergodic* if $\{(X_n, Y_n)\}$ is stationary and ergodic for all stationary and ergodic $\{X_n\}$
- For a stationary and ergodic channel

$$C = \sup_{p_X} \left\{ \operatorname*{l.inf.p}_{n\to\infty} \frac{1}{n} I_n \right\} = \sup_{p_X} \left\{ \operatorname*{l.p}_{n\to\infty} \frac{1}{n} I_n \right\} = \sup_{p_X} \bar{I}$$
$$= \lim_{n\to\infty} \frac{1}{n} \sup_{p(x_1^n)} I(X_1^n; Y_1^n)$$

Some Binary Channel Models

- A general binary channel

$$Y_m = X_m + Z_m$$

where Z_m is drawn according to an *arbitrary* binary random process

- **Capacity**

$$C = \sup_{p_X} \left\{ \operatorname{l.inf.p}_{n\to\infty} \frac{1}{n} I_n \right\} = 1 - \operatorname{l.sup.p}_{n\to\infty} \frac{1}{n} \log \frac{1}{p(Z_1^n)}$$

- $\{Z_n\}$ *stationary and memoryless*:

$$C = 1 - h(p)$$

where

$$h(p) = -p \log p - (1-p)\log(1-p) = H(Z_1)$$

with $p = \Pr(Z_n = 1)$, since

$$\operatorname{l.sup.p}_{n\to\infty} \frac{1}{n} \log \frac{1}{p(Z_1^n)} = H(Z_1)$$

- $\{Z_n\}$ *stationary and ergodic*:

$$C = 1 - \bar{H}(Z)$$

since

$$\operatorname*{l.\,sup.\,p}_{n\to\infty} \frac{1}{n} \log \frac{1}{p(Z_1^n)} = \lim_{n\to\infty} \frac{1}{n} H(Z_1^n) = \bar{H}(Z)$$

- $\{Z_n\}$ *stationary and nonergodic* (c.f. previous example): with probability q, $Z_n = 0$ for all n and with probability $(1-q)$, $\{Z_n\}$ is stationary and memoryless with $p = P(Z_n = 1) \Rightarrow$

$$C = 1 - h(p)$$

since

$$\operatorname*{l.\,sup.\,p}_{n\to\infty} \frac{1}{n} \log \frac{1}{p(Z_1^n)} = \max\{0, h(p)\}$$

  - Capacity determined by the "worst case" noise!
  - E.g., $p = 1/2 \Rightarrow C = 0$ (no rates achievable)!
- In this example $C^* = \max_{p(x)} I(X;Y)$ will not give the correct value for capacity ($C^* = p$)

Summary

- Shannon's formula

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

holds for *stationary and memoryless* channels

- For the class of *information stable channels*, it generalizes to

$$C = \lim_{n\to\infty} \sup_{p(x_1^n)} \frac{1}{n} I(X_1^n; Y_1^n)$$

(e.g., stationary and ergodic channels)

- The formula

$$\boxed{C = \sup_{p_X} \operatorname{l.inf.p}_{n\to\infty} \frac{1}{n} I_n}$$

holds for *any* channel $p(y_1^N | x_1^N)$

  - *Ergodicity* is the key to formulas based on mutual information; $n^{-1} I_n$ needs to converge to a non-random entity
  - E.g., in the nonergodic binary channel example

$$n^{-1} I_n \to \begin{cases} H(X_1), & \text{with prob. } q \\ 1/2, & \text{with prob. } (1-q) \end{cases}$$