A General Formula for Channel Capacity

1 Definitions

- Information variable $\omega \in \{1, \ldots, M\}$, $p(i) = \Pr(\omega = i)$
- Channel input $X \in \mathcal{X}$ and output $Y \in \mathcal{Y}$, finite alphabets
- Codewords $\{x_1^N(i) : i = 1, \ldots, M\}$, $x_n \in \mathcal{X}$
- Rate $R = N^{-1} \ln M$
- A sequence of channel uses,

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

  defined for for each N, including $N \to \infty$

  - a *discrete* channel with *completely arbitrary memory behavior*
- Decoder,

  $$\hat{\omega} = i \ \text{ if } \ Y_1^N \in F_i$$

  where $\{F_i\}$ is a partition of $\mathcal{Y}^N$
- Error probabilities,

  $$P_e^{(N)} = \sum_{i=1}^{M} \Pr\left(Y_1^N \in F_i^c | X_1^N = x_1^N(i)\right) p(i)$$
  $$\lambda^{(N)} = \max\left\{ \Pr\left(Y_1^N \in F_i^c | X_1^N = x_1^N(i)\right)\right\}_{i=1}^{M}$$
- Information density

  $$i_N(x_1^N; y_1^N) = \ln \frac{p(x_1^N, y_1^N)}{p(x_1^N)p(y_1^N)}$$
- Liminf in probability of $\{A_n\}$,

  $$\alpha = \text{liminfp}\,\{A_n\}$$

  = supremum of all α for which $\Pr(A_n \le \alpha) \to 0$ as $n \to \infty$
- Rate R *achievable* if there exists a sequence of codes such that $\lambda^{(N)} \to 0$ when $N \to \infty$
- C = supremum of all achievable rates

2 Feinstein's Lemma and a Converse

Lemma 1 *Given M and $a > 0$ and an input distribution $p(x_1^N)$, there exist $x_1^N(i) \in \mathcal{X}^N, i = 1, \ldots, M$, and a partition $F_1, \ldots, F_M$ of $\mathcal{Y}^N$ such that*

$$\Pr\left(Y_1^N \notin F_i | X_1^N = x_1^N(i)\right) \leq Me^{-a} + \Pr\left(i_N(X_1^N; Y_1^N) \leq a\right)$$

In particular, choosing $a = \ln M + N\gamma$, with $\gamma > 0$, gives

$$\Pr\left(Y_1^N \notin F_i | X_1^N = x_1^N(i)\right) \leq e^{-\gamma N} + \Pr\left(\frac{1}{N} i_N(X_1^N; Y_1^N) \leq \frac{1}{N} \ln M + \gamma\right)$$

Lemma 1 (Feinstein's Lemma [1]) implies that for any given $p(x_1^N)$ there exists a code of rate R such that, for any $\gamma > 0$ and $N > 0$

$$\lambda^{(N)} \leq e^{-\gamma N} + \Pr\left(\frac{1}{N} i_N(X_1^N; Y_1^N) \leq R + \gamma\right)$$

where

$$i_N(x_1^N; y_1^N) = \ln \frac{p(x_1^N, y_1^N)}{p(x_1^N)p(y_1^N)} = \ln \frac{p(y_1^N | x_1^N)}{\sum_{x_1^N} p(y_1^N | x_1^N) p(x_1^N)}$$

for the given $p(x_1^N)$ and $p(y_1^N|x_1^N)$ (the latter given by the channel in consideration).

Proof

We use the notation $x = x_1^N$, $y = y_1^N$, $\bar{X} = \mathcal{X}^N$ and $\bar{Y} = \mathcal{Y}^N$, for simplicity, where N is the fixed codeword length. Define $G = \{(x, y) : i_N(x, y) > a\}$. Set

$$\varepsilon = Me^{-a} + \Pr(i_N \leq a) = Me^{-a} + P(G^c)$$

and assume $\varepsilon < 1$ and hence also that $P(G^c) \leq \varepsilon < 1$ and therefore that

$$\Pr(i_N > a) = P(G) > 1 - \varepsilon > 0$$

Letting $G_x = \{y : (x, y) \in G\}$ this implies that in defining

$$A = \{x : P(G_x | x) > 1 - \varepsilon\}$$

it holds that $P(A) > 0$. Choose $x_1 \in A$ and let $F_1 = G_{x_1}$. Next choose if possible $x_2 \in A$ such that $P(G_{x_2} - F_1 | x_2) > 1 - \varepsilon$ and let $F_2 = G_{x_2} - F_1$. Continue in this way until either M points have been selected or all points in A have been exhausted. That is, given $\{x_j, F_j\}, j = 1, \ldots, i - 1$, find an $x_i \in A$ for which

$$P(G_{x_i} - \bigcup_{j<i} F_j | x_i) > 1 - \varepsilon$$

and let $F_i = G_{x_i} - \bigcup_{j<i} F_j$. If this terminates before M points have been collected, denote the final point's index by n. Observe that

$$P(F_i^c | x_i) \leq P(G_{x_i}^c | x_i) \leq \varepsilon, \;\; i = 1, \ldots, n$$

and hence the lemma will be proved if we can show that n cannot be strictly less than M.

Define $F = \bigcup_{i=1}^{n} F_i$ and consider the probability

$$P(G) = P(G \cap (\bar{X} \times F)) + P(G \cap (\bar{X} \times F^c))$$

The first term is bounded as

$$P(G \cap (\bar{X} \times F)) \le P(\bar{X} \times F) = P(F) = \sum_{i=1}^{n} P(F_i)$$

Let

$$f(x,y) = \frac{p(x,y)}{p(x)p(y)}$$

(i.e., $i_N = \ln f(x,y)$). We get

$$\begin{aligned} P(F_i) &= \sum_{y \in F_i} p(y) \le \sum_{y \in G_{x_i}} p(y) \le \sum_{y \in G_{x_i}} \frac{f(x_i,y)}{e^a} p(y) \\ &\le e^{-a} \sum_{y} p(y|x_i) = e^{-a} \end{aligned}$$

and hence

$$P(G \cap (\bar{X} \times F)) \le n e^{-a}$$

Now consider

$$\begin{aligned} P(G \cap (\bar{X} \times F^c)) &= \sum_{x} P(G \cap (\bar{X} \times F^c)|x)p(x) \\ &= \sum_{x} P(G_x \cap F^c|x)p(x) = \sum_{x} P(G_x - \bigcup_{i=1}^{n} F_i|x)p(x) \end{aligned}$$

Defining

$$B = \{x : P(G_x - \bigcup_{i=1}^{n} F_i|x) > 1 - \varepsilon\}$$

it must hold that $P(B) = 0$, or there would be a point x_{n+1} for which

$$P(G_{x_{n+1}} - \bigcup_{i=1}^{n+1} F_i|x_{n+1}) > 1 - \varepsilon$$

Hence

$$P(G \cap (A \times F^c)) \le 1 - \varepsilon$$

so we get

$$P(G) \le n e^{-a} + 1 - \varepsilon$$

From the definition of ε we have also that

$$P(G) = 1 - P(G^c) = 1 - \varepsilon + M e^{-a}$$

so $M \le n$ must hold, completing the proof.

Let a *reliable code sequence* be a sequence of codes that achieve $\lambda^{(N)} \to 0$ at a fixed rate $R < C$. Since

$$\bar{P}_e^{(N)} \triangleq \frac{1}{M} \sum_{i=1}^{M} P\big(F_i^c | x_1^N(i)\big) \le \lambda^{(N)}$$

it holds, for a reliable code sequence, that $\bar{P}_e^{(N)} \to 0$ for any $\{p(i)\}$. Hence if a sequence of codes gives

$$\bar{P}_e^{(N)} > 0$$

for all N, the sequence cannot be reliable. Thus, to prove a converse we can assume, without loss of generality, that $p(i) = M^{-1}$ and study the resulting average error probability $P_e^{(N)}$.

The following lemma is adopted from [2].

Lemma 2 *Assume that $\{x_1^N(i)\}_{i=1}^M$ is the codebook of any code used in encoding equiprobable information symbols $\omega \in \{1, \dots, M\}$, and let $\{F_i\}_{i=1}^M$ be the corresponding decoding sets. Then*

$$\begin{aligned} P_e^{(N)} &= \sum_{i=1}^{M} \frac{1}{M} \Pr\left(Y_1^N \notin F_i | X_1^N = x_1^N(i)\right) \\ &\geq \Pr\left(N^{-1} i_N(X_1^N; Y_1^N) \leq N^{-1} \ln M - \gamma\right) - e^{-\gamma N} \end{aligned}$$

for any $\gamma > 0$, and where $i_N(x_1^N; y_1^N)$ is evaluated with $p(x_1^N) = 1/M$.

Proof

As before, we use the notation $x = x_1^N$, $y = y_1^N$, where N is the fixed codeword length. Let $\varepsilon = P_e^{(N)}$, $\beta = e^{-\gamma N}$, and

$$L = \{(x, y) : p(x|y) \leq \beta\}$$

and note that

$$P(L) = \Pr\left(p(x|y) \leq e^{-\gamma N}\right) = \Pr(N^{-1}\, i_N \leq N^{-1} \ln M - \gamma)$$

We hence need to show that

$$P(L) \leq \varepsilon + \beta$$

holds for any code $\{x_i\}$, with $x_i = x_1^N(i)$ and decoding sets $\{F_i\}$. Letting

$$L_i = \{y : p(x_i|y) \leq \beta\}$$

we can write

$$\begin{aligned} P(L) &= \sum_i M^{-1} P(L_i|x_i) = \sum_i M^{-1} P(L_i \cap F_i^c|x_i) + \sum_i M^{-1} P(L_i \cap F_i|x_i) \\ &\leq \sum_i M^{-1} P(F_i^c|x_i) + \sum_i M^{-1} P(L_i \cap F_i|x_i) \\ &= \varepsilon + \sum_i \sum_{y \in L_i \cap F_i} p(x_i|y) p(y) \leq \varepsilon + \beta \sum_i \sum_{y \in L_i \cap F_i} p(y) \\ &\leq \varepsilon + \beta \sum_i \sum_{y \in F_i} p(y) \leq \varepsilon + \beta \end{aligned}$$

A General Formula for Channel Capacity [2]

Theorem 1

$$\boxed{C = \sup_{\{p(x_1^N)\}} \left\{ \operatorname{liminfp} \frac{1}{N} i_N(X_1^N; Y_1^N) \right\}}$$

where the supremum is over all possible sequences $\{p(x_1^N)\} = \{p(x_1^N)\}_{N=1}^{\infty}$.

Proof

Let

$$R^* = \operatorname{liminfp} \frac{1}{N} i_N(X_1^N; Y_1^N)$$

for any given $\{p(x_1^N)\}$, and let

$$C^* = \sup_{\{p(x_1^N)\}} R^*$$

For any $\delta > 0$ assume $R = R^* - \delta$. In Feinstein's lemma, fix N, let $\gamma = \delta/2$, and note that

$$\Pr\left(\frac{1}{N} i_N(X_1^N; Y_1^N) \leq R + \delta/2\right) = \Pr\left(\frac{1}{N} i_N(X_1^N; Y_1^N) \leq R^* - \delta/2\right)$$

and because of the definition of R^*

$$\lim_{N\to\infty} \Pr\left(\frac{1}{N} i_N(X_1^N; Y_1^N) \leq R^* - \delta/2\right) = 0$$

Thus R is an achievable rate for any $\{p(x_1^N)\}$ and $\delta > 0$, which means that $C \geq C^*$.

Now assume for $\gamma > 0$ that $R = C^* + 2\gamma$ is the rate of any code of length N that codes equally likely symbols, and note in that case that

$$\Pr\left(N^{-1} i_N(X_1^N; Y_1^N) \leq R - \gamma\right) = \Pr\left(N^{-1} i_N(X_1^N; Y_1^N) \leq C^* + \gamma\right)$$

As $N \to \infty$ this probability cannot vanish, due to the definition of C^*. Hence by Lemma 2, R is not achievable for any γ, which means that $C \leq C^*$.

3 Example

Assume that $p(y_1^N|x_1^N) = p(y_1|x_1)\cdots p(y_N|x_N)$ (stationary and memoryless channel). In [2, Theorem 10] it is shown that for such channels the $p(x_1^N)$ that achieves the supremum in the formula for C is of the form

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

That is, the optimal input distribution is stationary and memoryless. Hence, assuming this form for $p(x_1^N)$ it holds that

$$\operatorname{liminfp} \frac{1}{N} i_N(X_1^N; Y_1^N) = I(X; Y)$$

evaluated for $p(x) = p(x_1)$ and $p(y|x) = p(y_1|x_1)$, since the information density converges in probability to the mutual information [3]. Hence, we get Shannon's formula

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

(where the sup is a max, since $I(X;Y)$ is concave in $p(x)$).

References

[1] A. Feinstein, "A new basic theorem of information theory," *IEEE Transactions on Information Theory*, vol. 4, no. 4, pp. 2–22, Sept. 1954.

[2] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Transactions on Information Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.