Chapter 4

Channel Coding Theorems and Approximations of Output Statistics for Arbitrary Channels

Po-Ning Chen

Institute of Communications Engineering

National Chiao-Tung University

Hsin Chu, Taiwan 30010

Motivations

- Shannon's channel capacity [2] is usually derived under the assumption that the channel is memoryless.
- With moderate modification of the proof, this result was extended to stationary-ergodic channels for which the capacity formula becomes the maximization of the mutual information rate:

$$\lim_{n\to\infty} \sup_{X^n} \frac{1}{n} I(X^n; Y^n).$$

- Yet, for more general channels, such as non-stationary or non-ergodic channels, a more general expression for channel capacity needs to be derived.

General models for channels

- The channel transition probability in its most general form is denoted by $\{W^n = P_{Y^n|X^n}\}_{n=1}^{\infty}$, which is abbreviated by $\boldsymbol{W}$ for convenience.
- Similarly, the input and output random processes are respectively denoted by $\boldsymbol{X}$ and $\boldsymbol{Y}$.
- Throughout the text, we denote for convenience

$$P_{X^n,Y^n} = P_{X^n W^n},$$

where Y^n is the output of channel $W^n = P_{Y^n|X^n}$ under input X^n.

- Please refer also to Section 1.3 for the description of general channels.

Notations

- The sup- and inf- (mutual-)information rates are respectively defined by

$$\bar{I}(\boldsymbol{X};\boldsymbol{Y}) := \sup\{\theta \ : \ \underline{i}(\theta) < 1\}$$

and

$$\underline{I}(\boldsymbol{X};\boldsymbol{Y}) := \sup\{\theta \ : \ \bar{i}(\theta) \le 0\},$$

where

$$\underline{i}(\theta) := \liminf_{n\to\infty} \Pr\left\{\frac{1}{n} i_{X^n W^n}(X^n;Y^n) \le \theta\right\}$$

is the inf-spectrum of the normalized information density,

$$\bar{i}(\theta) := \limsup_{n\to\infty} \Pr\left\{\frac{1}{n} i_{X^n W^n}(X^n;Y^n) \le \theta\right\}$$

is the sup-spectrum of the normalized information density, and

$$i_{X^n W^n}(x^n;y^n) := \log \frac{P_{Y^n|X^n}(y^n|x^n)}{P_{Y^n}(y^n)}$$

is the *information density*.

Historical background

- In 1994, Verdú and Han have shown that the channel capacity in its most general form is

$$C := \sup_{\boldsymbol{X}} \underline{I}(\boldsymbol{X};\boldsymbol{Y}).$$

- In their proof, they showed the achievability part via Feinstein's lemma for the channel coding average error probability.
- More importantly, they provided a new converse based on an error lower bound for multihypothesis testing.
- In this chapter, we do not present the original proof of Verdú and Han in the converse theorem. Instead, we will first derive and illustrate in Section 4.3 a general lower bound on the minimum error probability of multihypothesis testing [Chen & Alajaji 2012].
- We then use a special case of the bound, which results the so-called Poor-Verdú bound [Poor & Verdú 1995], to complete the proof of the converse theorem.

Notations and definitions

Definition 4.1 (fixed-length data transmission code) An (n, M) fixed-length data transmission code for channel input alphabet $\mathcal{X}^n$ and output alphabet $\mathcal{Y}^n$ consists of

1. M messages intended for transmission;
2. an encoding function
$$f : \{1, 2, \ldots, M\} \to \mathcal{X}^n;$$
3. a decoding function
$$g : \mathcal{Y}^n \to \{1, 2, \ldots, M\},$$
which is (usually) a deterministic rule that assigns a guess to each possible received vector.

The channel inputs in $\{x^n \in \mathcal{X}^n \;:\; x^n = f(m) \text{ for some } 1 \leq m \leq M\}$ are the codewords of the data transmission code.

Notations and definitions

Definition 4.2 (average probability of error) The average probability of error for a $\mathcal{C}_n = (n, M)$ code with encoder $f(\cdot)$ and decoder $g(\cdot)$ transmitted over channel $W^n = P_{Y^n|X^n}$ is defined as

$$P_e(\mathcal{C}_n) = \frac{1}{M} \sum_{i=1}^{M} \lambda_i,$$

where

$$\lambda_i := \sum_{\left\{y^n \in \mathcal{Y}^n \,:\, g(y^n) \neq i\right\}} P_{Y^n|X^n}(y^n | f(i)).$$

We assume that the message set (of size M) is governed by a uniform distribution. Thus, under the average probability of error criterion, all codewords are treated equally (having a uniform prior distribution).

Definition 4.3 (channel capacity C) The channel capacity C is the supremum of all the rates R for which there exists a sequence of $\mathcal{C}_n = (n, M_n)$ channel block codes such that

$$\liminf_{n \to \infty} \frac{1}{n} \log M_n \geq R,$$

and

$$\limsup_{n \to \infty} P_e(\mathcal{C}_n) = 0.$$

Lemma 4.4 (Feinstein's Lemma) Fix a positive n. For every $\gamma > 0$ and input distribution P_{X^n} on $\mathcal{X}^n$, there exists an (n, M) block code for the transition probability $P_{W^n} = P_{Y^n|X^n}$ that its average error probability $P_e(\mathcal{C}_n)$ satisfies

$$P_e(\mathcal{C}_n) < \Pr\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) < \frac{1}{n}\log M + \gamma\right] + e^{-n\gamma}.$$

Proof:

Step 1: Notations. Define

$$\mathcal{G} := \left\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \ \frac{1}{n} i_{X^n W^n}(x^n; y^n) \geq \frac{1}{n}\log M + \gamma\right\}.$$

Let $\nu := e^{-n\gamma} + P_{X^n W^n}(\mathcal{G}^c)$.

Feinstein's Lemma obviously holds if $\nu \geq 1$, because then

$$P_e(\mathcal{C}_n) \leq 1 \leq \nu := \Pr\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) < \frac{1}{n}\log M + \gamma\right] + e^{-n\gamma}.$$

So we assume $\nu < 1$, which immediately results in

$$P_{X^n W^n}(\mathcal{G}^c) < \nu < 1,$$

or equivalently,

$$P_{X^n W^n}(\mathcal{G}) > 1 - \nu > 0. \quad (4.2.1)$$

Feinstein's Lemma

Therefore, denoting

$$\mathcal{A} := \{x^n \in \mathcal{X}^n \ : \ P_{Y^n|X^n}(\mathcal{G}_{x^n}|x^n) > 1 - \nu\}$$

with $\mathcal{G}_{x^n} := \{y^n \in \mathcal{Y}^n \ : \ (x^n, y^n) \in \mathcal{G}\}$, we have

$$P_{X^n}(\mathcal{A}) > 0,$$

because if $P_{X^n}(\mathcal{A}) = 0$,

$$(\forall \ x^n \text{ with } P_{X^n}(x^n) > 0) \ P_{Y^n|X^n}(\mathcal{G}_{x^n}|x^n) \le 1 - \nu$$

$$\Rightarrow \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) P_{Y^n|X^n}(\mathcal{G}_{x^n}|x^n) = P_{X^n W^n}(\mathcal{G}) \le 1 - \nu,$$

and a contradiction to (4.2.1) is obtained.

Feinstein's Lemma

Step 2: Encoder. Choose an x_1^n in $\mathcal{A}$ (Recall that $P_{X^n}(\mathcal{A}) > 0$.) Define $\Gamma_1 = \mathcal{G}_{x_1^n}$. (Then $P_{Y^n|X^n}(\Gamma_1|x_1^n) > 1 - \nu$.)

Next choose, if possible, a point $x_2^n \in \mathcal{X}^n$ without replacement (i.e., x_2^n cannot be identical to x_1^n) for which

$$P_{Y^n|X^n}\left(\mathcal{G}_{x_2^n} - \Gamma_1 \middle| x_2^n\right) > 1 - \nu,$$

and define $\Gamma_2 := \mathcal{G}_{x_2^n} - \Gamma_1$.

Continue in the following way as for codeword i: choose x_i^n to satisfy

$$P_{Y^n|X^n}\left(\mathcal{G}_{x_i^n} - \bigcup_{j=1}^{i-1} \Gamma_j \middle| x_i^n\right) > 1 - \nu,$$

and define $\Gamma_i := \mathcal{G}_{x_i^n} - \bigcup_{j=1}^{i-1} \Gamma_j$.

Repeat the above codeword selecting procedure until either M codewords are selected or all the points in $\mathcal{A}$ are exhausted.

Feinstein's Lemma

Step 3: Decoder. Define the decoding rule as

$$\phi(y^n) = \begin{cases} i, & \text{if } y^n \in \Gamma_i \\ \text{arbitrary}, & \text{otherwise.} \end{cases}$$

Step 4: Probability of error. For all selected codewords, the error probability given codeword i is transmitted, $\lambda_{e|i}$, satisfies

$$\lambda_{e|i} \le P_{Y^n|X^n}(\Gamma_i^c|x_i^n) < \nu.$$

(Note that $(\forall\ i)\ P_{X^n|X^n}(\Gamma_i|x_i^n) \ge 1 - \nu$ by Step 2.) Therefore, if we can show that the above codeword selecting procedures will not terminate before M, then

$$P_e(\mathcal{C}_n) = \frac{1}{M}\sum_{i=1}^{M}\lambda_{e|i} < \nu.$$

Step 5: Claim. The codeword selecting procedure in Step 2 will not terminate before M.

Proof: We will prove it by contradiction.

Suppose the above procedure terminates before M, say at $N < M$. Define the set

$$\mathcal{F} := \bigcup_{i=1}^{N} \Gamma_i \in \mathcal{Y}^n.$$

Consider the probability

$$P_{X^nW^n}(\mathcal{G}) = P_{X^nW^n}[\mathcal{G} \cap (\mathcal{X}^n \times \mathcal{F})] + P_{X^nW^n}[\mathcal{G} \cap (\mathcal{X}^n \times \mathcal{F}^c)]. \qquad (4.2.2)$$

Since for any $y^n \in \mathcal{G}_{x_i^n}$,

$$P_{Y^n}(y^n) \leq \frac{P_{Y^n|X^n}(y^n|x_i^n)}{M \cdot e^{n\gamma}},$$

we have

$$\begin{aligned} P_{Y^n}(\Gamma_i) &\leq P_{Y^n}(\mathcal{G}_{x_i^n}) \\ &\leq \frac{1}{M} e^{-n\gamma} P_{Y^n|X^n}(\mathcal{G}_{x_i^n}) \\ &\leq \frac{1}{M} e^{-n\gamma}. \end{aligned}$$

Feinstein's Lemma

So the 1st term of the right hand side in (4.2.2) can be upper bounded by

$$\begin{aligned} P_{X^nW^n}[\mathcal{G} \cap (\mathcal{X}^n \times \mathcal{F})] &\leq P_{X^nW^n}(\mathcal{X}^n \times \mathcal{F}) \\ &= P_{Y^n}(\mathcal{F}) = \sum_{i=1}^{N} P_{Y^n}(\Gamma_i) \leq N \times \frac{1}{M} e^{-n\gamma} = \frac{N}{M} e^{-n\gamma}. \end{aligned}$$

As for the 2nd term of the right hand side in (4.2.2), we can upper bound it by

$$\begin{aligned} P_{X^nW^n}[\mathcal{G} \cap (\mathcal{X}^n \times \mathcal{F}^c)] &= \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) P_{Y^n|X^n}(\mathcal{G}_{x^n} \cap \mathcal{F}^c | x^n) \\ &= \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) P_{Y^n|X^n}\left(\mathcal{G}_{x^n} - \bigcup_{i=1}^{N} \Gamma_i \middle| x^n \right) \\ &\leq \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)(1-\nu) \leq 1-\nu, \end{aligned}$$

where the last step follows since for all $x^n \in \mathcal{X}^n$,

$$P_{Y^n|X^n}\left(\mathcal{G}_{x^n} - \bigcup_{i=1}^{N} \Gamma_i \middle| x^n \right) \leq 1-\nu.$$

(Because otherwise we could find the $(N+1)$-th codeword.)

Feinstein's Lemma

Consequently,

$$P_{X^nW^n}(\mathcal{G}) \le \frac{N}{M}e^{-n\gamma} + 1 - \nu.$$

By definition of $\mathcal{G}$,

$$P_{X^nW^n}(\mathcal{G}) = 1 - \nu + e^{-n\gamma} \le \frac{N}{M}e^{-n\gamma} + 1 - \nu,$$

which implies $N \ge M$, resulting in a contradiction. □

Error bounds for multihypothesis testing

We next introduce the generalized Poor-Verdú bound parameterized by $\theta \geq 1$. Note that when $\theta = 1$, this bound reduces to the original Poor-Verdú bound in [Poor & Verdú 1995].

Lemma 4.5 (generalized Poor-Verdú bound [Chen & Alajaji 2012]**)** Suppose X and Y are random variables, where X takes values on a discrete (i.e., finite or coutably infinite) alphabet $\mathcal{X} = \{x_1, x_2, x_3, \ldots\}$ and Y takes on values in an arbitrary alphabet $\mathcal{Y}$. The minimum probability of error P_e in estimating X from Y satisfies

$$P_e \geq (1-\alpha) \cdot P_{X,Y} \left\{ (x,y) \in \mathcal{X} \times \mathcal{Y} : \ P^{(\theta)}_{X|Y}(x|y) \leq \alpha \right\} \tag{4.3.1}$$

for each $\alpha \in [0,1]$ and $\theta \geq 1$, where for each $y \in \mathcal{Y}$,

$$P^{(\theta)}_{X|Y}(x|y) := \frac{(P_{X|Y}(x|y))^\theta}{\sum_{x' \in \mathcal{X}} (P_{X|Y}(x'|y))^\theta}, \quad x \in \mathcal{X}, \tag{4.3.2}$$

is the tilted distribution of $P_{X|Y}(\cdot|y)$ with parameter θ.

Error bounds for multihypothesis testing

Proof: Fix $\theta \geq 1$. We only provide the proof for $0 < \alpha < 1$ since the lower bound trivially holds when $\alpha = 0$ and $\alpha = 1$.

- It is known that the estimate $e(Y)$ of X from observing Y that minimizes the error probability is the maximum *a posteriori* (MAP) estimate given by

$$e(Y) = \arg\max_{x \in \mathcal{X}} P_{X|Y}(x|Y). \tag{4.3.3}$$

Therefore, the error probability incurred in testing among the values of X is given by

$$\begin{aligned} 1 - P_e &= \Pr\{X = e(Y)\} \\ &= \int_{\mathcal{Y}} \left[\sum_{\{x \,:\, x = e(y)\}} P_{X|Y}(x|y) \right] dP_Y(y) \\ &= \int_{\mathcal{Y}} \left(\max_{x \in \mathcal{X}} P_{X|Y}(x|y) \right) dP_Y(y) \\ &= \int_{\mathcal{Y}} \left(\max_{x \in \mathcal{X}} f_x(y) \right) dP_Y(y) = E\left[\max_{x \in \mathcal{X}} f_x(Y) \right], \end{aligned}$$

where $f_x(y) := P_{X|Y}(x|y)$.

Error bounds for multihypothesis testing

- For a fixed $y \in \mathcal{Y}$, let $h_j(y)$ be the j-th element in the set

$$\{f_{x_1}(y), f_{x_2}(y), f_{x_3}(y), \ldots\}$$

such that its elements are listed in non-increasing order; i.e.,

$$h_1(y) \geq h_2(y) \geq h_3(y) \geq \cdots$$

and $\{h_1(y), h_2(y), h_3(y), \ldots\} = \{f_{x_1}(y), f_{x_2}(y), f_{x_3}(y), \ldots\}$. Then

$$1 - P_e = E[h_1(Y)]. \tag{4.3.4}$$

- For each $h_j(y)$ above, define $h_j^{(\theta)}(y)$ such that $h_j^{(\theta)}(y)$ is the respective element for $h_j(y)$, satisfying

$$h_j(y) = f_{x_j}(y) = P_{X|Y}(x_j|y) \ \Leftrightarrow\ h_j^{(\theta)}(y) = P_{X|Y}^{(\theta)}(x_j|y).$$

Since $h_1(y)$ is the largest among $\{h_j(y)\}_{j \geq 1}$, we note that

$$h_1^{(\theta)}(y) = \frac{h_1^{\theta}(y)}{\sum_{j\geq 1} h_j^{\theta}(y)} = \frac{1}{1 + \sum_{j\geq 2}[h_j(y)/h_1(y)]^{\theta}}$$

is non-decreasing in θ for each y; this implies that

$$h_1^{(\theta)}(y) \geq h_1(y) \quad \text{for } \theta \geq 1 \text{ and } y \in \mathcal{Y}. \tag{4.3.5}$$

Error bounds for multihypothesis testing

- For any $\alpha \in (0,1)$, we can write

$$
\begin{aligned}
& P_{X,Y}\left\{(x,y) \in \mathcal{X} \times \mathcal{Y} :\ P^{(\theta)}_{X|Y}(x|y) > \alpha\right\} \\
&= \int_{\mathcal{Y}} P_{X|Y}\left\{x \in \mathcal{X} :\ P^{(\theta)}_{X|Y}(x|y) > \alpha\right\} dP_Y(y) \\
&= \int_{\mathcal{Y}} \left(\sum_{j=1}^{\infty} h_j(y) \cdot \mathbf{1}\left(h_j^{(\theta)}(y) > \alpha\right)\right) dP_Y(y) \\
&\geq \int_{\mathcal{Y}} h_1(y) \cdot \mathbf{1}\left(h_1^{(\theta)}(y) > \alpha\right) dP_Y(y) \\
&\geq \int_{\mathcal{Y}} h_1(y) \cdot \mathbf{1}(h_1(y) > \alpha) dP_Y(y) \\
&= E[h_1(Y) \cdot \mathbf{1}(h_1(Y) > \alpha)], \qquad (4.3.6)
\end{aligned}
$$

where $\mathbf{1}(\cdot)$ is the indicator function and the second inequality follows from (4.3.5).

Error bounds for multihypothesis testing

- To complete the proof, we next relate $E[h_1(Y)\cdot\mathbf{1}(h_1(Y)>\alpha)]$ with $E[h_1(Y)]$, which is exactly $1-P_e$.

  For any $\alpha\in(0,1)$ and any random variable U with $\Pr\{0\le U\le 1\}=1$, the following inequality holds with probability one:

  $$U\le\alpha+(1-\alpha)\cdot U\cdot\mathbf{1}(U>\alpha).$$

  This can be easily proved by upper-bounding U in terms of α when $0\le U\le\alpha$, and $\alpha+(1-\alpha)U$, otherwise. Thus

  $$E[U]\le\alpha+(1-\alpha)E[U\cdot\mathbf{1}(U>\alpha)].$$

- Applying the above inequality to (4.3.6) by setting $U=h_1(Y)$, we obtain

  $$\begin{aligned}(1-\alpha)P_{X,Y}&\left\{(x,y)\in\mathcal{X}\times\mathcal{Y}\ :\ P^{(\theta)}_{X|Y}(x|y)>\alpha\right\}\\ &\ge\ E[h_1(Y)]-\alpha\\ &=\ (1-P_e)-\alpha\\ &=\ (1-\alpha)-P_e,\end{aligned}$$

  where the first equality follows from (4.3.4). This completes the proof. □

Error bounds for multihypothesis testing

- There are examples demonstrating that the generalized Poor-Verdú bound is tight when $\theta \to \infty$ (See the lecture note).
- For the verification of the general Shannon capacity, however, taking $\theta = 1$ is adequate.

Corollary 4.9 Every $\mathscr{C}_n = (n, M)$ code satisfies

$$P_e(\mathscr{C}_n) \geq \left(1 - e^{-n\gamma}\right) \Pr\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \frac{1}{n} \log M - \gamma\right]$$

for every $\gamma > 0$, where X^n places probability mass $1/M$ on each codeword, and $P_e(\mathscr{C}_n)$ denotes the error probability of the code.

Error bounds for multihypothesis testing

Proof: Taking $\alpha = e^{-n\gamma}$ and $\theta = 1$ in Lemma 4.5, and replacing X and Y in Lemma 4.5 by its n-fold counterparts, i.e., X^n and Y^n, we obtain

$$\begin{aligned}
P_e(\mathcal{C}_n) &\geq \left(1 - e^{-n\gamma}\right) P_{X^nW^n}\left[(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \ P_{X^n|Y^n}(x^n|y^n) \leq e^{-n\gamma}\right] \\
&= \left(1 - e^{-n\gamma}\right) P_{X^nW^n}\left[(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \ \frac{P_{X^n|Y^n}(x^n|y^n)}{1/M} \leq \frac{e^{-n\gamma}}{1/M}\right] \\
&= \left(1 - e^{-n\gamma}\right) P_{X^nW^n}\left[(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \ \frac{P_{X^n|Y^n}(x^n|y^n)}{P_{X^n}(x^n)} \leq \frac{e^{-n\gamma}}{1/M}\right] \\
&= \left(1 - e^{-n\gamma}\right) P_{X^nW^n}\Big[(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \ : \\
&\qquad\qquad \frac{1}{n}\log\frac{P_{X^n|Y^n}(x^n|y^n)}{P_{X^n}(x^n)} \leq \frac{1}{n}\log M - \gamma\Big] \\
&= \left(1 - e^{-n\gamma}\right) \Pr\left[\frac{1}{n} i_{X^nW^n}(X^n; Y^n) \leq \frac{1}{n}\log M - \gamma\right].
\end{aligned}$$

□

Capacity formulas for general channels

Definition 4.10 (ε-achievable rate) Fix $\varepsilon \in [0, 1]$. $R \geq 0$ is an ε-achievable rate if there exists a sequence of $\mathscr{C}_n = (n, M_n)$ channel block codes such that

$$\liminf_{n\to\infty} \frac{1}{n} \log M_n \geq R$$

and

$$\limsup_{n\to\infty} P_e(\mathscr{C}_n) \leq \varepsilon.$$

Definition 4.11 (ε-capacity C_ε) Fix $\varepsilon \in [0, 1]$. The supremum of ε-achievable rates is called the ε-capacity, C_ε.

- It is straightforward for the definition that C_ε is non-decreasing in ε, and $C_1 = \log |\mathcal{X}|$.

Capacity formulas for general channels

Observation 4.12 (capacity C) Note that channel capacity C is equal to the supremum of the rates that are ε-achievable for all $\varepsilon \in [0, 1]$:

$$C = \inf_{0 \le \varepsilon \le 1} C_\varepsilon = \lim_{\varepsilon \downarrow 0} C_\varepsilon = C_0.$$

Definition 4.13 (strong capacity C_{SC}) Define the *strong converse capacity* (or *strong capacity*) C_{SC} as the infimum of the rates R such that for all $\mathcal{C}_n = (n, M_n)$ channel block codes with

$$\liminf_{n \to \infty} \frac{1}{n} \log M_n \ge R,$$

we have

$$\liminf_{n \to \infty} P_e(\mathcal{C}_n) = 1.$$

ε-capacity

Theorem 4.14 (ε-capacity) For $0 < \varepsilon < 1$, the ε-capacity C_ε for arbitrary channels satisfies

$$C_\varepsilon = \sup_{\boldsymbol{X}} \underline{I}_\varepsilon(\boldsymbol{X}; \boldsymbol{Y}).$$

Proof:

1. $C_\varepsilon \geq \sup_{\boldsymbol{X}} \underline{I}_\varepsilon(\boldsymbol{X}; \boldsymbol{Y})$.

   Fix input $\boldsymbol{X}$. It suffices to show the existence of $\mathscr{C}_n = (n, M_n)$ data transmission code with rate

   $$\underline{I}_\varepsilon(\boldsymbol{X}; \boldsymbol{Y}) - \gamma < \frac{1}{n} \log M_n < \underline{I}_\varepsilon(\boldsymbol{X}; \boldsymbol{Y}) - \frac{\gamma}{2}$$

   and probability of decoding error satisfying

   $$\limsup_{n \to \infty} P_e(\mathscr{C}_n) \leq \varepsilon$$

   for every $\gamma > 0$. (Because if such code exists, then $\liminf_{n\to\infty}(1/n) \log M_n \geq \underline{I}_\varepsilon(\boldsymbol{X}; \boldsymbol{Y}) - \gamma$, which implies $C_\varepsilon \geq \underline{I}_\varepsilon(\boldsymbol{X}; \boldsymbol{Y}) - \gamma$ for arbitrarily small γ.)

From Lemma 4.4, there exists an $\mathscr{C}_n = (n, M_n)$ code whose error probability satisfies

$$\begin{aligned} P_e(\mathscr{C}_n) &< \Pr\left[\frac{1}{n} i_{X^n W^n}(X^n;Y^n) < \frac{1}{n}\log M_n + \frac{\gamma}{4}\right] + e^{-n\gamma/4} \\ &\le \Pr\left[\frac{1}{n} i_{X^n W^n}(X^n;Y^n) < \left(\underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y}) - \frac{\gamma}{2}\right) + \frac{\gamma}{4}\right] + e^{-n\gamma/4} \\ &\le \Pr\left[\frac{1}{n} i_{X^n W^n}(X^n;Y^n) < \underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y}) - \frac{\gamma}{4}\right] + e^{-n\gamma/4}. \end{aligned}$$

Since

$$\underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y}) := \sup\left\{R : \limsup_{n\to\infty} \Pr\left[\frac{1}{n} i_{W^n W^n}(X^n;Y^n) \le R\right] \le \varepsilon\right\},$$

we obtain

$$\limsup_{n\to\infty} \Pr\left[\frac{1}{n} i_{X^n W^n}(X^n;Y^n) < \underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y}) - \frac{\gamma}{4}\right] \le \varepsilon.$$

Hence, the proof of the direct part is completed by noting that

$$\begin{aligned} \limsup_{n\to\infty} P_e(\mathscr{C}_n) &\le \limsup_{n\to\infty} \Pr\left[\frac{1}{n} i_{X^n W^n}(X^n;Y^n) < \underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y}) - \frac{\gamma}{4}\right] \\ &\quad + \limsup_{n\to\infty} e^{-n\gamma/4} = \varepsilon. \end{aligned}$$

2. $C_\varepsilon \le \sup_{\boldsymbol{X}} \underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y})$.

- Suppose that there exists a sequence of $\mathscr{C}_n = (n, M_n)$ codes with rate strictly larger than $\sup_{\boldsymbol{X}} \underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y})$ and $\limsup_{n\to\infty} P_e(\mathscr{C}_n) \le \varepsilon$. Let the ultimate code rate for this code be $\sup_{\boldsymbol{X}} \underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y}) + 3\rho$ for some $\rho > 0$. Then for sufficiently large n,

$$\frac{1}{n}\log M_n > \sup_{\boldsymbol{X}} \underline{I}_\varepsilon(\boldsymbol{X};\boldsymbol{Y}) + 2\rho.$$

- Since the above inequality holds for every $\boldsymbol{X}$, it certainly holds if taking input $\hat{X}^n$ which places probability mass $1/M_n$ on each codeword, i.e.,

$$\frac{1}{n}\log M_n > \underline{I}_\varepsilon(\hat{\boldsymbol{X}};\hat{\boldsymbol{Y}}) + 2\rho, \tag{4.4.1}$$

where $\hat{\boldsymbol{Y}}$ is the channel output due to channel input $\hat{\boldsymbol{X}}$.

ε-capacity

- Then from Corollary 4.9, the error probability of the code satisfies

$$\begin{aligned} P_e(\mathcal{C}_n) &\geq \left(1 - e^{-n\rho}\right) Pr\left[\frac{1}{n} i_{\hat{X}^n W^n}(\hat{X}^n; \hat{Y}^n) \leq \frac{1}{n}\log M_n - \rho\right] \\ &\geq \left(1 - e^{-n\rho}\right) Pr\left[\frac{1}{n} i_{\hat{X}^n W^n}(\hat{X}^n; \hat{Y}^n) \leq \underline{I}_\varepsilon(\hat{\boldsymbol{X}}; \hat{\boldsymbol{Y}}) + \rho\right], \end{aligned}$$

  where the last inequality follows from (4.4.1), which by taking the limsup of both sides, we have

$$\varepsilon \geq \limsup_{n\to\infty} P_e(\mathcal{C}_n) \geq \limsup_{n\to\infty} Pr\left[\frac{1}{n} i_{\hat{X}^n W^n}(\hat{X}^n; \hat{Y}^n) \leq \underline{I}_\varepsilon(\hat{\boldsymbol{X}}; \hat{\boldsymbol{Y}}) + \rho\right] > \varepsilon,$$

  and a desired contradiction is obtained. □

General Shannon capacity and strong capacity

Theorem 4.15 (general channel capacity) The channel capacity C for arbitrary channel satisfies

$$C = \sup_{\boldsymbol{X}} \underline{I}(\boldsymbol{X};\boldsymbol{Y}).$$

Theorem 4.16 (general strong capacity)

$$C_{SC} := \sup_{\boldsymbol{X}} \bar{I}(\boldsymbol{X};\boldsymbol{Y}).$$

- Note that in the general formula for strong capacity, **sup-information rate** is used as contrary to the **inf-information rate** formula for Shannon capacity.

Examples

Example 4.17 (capacity) Let the input and output alphabets be $\{0, 1\}$, and let every output Y_i be given by:

$$Y_i = X_i \oplus N_i.$$

Assume the input process $\boldsymbol{X}$ and the noise process $\boldsymbol{N}$ are independent.

Then

$$\underline{H}(\boldsymbol{Y}) - \bar{H}(\boldsymbol{Y}|\boldsymbol{X}) \leq \underline{I}(\boldsymbol{X};\boldsymbol{Y}) \leq \bar{H}(\boldsymbol{Y}) - \bar{H}(\boldsymbol{Y}|\boldsymbol{X})$$

or equivalently,

$$\underline{H}(\boldsymbol{Y}) - \bar{H}(\boldsymbol{N}) \leq \underline{I}(\boldsymbol{X};\boldsymbol{Y}) \leq \bar{H}(\boldsymbol{Y}) - \bar{H}(\boldsymbol{N}).$$

By the channel symmetry, we obtain:

$$C = \log(2) - \bar{H}(\boldsymbol{N}) \text{ nats.}$$

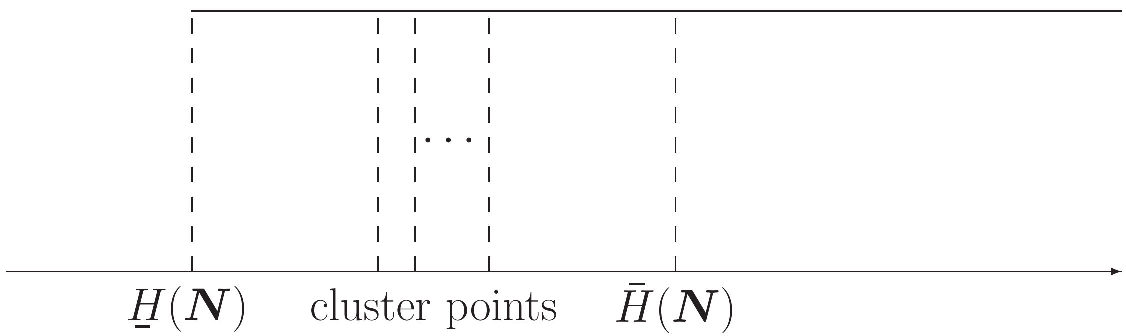
Examples

Case A) If $\boldsymbol{N}$ is a non-stationary binary independent sequence with

$$\Pr\{N_i = 1\} = p_i,$$

then

$$C = \log(2) - \limsup_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h_b(p_i) \quad \text{nats/channel usage.}$$



The ultimate CDFs of $-(1/n)\log P_{N^n}(N^n)$.

Case B) If $\boldsymbol{N}$ has the same distribution as the source process in Example 4.23, then $\bar{H}(\boldsymbol{N}) = \log(2)$ nats, which yields a zero channel capacity.

Examples

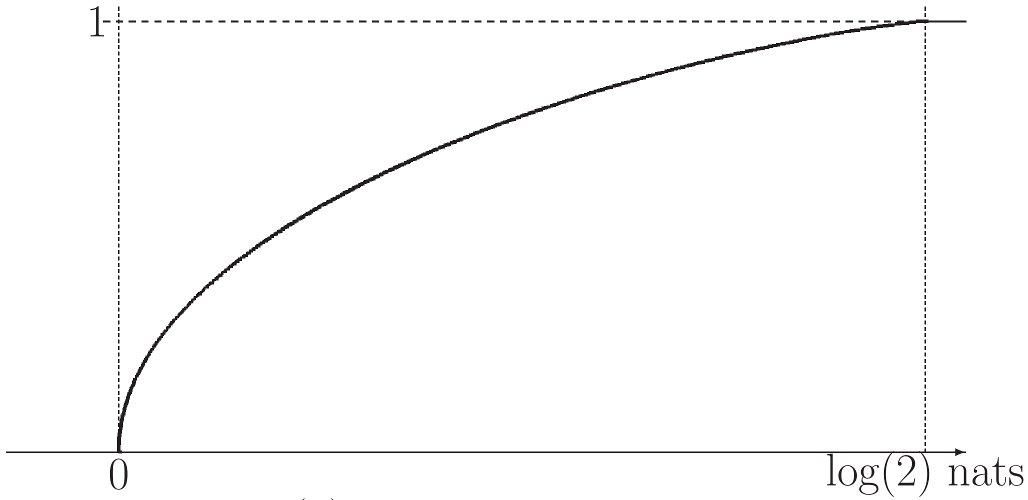
Example 4.18 (strong capacity)

Case A)

$$C_{SC} = 1 - \liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h_b(p_i).$$

Case B)

$$C_{SC} = \log(2) \text{ nats/channel usage.}$$



The ultimate CDF $i(\theta)$ of the normalized information density for Example 4.18-Case B).

Examples

Example 4.19 (ε-capacity) Consider the channel in Case B of Example 4.17.

$$C_\varepsilon = i^{-1}(\theta).$$

Capacity and resolvability for channels

- The channel capacity for discrete memoryless channel is shown to be:

$$C := \max_{P_X} I(P_X, Q_{Y|X}).$$

- Let $P_{\overline{X}}$ be the optimizer of the above maximization operation. Then

$$C := \max_{P_X} I(P_X, Q_{Y|X}) = I(P_{\overline{X}}, Q_{Y|X}).$$

- Here, the performance of the code is assumed to be the average error probability, namely

$$P_e(\mathcal{C}_n) = \frac{1}{M} \sum_{i=1}^{M} P_e(\mathcal{C}_n | x_i^n),$$

  if the code book is $\mathcal{C}_n := \{x_1^n, x_2^n, \ldots, x_M^n\}$.

- Due to the random coding argument, a deterministic good code with arbitrarily small error probability and rate less than channel capacity must exist.

- One can ask: What is the relationship between a good code and the optimizer $P_{\overline{X}}$? It is widely believed that if the code is good (with rate close to capacity and low error probability), then the output statistics $P_{\widetilde{Y}^n}$ – due to the equally-likely code – must approximate the output distribution, denoted by $P_{\overline{Y}^n}$, due to the input distribution achieving the channel capacity.

Theorem 4.20 (Han & Verdú 1993) For any channel $W^n = (Y^n|X^n)$ with finite input alphabet and capacity C that satisfies the strong converse (i.e., $C = C_{SC}$), the following statement holds.

Fix any $\gamma > 0$ and any sequence of $\{\mathcal{C}_n = (n, M_n)\}_{n=1}^{\infty}$ block codes with

$$\frac{1}{n}\log M_n \geq C - \gamma/2,$$

and vanishing error probability. Then

$$\frac{1}{n}\|\widetilde{Y}^n - \overline{Y}^n\| \leq \gamma \quad \text{for all sufficiently large } n,$$

where $\widetilde{Y}^n$ is the output due to the block code and $\overline{Y}^n$ is the output due the $\overline{X}^n$ that satisfies

$$I(\overline{X}^n; \overline{Y}^n) = \max_{X^n} I(X^n; Y^n).$$

To be specific,

$$P_{\widetilde{Y}^n}(y^n) = \sum_{x^n \in \mathcal{C}_n} P_{\widetilde{X}^n}(x^n) P_{W^n}(y^n|x^n) = \sum_{x^n \in \mathcal{C}_n} \frac{1}{M} P_{W^n}(y^n|x^n)$$

and

$$P_{\overline{Y}^n}(y^n) = \sum_{x^n \in \mathcal{X}^n} P_{\overline{X}^n}(x^n) P_{W^n}(y^n|x^n).$$

Capacity and resolvability for channels

- Note that the above theorem holds for arbitrary channels, not restricted to only discrete memoryless channels.
- One can wonder whether a result in the spirit of the above theorem can be proved for the input statistics rather than the output statistics.
- The answer is negative.
- Hence, the statement that *the statistics of any good code must approximate those that maximize the mutual information* is erroneously taken for granted.
  - However, we do not rule out the possibility of the existence of good codes that approximate those that maximize the mutual information.

Capacity and resolvability for channels

- To see this, simply consider the normalized entropy of $\overline{X}^n$ versus that of $\widetilde{X}^n$ (which is uniformly distributed over the codewords) for discrete memoryless channels:

$$
\begin{aligned}
\frac{1}{n}H(\overline{X}^n) - \frac{1}{n}H(\widetilde{X}^n) &= \left[\frac{1}{n}H(\overline{X}^n|\overline{Y}^n) + \frac{1}{n}I(\overline{X}^n;\overline{Y}^n)\right] - \frac{1}{n}\log(M_n) \\
&= \left[H(\overline{X}|\overline{Y}) + I(\overline{X};\overline{Y})\right] - \frac{1}{n}\log(M_n) \\
&= \left[H(\overline{X}|\overline{Y}) + C\right] - \frac{1}{n}\log(M_n).
\end{aligned}
$$

A good code with vanishing error probability exists for $(1/n)\log(M_n)$ arbitrarily close to C; hence, we can find a good code sequence to satisfy

$$
\lim_{n\to\infty}\left[\frac{1}{n}H(\overline{X}^n) - \frac{1}{n}H(\widetilde{X}^n)\right] = H(\overline{X}|\overline{Y}).
$$

Since the term $H(\overline{X}|\overline{Y})$ is in general positive, where a quick example is the BSC with crossover probability p, which yields

$$
\begin{aligned}
H(\overline{X}|\overline{Y}) &= H(\overline{X}) - I(\overline{X};\overline{Y}) \\
&= H(\overline{X}) - H(\overline{Y}) + H(\overline{Y}|\overline{X}) \\
&= H(\overline{Y}|\overline{X}) = -p\log(p) - (1-p)\log(1-p),
\end{aligned}
$$

the two input distributions can by no means resemble to each other.

Resolvability for channels

- The previous discussion motivates the necessity to find an equally-distributed (over a subset of input alphabet) input distribution that generates the output statistics, which is close to the output due to the input that maximizes the mutual information.

- Since such approximations are usually performed by computers, it may be natural to connect approximations of the input and output statistics with the concept of *resolvability*.

Resolvability for channels

- In a data transmission system as shown in Figure 4.3, suppose that the source, channel and output are respectively denoted by

$$X^n := (X_1, \ldots, X_n), \quad W^n := (W_1, \ldots, W_n), \quad \text{and} \quad Y^n := (Y_1, \ldots, Y_n),$$

where W_i has distribution $P_{Y_i|X_i}$.

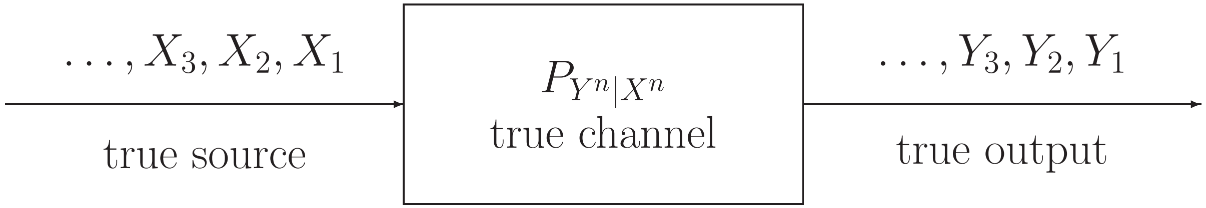


Figure 4.3: The communication system.

- To simulate the behavior of the channel, a computer-generated input may be necessary as shown in Figure 4.4.

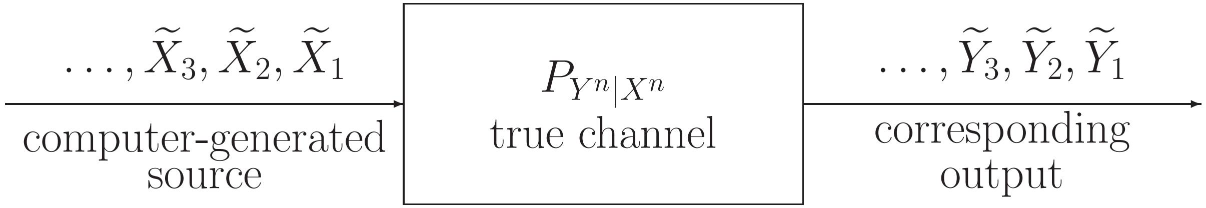


Figure 4.4: The simulated communication system.

Resolvability for channels

- As stated in Chapter 3, such computer-generated input is based on an algorithm formed by a few basic uniform random experiments, which has finite resolution.
- Our goal is to find a good computer-generated input $\widetilde{X}^n$ such that the corresponding output $\widetilde{Y}^n$ is very close to the true output Y^n.

Definition 4.21 (ε-resolvability for input $\boldsymbol{X}$ and channel $\boldsymbol{W}$) Fix $\varepsilon > 0$, and suppose that the (true) input random variable and (true) channel statistics are $\boldsymbol{X}$ and $\boldsymbol{W} = (\boldsymbol{Y}|\boldsymbol{X})$, respectively.

Then the ε-*resolvability* $S_\varepsilon(\boldsymbol{X}, \boldsymbol{W})$ for input $\boldsymbol{X}$ and channel $\boldsymbol{W}$ is defined by:

$$S_\varepsilon(\boldsymbol{X}, \boldsymbol{W}) := \min \Big\{ R \; : \; (\forall\, \gamma > 0)(\exists\, \tilde{\boldsymbol{X}} \text{ and } N)(\forall\, n > N)$$

$$\frac{1}{n} R(\widetilde{X}^n) < R + \gamma \text{ and } \|Y^n - \widetilde{Y}^n\| < \varepsilon \Big\},$$

where $P_{\widetilde{Y}^n} = P_{\widetilde{X}^n} P_{W^n}$.

- Note that if we take the channel W^n to be an identity channel for all n, namely $\mathcal{X}^n = \mathcal{Y}^n$ and $P_{Y^n|X^n}(y^n|x^n)$ is either 1 or 0, then the ε-resolvability for input $\boldsymbol{X}$ and channel $\boldsymbol{W}$ is reduced to source ε-resolvability for source $\boldsymbol{X}$ only:

$$S_\varepsilon(\boldsymbol{X}, \boldsymbol{W}_{\text{Identity}}) = S_\varepsilon(\boldsymbol{X}).$$

Similar reductions can be applied to all the following definitions.

Resolvability for channels

Definition 4.22 (ε-mean-resolvability for input $\boldsymbol{X}$ and channel $\boldsymbol{W}$) Fix $\varepsilon > 0$, and suppose that the (true) input random variable and (true) channel statistics are respectively $\boldsymbol{X}$ and $\boldsymbol{W}$.

Then the ε-*mean-resolvability* $\bar{S}_\varepsilon(\boldsymbol{X}, \boldsymbol{W})$ for input $\boldsymbol{X}$ and channel $\boldsymbol{W}$ is defined by:

$$\bar{S}_\varepsilon(\boldsymbol{X}, \boldsymbol{W}) := \min\Big\{R \;:\; (\forall\, \gamma > 0)(\exists\, \tilde{\boldsymbol{X}} \text{ and } N)(\forall\, n > N)$$

$$\frac{1}{n}H(\widetilde{X}^n) < R + \gamma \text{ and } \|Y^n, \widetilde{Y}^n\|_1 < \varepsilon\Big\},$$

where $P_{\widetilde{Y}^n} = P_{\widetilde{X}^n}P_{W^n}$ and $P_{Y^n} = P_{X^n}P_{W^n}$.

Definition 4.23 (resolvability and mean resolvability for input $\boldsymbol{X}$ and channel $\boldsymbol{W}$) The *resolvability* and *mean-resolvability* for input $\boldsymbol{X}$ and $\boldsymbol{W}$ are defined respectively as:

$$S(\boldsymbol{X}, \boldsymbol{W}) := \sup_{\varepsilon > 0} S_\varepsilon(\boldsymbol{X}, \boldsymbol{W}) \quad \text{and} \quad \bar{S}(\boldsymbol{X}, \boldsymbol{W}) := \sup_{\varepsilon > 0} \bar{S}_\varepsilon(\boldsymbol{X}, \boldsymbol{W}).$$

Resolvability for channels

Definition 4.24 (resolvability and mean resolvability for channel $\boldsymbol{W}$) The *resolvability* and *mean-resolvability* for channel $\boldsymbol{W}$ are defined respectively as:

$$S(\boldsymbol{W}) := \sup_{\boldsymbol{X}} S(\boldsymbol{X}, \boldsymbol{W}), \quad \text{and} \quad \bar{S}(\boldsymbol{W}) := \sup_{\boldsymbol{X}} \bar{S}(\boldsymbol{X}, \boldsymbol{W}).$$

Theorem 4.25 (Han & Verdú 1993)

$$S(\boldsymbol{W}) = C_{SC} = \sup_{\boldsymbol{X}} \bar{I}(\boldsymbol{X}; \boldsymbol{Y}) \quad \text{and} \quad \bar{S}(\boldsymbol{W}) = C = \sup_{\boldsymbol{X}} \underline{I}(\boldsymbol{X}; \boldsymbol{Y}).$$

- It is somewhat a reasonable inference that if no computer algorithms can produce a desired good output statistics under the number of random nats specified, then all codes should be bad codes for this rate.