Strong Converse, Feedback Channel Capacity and Hypothesis Testing*

Po-Ning Chen
Computer & Communication Research Laboratories
Industrial Technology Research Institute
Taiwan 310, Republic of China

Fady Alajaji
Department of Mathematics & Statistics
Queen's University
Kingston, ON K7L 3N6, Canada

Journal of the Chinese Institute of Engineers, to appear November 1995

Abstract

In light of recent results by Verdú and Han on channel capacity, we examine three problems: the strong converse condition to the channel coding theorem, the capacity of arbitrary channels with feedback and the Neyman-Pearson hypothesis testing type-II error exponent. It is first remarked that the strong converse condition holds if and only if the sequence of normalized channel information densities converges in probability to a constant. Examples illustrating this condition are also provided. A general formula for the capacity of arbitrary channels with output feedback is then obtained. Finally, a general expression for the Neyman-Pearson type-II error exponent based on arbitrary observations subject to a constant bound on the type-I error probability is derived.

Key Words: Strong converse, channel capacity, channels with feedback, hypothesis testing.

* Parts of this paper were presented at the 1995 Conference on Information Sciences and Systems, The John Hopkins University, Baltimore, MD, USA, March 1995.

Introduction

In this paper, we investigate three problems inspired by the recent work of Verdú and Han on the general capacity formula of *arbitrary* single-user channels [6]. We first address the strong converse condition obtained in [6] and provide examples of channels for which the strong converse holds. We next derive a general capacity formula for arbitrary single-user channels with output feedback. Finally, we analyze the Neyman-Pearson hypothesis testing problem based on arbitrary observations.

In [6], Verdú and Han give a necessary and sufficient condition for the validity of the strong converse to the channel coding theorem. They state that the strong converse holds if and only if the channel capacity is equal to the channel resolvability. We remark that if there exists an input distribution $P^*_{X^n}$ achieving the channel capacity, then the strong converse condition is actually equivalent to the convergence in probability to a constant (or in distribution to a degenerate random variable) of the sequence of normalized information densities according to a joint input-output distribution with $P^*_{X^n}$ as its induced marginal. We furthermore note that the expression of the strong capacity, which will be defined later, is given by the channel resolvability. We also obtain examples of discrete channels satisfying the strong converse condition.

The main tool used in [6] to derive a general expression for the (nonfeedback) channel capacity is a new approach to the (weak) converse of the coding theorem based on a simple lower bound on error probability. We utilize this result to generalize the capacity expression for channels with feedback. Feedback capacity is shown to equal the supremum, over all feedback encoding strategies, of the input-output *inf-information rate* which is defined as the liminf in probability of the normalized information density.

We finally consider the Neyman-Pearson hypothesis testing problem based on arbitrary observations. We derive a general expression for the type-II error exponent subject to a fixed bound on the type-I error probability. We observe that this expression is indeed the dual of the general ε-capacity formula given in [6].

On the strong converse of the single-user channel

1. Strong converse condition

Consider an arbitrary single-user channel with input alphabet $\mathcal{A}$ and output alphabet $\mathcal{B}$ and n-dimensional transition distribution given by $W^{(n)} = P_{Y^n|X^n} : \mathcal{A}^n \to \mathcal{B}^n$; $n = 1, 2, \ldots$.

Definition 1 ([6]) *An (n, M, ϵ) code has blocklength n, M codewords, and (average) error probability not larger than ϵ. $R \geq 0$ is an ϵ-achievable rate if for every $\gamma > 0$ there exists, for all sufficiently large n, (M, n, ϵ) codes with rate*

$$\frac{log_2\ M}{n} > R - \gamma.$$

The maximum ϵ-achievable rate is called the ϵ-capacity, C_ϵ. The channel capacity, C, is defined as the maximal rate that is ϵ-achievable for all $0 < \epsilon < 1$. It follows immediately from the definition that $C = \lim_{\epsilon \to 0} C_\epsilon$.

Definition 2 ([6]) *A channel with capacity C is said to satisfy the strong converse if for every $\delta > 0$ and every sequence of (n, M, λ_n) codes with rate*

$$\frac{\log_2\ M}{n} > C + \delta,$$

it holds that $\lambda_n \to 1$ as $n \to \infty$.

In [6], Verdú and Han derive a general formula for the operational capacity of *arbitrary* single-user channels (not necessarily stationary, ergodic, information stable, etc.). The (nonfeedback) capacity was shown to equal the supremum, over all input processes, of the input-output *inf-information rate* defined as the liminf in probability of the normalized information density:

$$C = \sup_{X^n} \underline{I}\,(X^n; Y^n), \tag{1}$$

where $X^n = (X_1, \ldots, X_n)$, for $n = 1, 2, \ldots$, is the block input vector and $Y^n = (Y_1, \ldots, Y_n)$ is the corresponding block output vector induced by X^n via the channel.

The symbol $\underline{I}\ (X^n;Y^n)$ appearing in (1) is the *inf-information rate* between X^n and Y^n and is defined as the *liminf in probability* of the sequence of normalized information densities $\frac{1}{n}\, i_{X^nY^n}(X^n;Y^n)$, where

$$i_{X^nY^n}(a^n;b^n) = \log_2 \frac{P_{Y^n|X^n}(b^n|a^n)}{P_{Y^n}(b^n)}. \tag{2}$$

Likewise, the *sup-information rate* denoted as $\bar{I}(X^n;Y^n)$ is defined as the *limsup in probability* of the sequence of normalized information densities.

The *liminf in probability* of a sequence [6] of random variables is defined as follows: If A_n is a sequence of random variables, its *liminf in probability* is the largest extended real number α such that for all $\xi > 0$, $\limsup_{n\to\infty} Pr[A_n \leq \alpha - \xi] = 0$. Similarly, its *limsup in probability* is the smallest extended real numbers β such that for all $\xi > 0$, $\limsup_{n\to\infty} Pr[A_n \geq \beta + \xi] = 0$. Note that these two quantities are always defined; if they are equal, then the sequence of random variables converges in probability to a constant (which is α).

In Theorem 6 in [6], Verdú and Han establish general expressions for ϵ-capacity. They also give a necessary and sufficient condition for the validity of the strong converse (Theorem 7 in [6]), which states that the strong converse condition is equivalent to the condition

$$\sup_{X^n} \underline{I}(X^n;Y^n) = \sup_{X^n} \bar{I}(X^n;Y^n), \tag{3}$$

i.e. $C = S$, where $S \stackrel{\Delta}{=} \sup_{X^n} \bar{I}(X^n;Y^n)$ denotes the channel *resolvability*, which is defined as the minimum number of random bits required per channel use in order to generate an input that achieves arbitrarily accurate approximation of the output statistics for *any* given input process [4]. Furthermore, if channel input alphabet is *finite*, then

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

Lemma 1 *If (3) holds and there exists $\check{X}^n$ such that*

$$\sup_{X^n} \underline{I}(X^n;Y^n) = \underline{I}(\check{X}^n;Y^n),$$

then

$$\underline{I}(\check{X}^n;Y^n) = \bar{I}(\check{X}^n;Y^n).$$

***Proof*:** We know that

$$\underline{I}(\tilde{X}^n;Y^n) = \sup_{X^n} \underline{I}(X^n;Y^n) = \sup_{X^n} \bar{I}(X^n;Y^n) \geq \bar{I}(\tilde{X}^n;Y^n).$$

But $\underline{I}(X^n;Y^n) \leq \bar{I}(\tilde{X}^n;Y^n)$, for all $\tilde{X}^n$. Hence

$$\underline{I}(\tilde{X}^n;Y^n) = \bar{I}(\tilde{X}^n;Y^n).$$

□

Remark: The above lemma states that if (3) holds and there exists an input distribution that achieves the channel capacity, then it *also achieves* the channel resolvability. However, the converse is not true in general; i.e., if (3) holds and there exists an input distribution that achieves the channel resolvability, then it *does not necessarily achieve* the channel capacity.

Observation 1 *If we assume that there exists an input distribution $P^*_{X^n}$ that achieves the channel capacity, then the following two conditions are equivalent:*

1. $\sup_{X^n} \underline{I}(X^n;Y^n) = \sup_{X^n} \bar{I}(X^n;Y^n)$.
2. $\frac{1}{n} i_{X^nW^n}(X^n;Y^n)$ *converges to a constant (which is the capacity C) in probability according to the joint input-output distribution $P_{X^nY^n}$, such that its induced marginal is $P^*_{X^n}$ and the induced conditional distribution $P_{Y^n|X^n}$ is given by the channel transition distribution.*

We will hereafter use the condition stated in the above observation to verify the validity of the strong converse. But first, we note the following result.

Define the strong converse capacity (or strong capacity) C_{SC} as the infimum of the rates R such that for all block codes with rate R and blocklength n,

$$\liminf_{n\to\infty} P_e^{(n)} = 1,$$

where $P_e^{(n)}$ is probability of decoding error. It follows from the definition that

$$C_{SC} = \lim_{\varepsilon\to 1} C_\varepsilon.$$

Lemma 2

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

Proof:

1. $C_{SC} \geq \sup_{X^n} \bar{I}(X^n; Y^n)$: From the definition of the strong converse capacity, we only need to show that if the probability of decoding error of a (sequence of) block code satisfies $\liminf_{n\to\infty} P_e^{(n)} = 1$, its rate must be greater than $\sup_{X^n} \bar{I}(X^n; Y^n)$.

Let $\widetilde{X}^n$ be the input distribution satisfying $\bar{I}(\widetilde{X}^n; Y^n) > \sup_{X^n} \bar{I}(X^n; Y^n) - \varepsilon$, and let $M = e^{nR}$. Also let $P_e^{(n)}$ satisfy $\liminf_{n\to\infty} P_e^{(n)} = 1$.

From Theorem 1 in [6] (also from Feinsteins's lemma), there exists an $(n, M, P_e^{(n)})$ code that satisfies

$$P_e^{(n)} \leq P\left[\frac{1}{n} i_{\widetilde{X}^n W^n}(\widetilde{X}^n; Y^n) \leq \frac{1}{n}\log M + \gamma\right] + \exp\{-\gamma n\},$$

for any $\gamma > 0$[1], which implies

$$(\forall\ \gamma > 0) \quad \liminf_{n\to\infty} P\left[\frac{1}{n} i_{\widetilde{X}^n W^n}(\widetilde{X}^n; Y^n) \leq R + \gamma\right] = 1.$$

The above result is identical to

$$(\forall\ \gamma > 0) \quad \limsup_{n\to\infty} P\left[\frac{1}{n} i_{\widetilde{X}^n W^n}(\widetilde{X}^n; Y^n) > R + \gamma\right] = 0.$$

Finally, by the definition of sup-information rate, R must be greater than $\bar{I}(\widetilde{X}^n; Y^n) > \sup_{X^n} \bar{I}(X^n; Y^n) - \varepsilon$. Since ε can be made arbitrarily small, we have the desired result.

2. $C_{SC} = \sup_{X^n} \bar{I}(X^n; Y^n)$: If $C_{SC} > \sup_{X^n} \bar{I}(X^n; Y^n)$, then there exists a code with rate $C_{SC} > R = \frac{1}{n}\log M > \sup_{X^n} \bar{I}(X^n; Y^n) + \varepsilon$ such that

$$\liminf_{n\to\infty} P_e^{(n)} < 1, \tag{4}$$

[1]To make it clear, we re-phrase Theorem 1 in [6] as follows.
Fix n and $0 < P_e^{(n)} < 1$, and also fix the input distribution $P_{\widetilde{X}^n}$ on $\mathcal{A}^n$. Then for every $\gamma > 0$, there exists an $(n, M, P_e^{(n)})$ code for the given transition probability W^n that satisfies

$$P_e^{(n)} \leq P\left[\frac{1}{n} i_{\widetilde{X}^n W^n}(\widetilde{X}^n; Y^n) \leq \frac{1}{n}\log M + \gamma\right] + \exp\{-\gamma n\}.$$

for some $\varepsilon > 0$. From [6,Theorem 4], every (n, M) code satisfies,

$$P_e^{(n)} \geq P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \frac{1}{n}\log M - \frac{\varepsilon}{2}\right] - \exp\{-\varepsilon n/2\},$$

where X^n places probability mass $1/M$ on each codeword. Hence,

$$\begin{aligned}
& \liminf_{n\to\infty} P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \frac{1}{n}\log M - \frac{\varepsilon}{2}\right] - \exp\{-\varepsilon n/2\} \\
= & \liminf_{n\to\infty} P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq R - \frac{\varepsilon}{2}\right] - \exp\{-\varepsilon n/2\} \\
\geq & \liminf_{n\to\infty} P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq \bar{I}(X^n; Y^n) + \varepsilon/2\right] - \exp\{-\varepsilon n/2\} \\
= & \ 1,
\end{aligned}$$

which implies $\liminf_{n\to\infty} P_e^{(n)} = 1$, and contradicts (4). □

It can be easily shown that for any input distribution X^n,

$$\underline{I}(X^n; Y^n) \leq \sup\{R : F_X(R) \leq \varepsilon\} \leq \bar{I}(X^n; Y^n),$$

where

$$F_X(R) \triangleq \limsup_{n\to\infty} P\left[\frac{1}{n} i_{X^n W^n}(X^n; Y^n) \leq R\right].$$

Hence, from Theorem 6 in [6], if we assume that $\sup_{X^n} \sup\{R : F_X(R) \leq \varepsilon\}$ is continuous in ε, we obtain that

$$C \leq C_\varepsilon \leq C_{SC}.$$

The above equation leads to the following result.

Corollary 1 *$C = S = C_{SC}$ iff $C_\varepsilon = C$ for all $\varepsilon \in (0, 1)$.*

2. Examples of channels satisfying the strong converse

A. Additive noise channel

Consider the channel with common input, noise, and output alphabet, $\mathcal{A} = \{0, 1, \ldots, q-1\}$, described by

$$Y_n = X_n \oplus Z_n,$$

where $\oplus$ denotes addition modulo q and X_n, Z_n and Y_n are respectively the input, noise, and output symbols of the channel at time n, $n = 1, 2, \ldots$. We assume that the input and noise sequences are independent of each other. We also assume that the noise process is stationary and ergodic.

Since the channel is symmetric, the input process that achieves (3) is uniform i.i.d. , which yields a uniform i.i.d. output process. It follows from the Shannon-McMillian theorem that the information spectrum converges to C where $C = \log q - H(Z_\infty)$. Here, $H(Z_\infty)$ denotes the noise entropy rate. Therefore, the strong converse holds, and $C_\varepsilon = C_{SC} = C$ for all $\varepsilon \in (0, 1)$.

Observation 2 *If the noise process is only stationary, then the strong converse does not hold in general. Indeed, by the ergodic decomposition theorem [2], we can show that the additive noise channel is an averaged channel whose components are q-ary channels with stationary ergodic additive noise. In this case, we obtain using Theorem 6 in [6], a general ϵ-capacity formula for this channel:*

$$C_\varepsilon = \log q - F_U^{-1}(1 - \varepsilon),$$

where U is a random variable with cumulative distribution function $F_U(\cdot)$[2] such that the sequence $-\frac{1}{n} \log P(Z^n)$ converges to U in probability. Furthermore, it is known that $U = H_\theta(Z_\infty)$ where $H_\theta(Z_\infty)$ is the entropy rate of the ergodic components θ defined on the space $(\Theta, \sigma(\Theta), G)$[3]. The distribution of U can hence be derived using the mixing distribution G of the average channel. Finally, we remark that

$$\lim_{\epsilon \to 0} C_\varepsilon = \log q - F_U^{-1}(1) = \log q - \mathrm{ess}_\Theta \sup H_\theta(Z_\infty) = C,$$

as expected.

[2]We assume the CDF $F_U(\cdot)$ admits an inverse. Otherwise, we can replace $F_U^{-1}(\cdot)$ by

$$F_U^{-1}(x) \stackrel{\Delta}{=} \sup\{y : F_U(y) < x\}.$$

[3]We assume that the probability space $(\Theta, \sigma(\Theta), G)$ satisfies certain regularity conditions [2].

B. Additive noise channel with input cost constraints

In general, the use of the channel is not free; we associate with each input letter x a nonnegative number $b(x)$, that we call the "cost" of x. The function $b(\cdot)$ is called the cost function. If we use the channel n consecutive times, i.e., we send an input vector $x^n = (x_1, x_2, \ldots, x_n)$, the cost associated with this input vector is "additive"; i.e.,

$$b(x^n) = \sum_{i=1}^{n} b(x_i).$$

For an input process $\{X_i\}_{i=1}^{\infty}$ with block input distribution $P^{(n)}(X^n = x^n)$ the *average cost* for sending X^n is defined by

$$E\left[b(X^n)\right] = \sum_{x^n} P^{(n)}(x^n)\ b(x^n) = \sum_{i=1}^{n} E\left[b(X_i)\right].$$

We assume that the cost function is "bounded"; i.e., there exists a finite $b_{\max}$ such that $b(x) \leq b_{\max}$ for all x in the set $\{0, \ldots, q-1\}$.

Definition 3 *An n-dimensional input random vector $X^n = (X_1, X_2, \ldots, X_n)$ that satisfies*

$$\frac{1}{n} E\left[b(X^n)\right] \leq \beta,$$

is called a β-admissible input vector. We denote the set of n-dimensional β-admissible input distributions by $\tau_n(\beta)$:

$$\tau_n(\beta) = \left\{ P^{(n)}(X^n) : \frac{1}{n} E\left[b(X^n)\right] \leq \beta \right\}.$$

Recall that a channel is said to be stationary if for every stationary input, the joint input-output process is stationary. Furthermore, a channel is said to be ergodic if for every ergodic input process, the joint input-output process is ergodic. It is known that a channel with stationary mixing additive noise is ergodic [2,5].

Lemma 3 *If the noise process is stationary and mixing, then the strong converse holds:*

$$C_\varepsilon(\beta) = C(\beta) = \lim_{n \to \infty}\ C_n(\beta),$$

where $C_n(\beta)$ is the n'th capacity-cost function given by

$$C_n(\beta) \stackrel{\triangle}{=} \max_{P^{(n)}(X^n) \in \tau_n(\beta)} \frac{1}{n} I(X^n; Y^n).$$

Proof: Since the channel is a causal, historyless[4] and stationary ergodic channel, and the cost function is additive and bounded, then there exists a stationary ergodic input process that achieves $C(\beta)$. This follows from the dual result on the distortion rate function $D(R)$ of stationary ergodic sources, which states that for a stationary ergodic source with additive and bounded distortion measure, there exists a stationary ergodic input-output process $P_{X^nY^n}$ that achieves $D(R)$ such that the induced marginal P_{X^n} is the source distribution [2,3].

Therefore, if we form the joint input-output process $\{(X_n, Y_n)\}_{n=1}^{\infty}$ using the stationary ergodic input process that achieves $C(\beta)$, we obtain that $\{(X_n, Y_n)\}_{n=1}^{\infty}$ is stationary ergodic. Hence, $\frac{1}{n} i_{X^nY^n}(X^n; Y^n)$ converges to $C(\beta)$ in probability. □

General capacity formula with feedback

Consider a discrete channel with output feedback. By this we mean that there exists a "return channel" from the receiver to the transmitter; we assume this return channel is noiseless, delayless, and has large capacity. The receiver uses the return channel to inform the transmitter what letters were actually received; these letters are received at the transmitter before the next letter is transmitted, and therefore can be used in choosing the next transmitted letter.

A feedback code with blocklength n and rate R consists of sequence of encoders

$$f_i : \{1, 2, \ldots, 2^{nR}\} \times \mathcal{B}^{i-1} \to \mathcal{A}$$

for $i = 1, 2, \ldots, n$, along with a decoding function

$$g : \mathcal{B}^n \to \{1, 2, \ldots, 2^{nR}\},$$

[4]Recall that a channel is said to be causal (with no anticipation) if for a given input and a given input-output history, its current output is independent of future inputs. Furthermore, a channel is said to be historyless (with no input memory) if its current output is independent of previous inputs. Refer to [2] for more rigorous definitions of causal and historyless channels.

where $\mathcal{A}$ and $\mathcal{B}$ are the input and output alphabets, respectively. The interpretation is simple: If the user wishes to convey message $V \in \{1, 2, \ldots, 2^{nR}\}$ then the first code symbol transmitted is $X_1 = f_1(V)$; the second code symbol transmitted is $X_2 = f_2(V, Y_1)$, where Y_1 is the channel's output due to X_1. The third code symbol transmitted is $X_3 = f_3(V, Y_1, Y_2)$, where Y_2 is the channel's output due to X_2. This process is continued until the encoder transmits $X_n = f_n(V, Y_1, Y_2, \ldots, Y_{n-1})$. At this point the decoder estimates the message to be $g(Y^n)$, where $Y^n = [Y_1, Y_2, \ldots, Y_n]$.

We assume that V is uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$. The probability of decoding error is thus given by:

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum_{k=1}^{2^{nR}} Pr\{g(Y^n) \neq V | V = k\} = Pr\{g(Y^n) \neq V\}.$$

We say that a rate R is *achievable (admissible)* if there exists a sequence of codes with blocklength n and rate R such that

$$\lim_{n \to \infty} P_e^{(n)} = 0.$$

We will denote the capacity of the channel with feedback by C_{FB}. As before, C_{FB} is the supremum of all admissible feedback code rates.

Lemma 4 *The general capacity formula of an arbitrary channel with feedback is*

$$C_{FB} = \sup_{X^n} \underline{I}(V; Y^n),$$

where the supremum is taken over all possible feedback encoding schemes.[5]

***Proof*:**

1. $C_{FB} \leq \sup_{(f_1, \ldots, f_n)} \underline{I}(V; Y^n)$.

We first state the following result.

[5] $$\sup_{X^n} \underline{I}(V; Y^n) = \sup_{X^n = (f_1(V), f_2(V, Y_1), \ldots, f_n(V, Y^{n-1}))} \underline{I}(V; Y^n) = \sup_{(f_1, f_2, \ldots, f_n)} \underline{I}(V; Y^n).$$

Proposition 1 *For a feedback code of blocklength n and size M, the probability of error satisfies*

$$P_e^{(n)} \;\geq\; P\left[\frac{1}{n}\imath_{WY^n}(W;Y^n) \;\leq\; \frac{1}{n}\log M - \gamma\right] - \exp\{-\gamma n\},$$

for every $\gamma > 0$, where $P_W(W=w) = 1/M$ for all w.

The proof of the proposition is as follows. Let $\beta = \exp\{-\gamma n\}$. Define

$$\begin{aligned}
\mathcal{L} &\stackrel{\triangle}{=} \left\{(w,b^n) \in \{1,2,\ldots,M\}\times\mathcal{Y}^n \;:\; P_{W|Y^n}(w|b^n) \leq \beta\right\} \\
&= \left\{(w,b^n) \in \{1,2,\ldots,M\}\times\mathcal{Y}^n \;:\; \frac{1}{n}\imath_{WY^n}(w;b^n) \;\leq\; \frac{1}{n}\log M - \gamma\right\} \\
&= \cup_{w=1}^{M}\{w\}\times\mathcal{B}_w,
\end{aligned}$$

where $\mathcal{B}_w \stackrel{\triangle}{=} \{b^n \in \mathcal{Y}^n \;:\; P_{W|Y^n}(w|b^n) \leq \beta\}$. By defining $\mathcal{D}_w \in \mathcal{Y}^n$ be the decoding set corresponding to w, we obtain

$$\begin{aligned}
P_{WY^n}(\mathcal{L}) &= \sum_{w=1}^{M} P_{WY^n}(\{w\}\times\mathcal{B}_w) \\
&= \sum_{w=1}^{M} P_{WY^n}(\{w\}\times(\mathcal{B}_w\cap\mathcal{D}_w^c)) + \sum_{w=1}^{M} P_{WY^n}(\{w\}\times(\mathcal{B}_w\cap\mathcal{D}_w)) \\
&= \sum_{w=1}^{M}\frac{1}{M}P_{Y^n|W}(\mathcal{B}_w\cap\mathcal{D}_w^c|w) + \sum_{w=1}^{M} P_{WY^n}(\{w\}\times(\mathcal{B}_w\cap\mathcal{D}_w)) \\
&\leq \sum_{w=1}^{M}\frac{1}{M}P_{Y^n|W}(\mathcal{D}_w^c|w) + \beta P_{Y^n}(\cup_{w=1}^{M}\mathcal{D}_w), \\
&\qquad\qquad \text{because } \mathcal{D}_w \text{ are pair–wise disjoint.} \\
&\leq P_e^{(n)} + \beta.
\end{aligned}$$

Based on this proposition, we can show that

$$C_{FB} \leq \sup_{(f_1,\ldots,f_n)} \underline{I}(V;Y^n)$$

using proof-by-contradiction [6].

2. $C_{FB} \geq \sup_{(f_1,\ldots,f_n)} \underline{I}(V;Y^n)$.

This follows directly using Feinstein's lemma as in [6]. □

General formula for the Neyman-Pearson hypothesis testing error exponent

In this section, we consider a Neyman-Pearson hypothesis testing problem for testing a null hypothesis $H_0 : P_{X^n}$ against an alternative hypothesis $H_1 : Q_{X^n}$ based on a sequence of random observations $X^n = (X_1, \ldots, X_n)$, which is supposed to exhibit a probability distribution of either P_{X^n} or Q_{X^n}. Upon receipt of the n observations, a final decision about the nature of the random observations is made so that the type-II error probability β_n, subject to a *fixed* upper bound ε on the type-I error probability α_n, is minimized. The type-I error probability is defined as the probability of accepting hypothesis H_1 when actually H_0 is true; while the type-II error probability is defined as the probability of accepting hypothesis H_0 when actually H_1 is true [1].

For arbitrary observations (not necessarily stationary, ergodic), we derive a general formula for the type-II error exponent subject to a constant upper bound ε on the type-I error probability. This is given in the following lemma.

Lemma 5 *Given a sequence of random observations $X^n = (X_1, \ldots, X_n)$ which is assumed to have a probability distribution either P_{X^n} or Q_{X^n}, the type-II error exponent satisfies*

$$\sup\{D : \underline{F}(D) < \varepsilon\} \ \leq \ \limsup_{n\to\infty} -\frac{1}{n}\log \beta_n^*(\varepsilon) \ \leq \ \sup\{D : \underline{F}(D) \leq \varepsilon\}, \tag{5}$$

$$\sup\{D : \bar{F}(D) < \varepsilon\} \ \leq \ \liminf_{n\to\infty} -\frac{1}{n}\log \beta_n^*(\varepsilon) \ \leq \ \sup\{D : \bar{F}(D) \leq \varepsilon\}, \tag{6}$$

where

$$\underline{F}(D) \triangleq \liminf_{n\to\infty} P\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} \leq D\right], \text{ and } \bar{F}(D) \triangleq \limsup_{n\to\infty} P\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} \leq D\right],$$

and $\beta_n^*(\varepsilon)$ represents the minimum type-II error probability subject to a fixed type-I error bound $\varepsilon \in (0, 1)$.

Proof: We first prove the lower bound of $\limsup_{n\to\infty} -(1/n)\log \beta_n^*(\varepsilon)$. For any D satisfying $\underline{F}(D) < \varepsilon$, there exists $\delta > 0$ such that $\underline{F}(D) < \varepsilon - 2\delta$; and hence, by the definition of $\underline{F}(D)$, ($\exists$

a subsequence $\{n_j\}$ and N) such that for $j > N$,

$$P\left[\frac{1}{n_j}\log\frac{P(X^{n_j})}{Q(X^{n_j})} \le D\right] \;\le\; \varepsilon - \delta \;<\; \varepsilon.$$

$$\begin{aligned}
\therefore\quad \beta^*_{n_j}(\varepsilon) &\le Q\left[\frac{1}{n_j}\log\frac{P(X^{n_j})}{Q(X^{n_j})} > D\right] \\
&\le P\left[\frac{1}{n_j}\log\frac{P(X^{n_j})}{Q(X^{n_j})} > D\right]\cdot\exp\{-n_jD\} \\
&\le \exp\{-n_jD\}\,. \qquad (7)
\end{aligned}$$

Therefore,

$$\limsup_{n\to\infty} -\frac{1}{n}\log\beta^*_n(\varepsilon) \;\ge\; \limsup_{j\to\infty} -\frac{1}{n_j}\log\beta^*_{n_j}(\varepsilon) \;\ge\; D,$$

for any D with $\underline{F}(D) < \varepsilon$.

For the proof of the upper bound of $\limsup_{n\to\infty} -(1/n)\log\beta^*_n(\varepsilon)$, let $\mathcal{U}_n$ be the optimal acceptance region for alternative hypothesis under likelihood ratio partition, which is defined as follows.

$$\mathcal{U}_n \triangleq \left\{\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} < \tau_n\right\} + \eta_n\left\{\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} = \tau_n\right\}, \qquad (8)$$

for some τ_n and possible randomization factor $\eta_n \in [0,1)$. Then $P(\mathcal{U}_n) = \varepsilon$.

Let $\underline{D} = \sup\{D : \underline{F}(D) \le \varepsilon\}$. Then $\underline{F}(\underline{D}+\delta) > \varepsilon$ for any $\delta > 0$. Hence, $(\exists\ \gamma = \gamma(\delta) > 0)$, $\underline{F}(\underline{D}+\delta) > \varepsilon + \gamma$.

By the definition of $\underline{F}(\underline{D}+\delta)$, $(\exists\ N)(\forall\ n > N)$

$$P\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} \le \underline{D}+\delta\right] \;>\; \varepsilon + \frac{\gamma}{2}.$$

Therefore,

$$\begin{aligned}
\beta^*_n(\varepsilon) &= Q\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} > \tau_n\right] + (1-\eta_n)\cdot Q\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} = \tau_n\right] \\
&\ge Q\left[\underline{D}+\delta \ge \frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} > \tau_n\right] + (1-\eta_n)\cdot Q\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} = \tau_n\right] \\
&\ge \left(P\left[\underline{D}+\delta \ge \frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} > \tau_n\right] + (1-\eta_n)\cdot P\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} = \tau_n\right]\right) \\
&\qquad \times\exp\{-n(\underline{D}+\delta)\}
\end{aligned}$$

$$
\begin{aligned}
&= \left(P\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} \le \underline{D}+\delta\right] - P\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} < \tau_n\right] \right. \\
&\quad \left. -\eta_n P\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} = \tau_n\right]\right) \times \exp\{-n(\underline{D}+\delta)\} \\
&\ge \left(\varepsilon + \frac{\gamma}{2} - \varepsilon\right)\exp\{-n(\underline{D}+\delta)\}, \quad \text{for } n > N \\
&= \frac{\gamma}{2}\exp\{-n(\underline{D}+\delta)\}, \quad \text{for } n > N. \qquad (9)
\end{aligned}
$$

$$
\therefore \quad \limsup_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) \ \le \ \underline{D}+\delta.
$$

Since δ can be made arbitrarily small,

$$
\limsup_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) \ \le \ \underline{D}.
$$

Similarly, to prove the lower bound of $\liminf_{n\to\infty} -(1/n)\log\beta_n^*(\varepsilon)$, we first note that for any D satisfying $\bar{F}(D) < \varepsilon$, $(\exists\ \delta > 0)$ such that $\bar{F}(D) < \varepsilon - 2\delta$; and hence, by the definition of $\bar{F}(D)$, $(\exists\ N)(\forall\ n > N)$,

$$
P\left[\frac{1}{n}\log\frac{P(X^n)}{Q(X^n)} \le D\right] \ \le \ \varepsilon - \delta \ < \ \varepsilon.
$$

By following the same procedure of (7), we have for $n > N$,

$$
\beta_n^*(\varepsilon) \ \le \ \exp\{-nD\},
$$

Therefore,

$$
\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) \ \ge \ D,
$$

for any D with $\bar{F}(D) < \varepsilon$.

Then for the proof of the upper bound of $\liminf_{n\to\infty} -(1/n)\log\beta_n^*(\varepsilon)$, let $\bar{D} = \sup\{D : \bar{F}(D) \le \varepsilon\}$. Then $\bar{F}(\bar{D}+\delta) > \varepsilon$ for any $\delta > 0$. Hence, $(\exists\ \gamma = \gamma(\delta) > 0)$, $\bar{F}(\bar{D}+\delta) > \varepsilon + \gamma$.

By the definition of $\bar{F}(\bar{D}+\delta)$, $(\exists$ a subsequence $\{n_j\}$ and $N)$ such that for $j > N$,

$$
P\left[\frac{1}{n_j}\log\frac{P(X^{n_j})}{Q(X^{n_j})} \le \bar{D}+\delta\right] \ > \ \varepsilon + \frac{\gamma}{2}.
$$

Therefore, by following the same procedure as (9), we have for $j > N$,

$$
\beta_{n_j}^*(\varepsilon) \ \ge \ \frac{\gamma}{2}\exp\left\{-n_j(\bar{D}+\delta)\right\}
$$

$$\therefore \quad \liminf_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) \;\le\; \liminf_{j\to\infty} -\frac{1}{n_j}\log\beta_{n_j}^*(\varepsilon) \;\le\; \bar{D}+\delta.$$

Since δ can be made arbitrarily small,

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) \;\le\; \bar{D}.$$

□

Remarks:

- Both $\bar{F}(D)$ and $\underline{F}(D)$ are non-decreasing; hence, the number of discontinuous points for both functions is countable.

- When the normalized log-likelihood ratio converges in probability to a constant D_c under null distribution which is the case for most detection problems of interest, the type-II error exponent is that constant D_c, and is independent of the type-I error bound ε. For example, in a special case of i.i.d. data source with $|E_P[\log P(X)/Q(X)]| < \infty$, both functions degenerate to the form

$$\bar{F}(D) = \underline{F}(D) = 1 \qquad \text{if } D > D_c$$
$$\bar{F}(D) = \underline{F}(D) = 0 \qquad \text{if } D < D_c,$$

  where $D_c \triangleq E_P[\log P(X)/Q(X)]$. As a result, for $\varepsilon \in (0,1)$,

$$\limsup_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) = \liminf_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) = D_c.$$

- The significance of the general type-II error exponent formula of fixed level becomes transparent when the spectrum (the cumulative distribution function) of the normalized log-likelihood ratio converges in probability under P (which is weaker than convergence in mean) to a random variable Z with invertible cumulative distribution function $F(\cdot)$. In this case, the type-II error exponent can be explicitly written as

$$\lim_{n\to\infty} -\frac{1}{n}\log\beta_n^*(\varepsilon) = F^{-1}(\varepsilon),$$

for $\varepsilon \in (0,1)$. A more extreme case is that Z is almost surely a constant which is

$$\lim_{n\to\infty} \frac{1}{n} D\left(P_{X^n} \| Q_{X^n}\right),$$

if the limit exists, where $D(\cdot\|\cdot)$ is the Kullback-Leibler divergence of two probability measures. This result coincides with that obtained from Stein's Lemma. This is also the counterpart result of the strong converse condition (i.e., the ε-capacity is independent of ε) for discrete memoryless channels (DMC) [6].

Summary

In this paper, we considered three different problems related to the work of Verdú and Han on channel capacity [6]. Pertinent observations concerning the validity of the strong converse to the channel coding theorem, as well as examples of channels for which the strong converse holds, were provided. General expressions for the feedback capacity of arbitrary channels and the Neyman-Pearson type-II error exponent of constant test level were also derived.

References

1. R. E. Blahut, *Principles and Practice of Information Theory*, Addison Wesley, New York (1987).

2. R. M. Gray, *Entropy and Information Theory*, Springer-Verlag New York Inc. (1990).

3. R. M. Gray, *Source Coding Theory*, Kluwer Academic Publishers, Norwell, MA (1990).

4. T. S. Han and S. Verdú, "Approximation Theory of Output Statistics", *IEEE Transactions on Information Theory*, Vol. 39, No. 3, pp. 752-772 (1993).

5. M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*, Holden-Day, San Francisco (1964).

6. S. Verdú and T. S. Han, "A General Formula for Channel Capacity", *IEEE Transactions on Information Theory*, vol. 40, pp. 1147-1157, July 1994.