Asymptotic Neyman-Pearson Games for Converse to the Channel Coding Theorem

Pierre Moulin
Dept of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

***Abstract*—Upper bounds have recently been derived on the maximum volume of length-n codes for memoryless channels subject to either a maximum or an average decoding error probability ϵ. These bounds are expressed in terms of a minmax game whose variables are n-dimensional probability distributions and whose payoff function is the power of a Neyman-Pearson test at significance level $1-\epsilon$. We derive the exact asymptotics (as $n \to \infty$) of this game by relating it to a problem that admits an asymptotic saddlepoint with an equalizer property.**

I. Introduction

Strassen [1], Polyanskiy *et al.* [2], and Hayashi [3] have derived refined asymptotics for coding on memoryless channels. For any length-n code with tolerable decoding error probability ϵ, they found that the maximum volume of the code takes the form

$$M^*(n,\epsilon) = \exp\{nC - \sqrt{nV}\, Q^{-1}(\epsilon) + O(\log n)\} \quad \text{as } n \to \infty \tag{1}$$

under both the maximum and average error probability criteria, subject to some regularity conditions on the channel law. In (1), C is channel capacity, V is channel dispersion, and $Q(x) \triangleq \int_x^\infty (2\pi)^{-1/2} \exp\{-t^2/2\}\, dt$. The $O(\log n)$ term is equal to $\frac{1}{2}\log n$ for symmetric channels [1, footnote p.692].

Our recent paper [4] sharpened (1) using strong large deviations analysis and exact central limit asymptotics (again under regularity conditions on the channel law). Under the average error probability criterion, we have

$$\begin{aligned} \underline{A}_\epsilon + o(1) \le \log M^*(n,\epsilon) - \left[nC - \sqrt{nV}\, Q^{-1}(\epsilon) + \frac{1}{2}\log n\right] \\ \le \overline{A}_\epsilon + o(1) \end{aligned} \tag{2}$$

where $\underline{A}_\epsilon$ and $\overline{A}_\epsilon$ are two constants. For symmetric channels, $\overline{A}_\epsilon = \underline{A}_\epsilon + 1$.

The lower bound is achieved using iid random codes and ML decoding. The upper bound is based on a *metaconverse* [2] taking the form of a maxmin optimization problem whose variables are n-dimensional probability distributions on the channel input and output sequences and whose payoff function is the power of a Neyman-Pearson test at significance level $1-\epsilon$.

This paper derives the upper bound of (2) via the asymptotics of the above Neyman-Pearson game. A more tedious approach was briefly sketched in [4], involving a decomposition of the code into five subcodes. The proof in this paper starts from a converse for constant-composition codes (Thm 2.2) and then derives a converse for general codes under the maximum error probability criterion (Thm 3.1) and finally, a converse under the average error criterion (Thm 4.3). The upper bound of (2) is shown to hold for all three problems.

Notation. We use uppercase letters for random variables (rv's), lowercase letters for their individual values, calligraphic letters for alphabets, and boldface letters for sequences. The set of all probability distributions over a finite set $\mathcal{X}$ is denoted by $\mathscr{P}(\mathcal{X})$. Mathematical expectation with respect to probability distribution P is denoted by the symbol $\mathbb{E}_P$. Given a distribution P on the rv X and a conditional distribution W on another rv Y given X, we denote by $P \times W$ the joint distribution on (X,Y) and by (PW) the marginal distribution on Y. The indicator function of a set $\mathcal{A}$ is denoted by $\mathbb{1}\{x \in \mathcal{A}\}$. All logarithms are natural logarithms. The notations $f(n) = o(g(n))$ (small oh) and $f(n) = O(g(n))$ (big oh) indicate that $\lim_{n\to\infty} \frac{f(n)}{g(n)}$ is zero and finite, respectively.

A. Definitions

Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite alphabets and $(W, \mathcal{X}, \mathcal{Y})$ a discrete memoryless channel. The Kullback-Leibler divergence between two distributions P and Q on a common alphabet is denoted by $D(P\|Q) \triangleq \mathbb{E}_P[\log \frac{P(X)}{Q(X)}]$, divergence variance by $V(P\|Q) \triangleq \mathbb{E}_P[\log \frac{P(X)}{Q(X)}]^2 - D^2(P\|Q)$, and divergence third central moment by $T(P\|Q) \triangleq \mathbb{E}_P[\log \frac{P(X)}{Q(X)} - D(P\|Q)]^3$. Given a $\mathcal{X}$-valued rv X distributed as P and two conditional distributions W and Q on a $\mathcal{Y}$-valued rv Y given X, we denote by $D(W\|Q|P) = \mathbb{E}_{P\times W} \log \frac{W(Y|X)}{Q(Y|X)}$ the conditional KL divergence between W and Q given P, and likewise by $V(W\|Q|P)$ and $T(W\|Q|P)$ the conditional divergence variance and the conditional divergence third central moment. We also define the conditional skewness $S(W\|Q|P) \triangleq T(W\|Q|P)/V(W\|Q|P)^{3/2}$.

A real rv L is of the lattice type if there exist numbers d and l_0 such that L belongs to the lattice $\{l_0 + kd,\ k \in \mathbb{Z}\}$ with probability 1. The largest d for which this holds is called the *span* of the lattice, and l_0 is the *offset*. The sum of a nonlattice rv and a lattice rv is a nonlattice rv. The sum of two lattice rv's is a lattice rv if and only if the ratio of their spans is a rational number. For each $d > 0$, let $\mathscr{P}_{\text{lat}}(d) \triangleq \{Q \in \mathscr{P}(\mathcal{Y}) \ :\ \log \frac{W(Y|X)}{Q(Y)}$ is a lattice rv with span $d\}$.

The empirical distribution (n-type) on $\mathcal{X}$ of a sequence $\mathbf{x} \in$

$\mathcal{X}^n$ is defined by $\hat{P}_{\mathbf{x}}(x) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x_i = x\}$, $x \in \mathcal{X}$. We denote by $T[P]$ the set of all sequences of type P (type class), by $U_{\mathbf{X}|P}$ the uniform distribution over type class $T[P]$, and by $\mathscr{P}_n(\mathcal{X}) \subset \mathscr{P}(\mathcal{X})$ the set of n-types over $\mathcal{X}$.

Define $l(x,y) = \log \frac{W(y|x)}{(PW)(y)}$, $x \in \mathcal{X}, y \in \mathcal{Y}$. For some DMCs with capacity-achieving distribution, the rv $l(X,Y)$ is of the lattice type. Due to space constraints, this case is not treated here.

Let $W_x \triangleq \{W(\cdot|x)\} \in \mathscr{P}(\mathcal{Y})$ for each $x \in \mathcal{X}$. We define the following moments of the rv $l(X,Y)$ with respect to the joint distribution $P \times W$: the mean (= mutual information) $I(P;W)$, the *conditional* information variance (given X) $V(P,W) = \sum_{x\in\mathcal{X}} P(x)V(W_x\|(PW))$ the *conditional* third central moment (given X) $T(P,W) = \sum_{x\in\mathcal{X}} P(x)T(W_x\|(PW))$, and the *conditional* skewness $S(P,W) = \frac{T(P,W)}{[V(P,W)]^{3/2}}$.

We also define the reverse channel $\breve{W}_y(x) = \frac{W(y|x)P(x)}{(PW)(y)}$ via Bayes' rule; the Fisher information matrix

$$J_{xx'}(P,W) \triangleq -\frac{\partial^2 I(P,W)}{\partial P(x)\partial P(x')} = \sum_y \frac{W(y|x)W(y|x')}{(PW)(y)} \quad (3)$$

($x, x' \in \mathcal{X}$) which satisfies $\sum_x P(x)J_{xx'}(P;W) = 1\, \forall x'$; and the $|\mathcal{X}|-1$ dimensional linear space

$$\mathcal{L}(\mathcal{X}) \triangleq \left\{ h \in \mathbb{R}^{|\mathcal{X}|} \ : \ \sum_{x\in\mathcal{X}} h(x) = 0 \right\}.$$

Let P^* be the capacity-achieving input distribution, assumed to be unique (see **(A1)** below). Define the vectors

$$\begin{aligned}
\breve{v}(x) &\triangleq -2[\mathbb{E}_{W_x} D(\breve{W}_Y\|P^*) - I(P^*;W)] && (4)\\
\tilde{v}(x) &\triangleq V(W_x\|(P^*W)) - V(P^*;W) && (5)\\
v(x) &\triangleq \tilde{v}(x) + \breve{v}(x), \quad x \in \mathcal{X} && (6)\\
&= \partial V(P;W)/\partial P(x) + \text{constant}
\end{aligned}$$

all of which have zero mean under P^*. We also denote by $J^\dagger$ the pseudo-inverse of $J(P^*;W)$ and let

$$\|v\|_{J^\dagger} \triangleq vJ^\dagger v \triangleq \sqrt{\sum_{x,x'} v(x)v(x')J^\dagger_{xx'}}.$$

Next, define the constant

$$A_{\text{ns}} \triangleq \frac{1}{V(P,W)} \left(\|v\|_{J^\dagger} - \|\breve{v}\|_{J^\dagger}\right) \quad (7)$$

which is nonzero only for nonsymmetric channels, hence the subscript "ns". Finally, define

$$\begin{aligned}
h^* &\triangleq \frac{t_\epsilon}{2\sqrt{V(P^*;W)}} J^\dagger v && \in \mathcal{L}(\mathcal{X}) && (8)\\
P_n^* &\triangleq P^* + n^{-1/2}h^* && \in \mathscr{P}(\mathcal{X}) && (9)\\
\tilde{h}^* &\triangleq \frac{t_\epsilon}{2\sqrt{V(P^*;W)}} J^\dagger \tilde{v} && \in \mathcal{L}(\mathcal{X}) && (10)\\
\tilde{P}_n^* &\triangleq P^* + n^{-1/2}\tilde{h}^* && \in \mathscr{P}(\mathcal{X}) && (11)\\
Q_n &\triangleq (\tilde{P}_n^* W) && \in \mathscr{P}(\mathcal{Y}). && (12)
\end{aligned}$$

B. Achievable Rates

The message m to be transmitted is drawn uniformly from the message set $\mathcal{M}_n = \{1, 2, \cdots, M\}$. A code is a pair of encoder mapping $f_n : \mathcal{M}_n \to \mathsf{F} \subseteq \mathcal{X}^n$, $\mathbf{x}(m) = f_n(m)$, and decoder mapping $g_n : \mathcal{Y}^n \to \mathcal{M}_n$, $\hat{m} = g_n(\mathbf{y})$. The code has volume (or size) M and rate $R_n = \frac{1}{n}\log M$. Shannon capacity is denoted by $C = \max_{P\in\mathscr{P}(\mathcal{X})} I(P;W)$.

Assume the following:

- **(A1)** The capacity-achieving distribution P^* is unique and $\mathcal{X}$ is its support set: $P^*(x) > 0\ \forall x \in \mathcal{X}$.
- **(A2)** $V(P^*;W) > 0$.
- **(A3)** $\ln \frac{W(Y|X)}{(P^*W)(Y)}$ is a nonlattice rv.

Let $t_\epsilon \triangleq Q^{-1}(\epsilon)$, $V = V(P^*,W)$, $S = S(P^*;W)$. Then the constant $\overline{A}_\epsilon$ of (2) is given by

$$\overline{A}_\epsilon = \frac{t_\epsilon^2}{8} A_{\text{ns}} - \frac{S\sqrt{V}}{6}(t_\epsilon^2 - 1) + \frac{1}{2}t_\epsilon^2 + \frac{1}{2}\log(2\pi V) \quad (13)$$

and the upper bound of (2) holds under Assumptions **(A1)**—**(A3)**. The lower bound is achieved by iid random codes drawn from the distribution P_n^* of (9) and ML decoding.

C. Minimax Converse

Our derivations are based on results from [4] on strong large deviations for binary hypothesis testing as well as on two theorems from [2] which are stated below.

Theorem 1.1: [2, Thm 31 p.2319]. The volume M_{F} of any code with codewords in $\mathsf{F} \subseteq \mathcal{X}^n$ and *maximum* error probability ϵ satisfies

$$M_{\mathsf{F}} \le \inf_{Q_{\mathbf{Y}}} \sup_{\mathbf{x}\in\mathsf{F}} \frac{1}{\beta_{1-\epsilon}(W^n(\cdot|\mathbf{x}), Q_{\mathbf{Y}})}$$

where the supremum is over all feasible codewords, and the infimum is over all probability distributions over $\mathcal{Y}^n$.

Theorem 1.2: [2, Thm 27 p. 2318]. The volume M_{F} of any code with codewords in $\mathsf{F} \subseteq \mathcal{X}^n$ and *average* error probability ϵ satisfies

$$M_{\mathsf{F}} \le \sup_{P_{\mathbf{X}}} \inf_{Q_{\mathbf{Y}}} \frac{1}{\beta_{1-\epsilon}(P_{\mathbf{XY}}, P_{\mathbf{X}} \times Q_{\mathbf{Y}})} \quad (14)$$

where the sup is over all probability distributions over F, and the inf is over all probability distributions over $\mathcal{Y}^n$.

While the order of sup and inf can be exchanged in (14) [6], deriving the asymptotics of this game is the topic of Sec. IV.

The following theorem of [2] will be refined and extended to the average error probability criterion in the next section.

Theorem 1.3: [2, Thm 48 p. 2331]. The volume M of any constant-composition code in $\mathcal{X}^n$ with maximal error probability ϵ satisfies $\log M \le nC - \sqrt{nV}\, t_\epsilon + \frac{1}{2}\log n + F$ for some constant F.

II. Converse for Constant-Composition Codes

For each $\delta > 0$ and $P \in \mathscr{P}(\mathcal{X})$, define a subset of distributions $\mathcal{H}_\delta(P) \subseteq \mathscr{P}(\mathcal{Y})$ as follows. If $\min_{x\in\mathcal{X}} P(x) < \delta$, let $\mathcal{H}_\delta(P) \triangleq \emptyset$. Otherwise let

$$\begin{aligned}
\mathcal{H}_\delta(P) \triangleq \{&Q \in \mathscr{P}(\mathcal{Y}) : \ D(W\|Q|P) < \infty,\\
&\delta \le V(W\|Q|P) < \infty,\ T(W\|Q|P) < \infty\}.
\end{aligned}$$

These sets are nested (increasing as $\delta \downarrow 0$), as are the sets

$$\begin{aligned} \mathcal{R}_\delta &\triangleq \{P \in \mathscr{P}(\mathcal{X}) : (PW) \in H_\delta(P)\} \\ &= \{P \in \mathscr{P}(\mathcal{X}) : \delta \le \min_{x \in \mathcal{X}} P(x),\ \delta \le V(P;W)\}. \end{aligned}$$

By Assumptions **(A1)**, **(A2)**, there exists $\delta > 0$ such that

$$P^* \in \mathcal{R}_\delta \quad \text{and} \quad \sup_{P \notin \mathcal{R}_\delta} I(P;W) < C - \delta. \tag{15}$$

Distributions not in $\mathcal{R}_\delta$ will be given a special treatment; they are far from P^*.

Define the following functions of $P \in \mathscr{P}(\mathcal{X})$ and $Q \in \mathscr{P}(\mathcal{Y})$. First assume that $Q \notin \cup_{d>0} \mathscr{P}_{\text{lat}}(d)$, i.e., $\log[W(Y|X)/Q(Y)]$ is not a lattice rv. Then

$$\begin{aligned} F_\epsilon(W\|Q|P) &\triangleq \frac{1}{2}t_\epsilon^2 - \frac{1}{6}S(W\|Q|P)\sqrt{V(W\|Q|P)}(t_\epsilon^2 - 1) \\ &\quad + \frac{1}{2}\log(2\pi V(W\|Q|P)), \end{aligned} \tag{16}$$

$$\zeta_{n,\delta}(P,Q) \triangleq \begin{cases} nD(W\|Q|P) - \sqrt{nV(W\|Q|P)}\,t_\epsilon + F_\epsilon(W\|Q|P) \\ \qquad : \ Q \in \mathcal{H}_\delta(P) \\ nD(W\|Q|P) + \sqrt{\frac{nV(W\|Q|P)}{1-\epsilon}} + \log\frac{1-\epsilon}{2} \\ \qquad : \ Q \notin \mathcal{H}_\delta(P), \end{cases} \tag{17}$$

$$\begin{aligned} F_\epsilon(P;W) &\triangleq \frac{1}{2}t_\epsilon^2 - \frac{1}{6}S(P;W)\sqrt{V(P;W)}(t_\epsilon^2 - 1) \\ &\quad + \frac{1}{2}\log(2\pi V(P;W)), \end{aligned} \tag{18}$$

$$\zeta_{n,\delta}(P;W) \triangleq \begin{cases} nI(P;W) - \sqrt{nV(P;W)}t_\epsilon + F_\epsilon(P;W) \\ \qquad : \ P \in \mathcal{R}_\delta \\ nI(P;W) + \sqrt{\frac{nV(P;W)}{1-\epsilon}} + \log\frac{1-\epsilon}{2} \\ \qquad : \ P \notin \mathcal{R}_\delta \end{cases} \tag{19}$$

and the constant (recall (13))

$$F_\epsilon \triangleq F_\epsilon(P^*;W) = \overline{A}_\epsilon - \frac{t_\epsilon^2}{8}A_{\text{ns}}. \tag{20}$$

Observe that

$$\zeta_{n,\delta}(P,Q) = \zeta_{n,\delta}(P;W) \quad \text{for } Q = (PW). \tag{21}$$

If $Q \in \mathscr{P}_{\text{lat}}(d)$ for some $d > 0$, the same definitions apply, with a constant term $l(d) \triangleq \ln\frac{d}{1-\exp\{-d\}}$ added to the right side of (16). The function $l(d)$ is continuous and increases from 0 to ∞ as d increases from 0 to ∞. It may be shown that $\sup\{l(d) \ : \ \exists P \in \mathscr{P}(\mathcal{X}) \ : \ (PW) \in \mathscr{P}_{\text{lat}}(d),\ \max_x |P(x) - P^*(x)| \le \delta\} \downarrow 0$ as $\delta \downarrow 0$. Hence (19) holds up to an $o(1)$ term in a vanishing neighborhood of P^*, including the subset associated with lattice rv's ($\exists d > 0 \ : (PW) \in \mathscr{P}_{\text{lat}}(d)$).

The proposition below follows from [4, Prop. 2.2] in the case $Q \in \mathcal{H}_\delta(P)$ and coincides with [1, Thm 1.1] in the iid case. Prop. 2.1 strengthens [1, Thm 3.1] and [2, Lemma 58].

Proposition 2.1: For any sequence $\mathbf{x}$ of type $P \in \mathscr{P}_n(\mathcal{X})$ and any distribution $Q \in \mathscr{P}(\mathcal{Y})$, the following inequality holds:

$$-\log\beta_{1-\epsilon}(W^n(\cdot|\mathbf{x}), Q^n) \le \zeta_{n,\delta}(P,Q) + \frac{1}{2}\log n + o(1). \tag{22}$$

Sketch of the proof. Define $\overline{D}_n = \frac{1}{n}\sum_{i=1}^n D(W_{x_i}\|Q) = D(W\|Q|P)$ and likewise $\overline{V}_n = \frac{1}{n}\sum_{i=1}^n V(W_{x_i}\|Q) = V(W\|Q|P)$, $\overline{T}_n = \frac{1}{n}\sum_{i=1}^n T(W_{x_i}\|Q) = T(W\|Q|P)$, $\overline{S}_n = \overline{T}_n \overline{V}_n^{-3/2} = S(W\|Q|P)$, and

$$na_n \triangleq n\overline{D}_n - \sqrt{n\overline{V}_n}t_\epsilon - \frac{1}{6}\overline{S}_n\sqrt{\overline{V}_n}(t_\epsilon^2 - 1).$$

If $Q \in \mathcal{H}_\delta(P)$, let $T_n \triangleq \sum_{i=1}^n \ln\frac{W(Y_i|x_i)}{Q(Y_i)}$. By [4, Prop. 2.2] we have

$$Q^n[T_n \ge na_n] = \frac{\exp\{-na_n - \frac{1}{2}t_\epsilon^2 + o(1)\}}{\sqrt{2\pi\overline{V}_n n}} \tag{23}$$

and by the Cornish-Fisher formula [4, (21)] we have

$$W^n[T_n \ge na_n|\mathbf{X} = \mathbf{x}] = 1 - \epsilon + o(n^{-1/2}) \tag{24}$$

when $Q \notin \cup_{d>0}\mathscr{P}_{\text{lat}}(d)$, i.e., $\log[W(Y|X)/Q(Y)]$ is not a lattice rv. If $\exists d > 0$ such that $Q \in \mathscr{P}_{\text{lat}}(d)$ then (23) holds if the right side is multiplied by a sequence γ_n that can be explicitly identified, is bounded from above and below, and takes the value $d/(1 - e^{-d}) \ge 1$ for na_n in the lattice. The inequality (22) follows from (23) (24) and the definitions (16) and (17). If $Q \notin \mathcal{H}_\delta(P)$, the inequality (22) follows from [2, Lemma 59 p.2341]. □

Theorem 2.2: The volume $M[P]$ of any code of constant composition $P \in \mathscr{P}_n(\mathcal{X})$ and (maximum or average) error probability ϵ satisfies

$$\begin{aligned} \log M[P] &\le \zeta_{n,\delta}(P;W) + \frac{1}{2}\log n + o(1) \qquad &(25) \\ &\le nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\log n + \overline{A}_\epsilon + o(1). &(26) \end{aligned}$$

In (26), equality is achieved at $P = P_n^*$ of (9).

Proof. Fix $Q = (PW)$ and $Q_{\mathbf{Y}} = Q^n$. Under the maximum-error probability criterion, (25) follows from Theorem 1.1, Prop. 2.1, and (21). Since $\beta_{1-\epsilon}(W^n(\cdot|\mathbf{x}), Q^n)$ is the same for all $\mathbf{x}$ of type P, (25) also holds under the average-error probability criterion [2, Lemma 29 p. 2318]. The upper bound (26) on $\zeta_{n,\delta}(P;W)$ for $P \in \mathscr{P}_n(\mathcal{X}) \cap \mathcal{R}_\delta$ is given in [4]. The upper bound also holds (and is loose) for $P \in \mathscr{P}_n(\mathcal{X}) \cap \mathcal{R}_\delta^c$ owing to (15) (19). □

III. General Codes, Maximum Error Probability

Theorem 3.1: The volume M of any code with codewords in $\mathcal{X}^n$ and maximum error probability ϵ satisfies

$$M \le \exp\{nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\log n + \overline{A}_\epsilon + o(1)\}. \tag{27}$$

The theorem is proved at the end of this section. First we make some remarks and give some definitions and lemmas.

Fix any $\mathsf{F} \subseteq \mathcal{X}^n$ and let $\mathscr{P}_0$ be any subset of $\mathscr{P}(\mathcal{X})$ such that $\mathbf{x} \in \mathsf{F} \Rightarrow \hat{P}_{\mathbf{x}} \in \mathscr{P}_0$. It follows from Theorem 1.1 with $Q_{\mathbf{Y}} = Q^n$ and Prop. 2.1 that the volume M_{F} of any such code satisfies (for any $\delta > 0$)

$$M_{\mathsf{F}} \le \exp\left\{\inf_{Q \in \mathscr{P}(\mathcal{Y})}\sup_{P \in \mathscr{P}_0}\zeta_{n,\delta}(P,Q) + \frac{1}{2}\log n + o(1)\right\}.$$

At first sight this suggests seeking a solution to the minmax game with payoff function $\zeta_{n,\delta}(P,Q)$ over $\mathscr{P}_0 \times \mathscr{P}(\mathcal{Y})$. Assume $P^* \in \mathscr{P}_0$. Then a version of this game with payoff

$$\lim_{n\to\infty}\frac{1}{n}\zeta_{n,\delta}(P,Q) = D(W\|Q|P) = \sum_{x \in \mathcal{X}} P(x)D(W_x\|Q) \tag{28}$$

admits the well-known *equalizer* saddlepoint solution $(P^*, (P^*W))$. Indeed (28) is linear in P and convex in Q, and

$$D(W\|(P^*W)|P) = D(W\|(P^*W)|P^*) \le D(W\|Q|P^*) \quad (29)$$

$\forall P, Q$, where equality holds because $D(W_x\|(P^*W)) = I(P^*; W)$ for all $x \in \mathrm{supp}\{P^*\} = \mathcal{X}$. Owing to the equalizer property, we have

$$\begin{aligned} \inf_{Q\in\mathscr{P}(\mathcal{Y})} \sup_{P\in\mathscr{P}_0} D(W\|Q|P) &= \sup_{P\in\mathscr{P}_0} \inf_{Q\in\mathscr{P}(\mathcal{Y})} D(W\|Q|P) \\ &= I(P^*; W) \end{aligned}$$

even if $\mathscr{P}_0$ is not a convex set.

For finite n our game generally admits no saddlepoint because the payoff function $\zeta_{n,\delta}(P, Q)$ is nonconcave in P. However

- In the symmetric case where $V(W_x\|(P^*W)) = V(P^*; W)$ and $F(W_x\|(P^*W)) = F(P^*; W)$ for all $x \in \mathcal{X}$, the game clearly admits an *asymptotic saddlepoint* solution $(P^*, (P^*W))$ in the sense that
$$\zeta_{n,\delta}(P, (P^*W)) = \zeta_{n,\delta}(P^*, (P^*W)) \le \zeta_{n,\delta}(P^*, Q) + o(1)$$
$\forall P \in \mathscr{P}(\mathcal{X})$, $\forall Q$, and the asymptotic value of the game is $\zeta_{n,\delta}(P^*, (P^*W)) = \zeta_{n,\delta}(P^*; W)$.
- In the nonsymmetric case, we shall see (in Lemma 3.4) there still exists an *asymptotic saddlepoint* (P_n^*, Q_n) if the maximization over P is restricted to a suitably defined vanishing neighborhood $\mathscr{P}_1$ of P^*.

Instead of applying Theorem 1.1 with $\mathsf{F} = \mathcal{X}^n$ directly, we define a set of subcodes with maximum error probability ϵ each, and derive the asymptotics for these subcodes. The upper bound on M is the sum of the upper bounds on the volume of the subcodes.

Define the "good" class of codewords

$$\mathsf{F}_1 \triangleq \left\{ \mathbf{x} \in \mathcal{X}^n : \zeta_{n,\delta}(\hat{P}_{\mathbf{x}}; W) \ge \zeta_{n,\delta}(P_n^*; W) - \frac{\sqrt{n}}{\log^2 n} \right\} \quad (30)$$

and the corresponding "good" class of distributions over $\mathcal{X}$

$$\mathscr{P}_1 \triangleq \left\{ P \in \mathscr{P}(\mathcal{X}) : \zeta_{n,\delta}(P; W) \ge \zeta_{n,\delta}(P_n^*; W) - \frac{\sqrt{n}}{\log^2 n} \right\} \quad (31)$$

Hence $\mathscr{P}_1 \subset \mathcal{R}_\delta$ for n large enough, and $\mathbf{x} \in \mathsf{F}_1 \Leftrightarrow \hat{P}_{\mathbf{x}} \in \mathscr{P}_1$.

Lemma 3.2: Fix $h, \tilde{h} \in \mathcal{L}(\mathcal{X})$ and let $n^{-1/2} \le \epsilon_n \ll 1$,

$$P_n = P^* + \epsilon_n h, \quad \tilde{P}_n = P^* + n^{-1/2}\tilde{h}, \quad Q_n = (\tilde{P}_n W). \quad (32)$$

Then

$$\begin{aligned} \zeta_{n,\delta}(P_n, Q_n) = \zeta_{n,\delta}(P^*; W) + \frac{1}{2}\tilde{h}J\tilde{h} - \tilde{h}\breve{v}\frac{t_\epsilon}{2\sqrt{V(P^*; W)}} \\ -\epsilon_n\sqrt{n}\, h\left(J\tilde{h} + \tilde{v}\frac{t_\epsilon}{2\sqrt{V(P^*; W)}}\right) + O(\epsilon_n^2\sqrt{n}). \end{aligned}$$

Proof: By **(A1)** and **(A2)**, the function $\zeta_{n,\delta}$ is twice differentiable at $(P^*, (P^*W))$. The claim follows by Taylor series expansion of the function $\zeta_{n,\delta}$ at that point. □

Now let $\tilde{g}$ and $\breve{g}$ be two functions over $\mathcal{X}$ and consider the game with payoff function

$$\mathcal{E}(h, \tilde{h}) = \frac{1}{2}\tilde{h}J\tilde{h} - \tilde{h}\breve{g} - h[J\tilde{h} + \tilde{g}], \quad h, \tilde{h} \in \mathcal{L}(\mathcal{X}) \quad (33)$$

to be maximized over h and minimized over $\tilde{h}$. The payoff function is linear in h and convex quadratic in $\tilde{h}$. Let $g = \tilde{g} + \breve{g}$.

Lemma 3.3: The game (33) admits the saddlepoint

$$h^* = -J^\dagger g, \quad \tilde{h}^* = -J^\dagger \tilde{g}$$

and its value is $\mathcal{E}^* = \frac{1}{2}\|g\|_{J^\dagger}^2 - \frac{1}{2}\|\breve{g}\|_{J^\dagger}^2$. Moreover the saddlepoint satisfies the equalizer property

$$\mathcal{E}(h, \tilde{h}^*) = \mathcal{E}(h^*, \tilde{h}^*) \le \mathcal{E}(h^*, \tilde{h}) \quad \forall h, \tilde{h} \in \mathcal{L}(\mathcal{X}). \quad (34)$$

The following lemma shows that the payoff function $\zeta_{n,\delta}(P, Q_n)$ is constant (up to a vanishing term) over $P \in \mathscr{P}_1$. The proof follows from Lemmas 3.2 and 3.3.

Lemma 3.4: The game with payoff function $\zeta_{n,\delta}(P, Q)$ with $P \in \mathscr{P}_1 \subset \mathcal{R}_\delta$ admits an asymptotic saddlepoint (P_n^*, Q_n) that satisfies the asymptotic equalizer property

$$\begin{aligned} \zeta_{n,\delta}(P, Q_n) + O\left(\frac{1}{\log^2 n}\right) &= \zeta_{n,\delta}(P_n^*, Q_n) \\ &\le \zeta_{n,\delta}(P_n^*, Q). \quad (35) \end{aligned}$$

The asymptotic value of the game is $\zeta_{n,\delta}(P_n^*, Q_n) = \zeta_{n,\delta}(P^*; W) + \frac{1}{8}t_\epsilon^2 A_{\mathrm{ns}}$ where A_{ns} is defined in (7).

Proof of Theorem 3.1. Denote by $M[\mathscr{P}_1]$ the volume of a subcode with codewords in F_1 and maximum error probability ϵ. Then

$$\begin{aligned} M[\mathscr{P}_1] &\overset{(a)}{\le} \sup_{\mathbf{x}\in\mathsf{F}_1} \frac{1}{\beta_{1-\epsilon}(W^n(\cdot|\mathbf{x}), Q_n^n)} \\ &\overset{(b)}{=} \exp\left\{ \max_{P\in\mathscr{P}_1} \zeta_{n,\delta}(P, Q_n) + \frac{1}{2}\log n + o(1) \right\} \\ &\overset{(c)}{=} \exp\left\{ \zeta_{n,\delta}(P^*; W) + \frac{1}{2}\log n + \frac{t_\epsilon^2}{8}A_{\mathrm{ns}} + o(1) \right\} \end{aligned}$$

where inequality (a) and equalities (b) and (c) follow from Theorem 1.1, and Prop. 2.1 and Lemma 3.4, respectively.

For the codewords not in F_1 we have up to $(n+1)^{|\mathcal{X}|-1}$ types. By Theorem 2.2, the cardinality of each constant-composition subcode with type $P \notin \mathscr{P}_1$ is upper-bounded by

$$\begin{aligned} M[P] &\le \exp\{\zeta_{n,\delta}(P; W) + \frac{1}{2}\log n + o(1)\} \\ &\le \exp\{\zeta_{n,\delta}(P^*; W) - \frac{\sqrt{n}}{\log^2 n} + \frac{1}{2}\log n + o(1)\} \end{aligned}$$

where the last inequality follows from (31). The cardinality of the union of such subcodes is therefore upper bounded by

$$\begin{aligned} M[\mathscr{P}_1^c] &= \sum_{P\notin\mathscr{P}_1} M[P] \le (n+1)^{|\mathcal{X}|-1} \max_{P\notin\mathscr{P}_1} M[P] \\ &\le \exp\left\{ \zeta_{n,\delta}(P^*; W) - \frac{\sqrt{n}}{\log^2 n} + \left(|\mathcal{X}| - \frac{1}{2}\right)\log n + o(1) \right\}. \end{aligned}$$

Finally,

$$\begin{aligned}
M &\leq M[\mathscr{P}_1] + M[\mathscr{P}_1^c] \\
&= \exp\left\{\zeta_{n,\delta}(P^*;W) + \frac{1}{2}\log n + \frac{t_\epsilon^2}{8}A_{\mathrm{ns}} + o(1)\right\} \\
&= \exp\left\{nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\log n + F_\epsilon + \frac{t_\epsilon^2}{8}A_{\mathrm{ns}} + o(1)\right\}
\end{aligned}$$

which proves the claim. □

IV. General Codes, Average Error Probability

Each codeword $\mathbf{x} \in \mathcal{X}^n$ has a type $\hat{P}_{\mathbf{x}} \in \mathscr{P}_n(\mathcal{X})$. For constant-composition codes, $\hat{P}_{\mathbf{x}}$ is the same for all codewords. For a more general code, $\hat{P}_{\mathbf{x}}$ is not fixed but has a nondegenerate empirical distribution π_n over $\mathscr{P}(\mathcal{X})$. That is, $\pi_n(\mathcal{A}) = \frac{1}{M}\sum_{1\leq m\leq M}\mathbb{1}\{\hat{P}_{\mathbf{x}(m)} \in \mathcal{A}\}$ for any collection $\mathcal{A}$ of types. We refer to π_n as the type distribution of the code.

By [4, Prop. 4.4] there is no loss of optimality in restricting the maximization over $P_{\mathbf{X}}$ in Theorem 1.2 to permutation-invariant distributions of the form $P_{\mathbf{X}} = \int_{\mathscr{P}_1}\pi_n(dP)U_{\mathbf{X}|P}$.

Theorem 4.1: The volume $M[\mathscr{P}_1]$ of any code with codewords in F_1 and average error probability ϵ satisfies

$$\log M[\mathscr{P}_1] \leq nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\log n + \overline{A}_\epsilon + o(1).$$

Proof. Fix $Q_{\mathbf{Y}} = Q_n^n$, the n-fold product of $Q_n \in \mathscr{P}(\mathcal{Y})$ defined in (12). We have the asymptotic equalizer property

$$\begin{aligned}
&\forall P \in \mathscr{P}_1 : \ \beta_{1-\epsilon}(U_{\mathbf{X}|P}\times W^n, U_{\mathbf{X}|P}\times Q_n^n) \\
&\overset{(a)}{=} \ \beta_{1-\epsilon}(W^n(\cdot|\mathbf{x}), Q_n^n) \quad \forall \mathbf{x} \ : \ \hat{P}_{\mathbf{x}} = P \\
&\overset{(b)}{=} \ \exp\{-n\zeta_{n,\delta}(P,Q_n) - \frac{1}{2}\log n + o(1)\} \\
&\overset{(c)}{=} \ \exp\{-nC + \sqrt{nV}t_\epsilon - \frac{1}{2}\log n - \overline{A}_\epsilon\}\,[1+o(1)] \\
&\triangleq \ \beta_{1-\epsilon,n}\,[1+o(1)]
\end{aligned}$$

where (a) follows from [2, Lemma 29 p. 2318], (b) from Prop. 2.1, and (c) from Lemma 3.4 and the fact that $P \in \mathscr{P}_1$. Then for any distribution π_n over $\mathscr{P}_1$ we can show using a variation of [2, Lemma 29 p. 2318] that

$$\beta_{1-\epsilon}(P_{\mathbf{X}}\times W^n, P_{\mathbf{X}}\times Q_n^n) = \beta_{1-\epsilon,n}\,[1+o(1)]$$

with $P_{\mathbf{X}} = \int_{\mathscr{P}_1}\pi_n(dP)U_{\mathbf{X}|P}$. Application of Theorem 1.2 proves the claim. □

Theorem 4.2: The volume $M[\mathscr{P}_1^c]$ of any code with codewords in $\mathcal{X}^n \setminus \mathsf{F}_1 = \mathsf{F}(\mathscr{P}_1^c)$ and average error probability ϵ satisfies

$$\log M[\mathscr{P}_1^c] \leq nC - \sqrt{nV}t_\epsilon - \frac{\sqrt{n}}{\log^2 n} + O\left(\frac{\sqrt{n}}{\log^3 n}\right).$$

Sketch of the proof. There are $J < (n+1)^{|\mathcal{X}|-1}$ codeword types in $\mathscr{P}_1^c$. For each such type P_j, $1 \leq j \leq J$, denote by M_j the number of codewords of type P_j and by ϵ_j the average decoding error probability conditioned on the codeword having type P_j. Assume momentarily that $M_j = 0$ for the "bad types" $P_j \notin \mathcal{R}_\delta$. We have $M = \sum_j M_j$ and $\epsilon = \sum_j \epsilon_j \frac{M_j}{M}$. We show there exists j such that $\epsilon_j \leq \epsilon\left[1 + J\exp\left(-\frac{\sqrt{n}}{\log^3 n}\right)\right]$ and $\frac{M_j}{M} \geq \exp\left(-\frac{\sqrt{n}}{\log^3 n}\right)$. Hence

$$\begin{aligned}
\log M &\leq \log M_j + \frac{\sqrt{n}}{\log^3 n} \\
&\overset{(a)}{\leq} nI(P_j;W) - \sqrt{nV(P_j;W)}\mathcal{Q}^{-1}(\epsilon_j) + O\left(\frac{\sqrt{n}}{\log^3 n}\right) \\
&\overset{(b)}{\leq} nI(P_j;W) - \sqrt{nV(P_j;W)}\mathcal{Q}^{-1}(\epsilon) + O\left(\frac{\sqrt{n}}{\log^3 n}\right) \\
&\overset{(c)}{\leq} nC - \sqrt{nV}\mathcal{Q}^{-1}(\epsilon) - \frac{\sqrt{n}}{\log^2 n} + O\left(\frac{\sqrt{n}}{\log^3 n}\right)
\end{aligned}$$

where (a) follows from Theorem 2.2, (b) from the upper bound on ϵ_j and the fact that the function $-\mathcal{Q}^{-1}(\epsilon)$ is increasing, and (c) from the fact that the type $P_j \notin \mathscr{P}_1$. The same upper bound (c) can be shown to hold if $M_j > 0$ for bad types $P_j \notin \mathcal{R}_\delta$. □

Theorem 4.3: The volume M of any code with codewords in $\mathcal{X}^n$ and average error probability ϵ satisfies

$$\log M \leq nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\log n + \overline{A}_\epsilon + o(1).$$

Sketch of the proof. Any (M,ϵ) code is the union of a (M_1,ϵ_1) subcode with codewords in F_1 and a (M_2,ϵ_2) subcode with codewords in $\mathcal{X}^n \setminus \mathsf{F}_1$ where $M = M_1 + M_2$ and $\epsilon = \epsilon_1\frac{M_1}{M_1+M_2} + \epsilon_2\frac{M_2}{M_1+M_2}$. Theorems 4.1 and 4.2 yield

$$\log M_1 \leq nC - \sqrt{nV}\mathcal{Q}^{-1}(\epsilon_1) + \frac{1}{2}\log n + \overline{A}(\epsilon_1) + o(1) \tag{36}$$

$$\log M_2 \leq nC - \sqrt{nV}\mathcal{Q}^{-1}(\epsilon_2) - \frac{\sqrt{n}}{\log^2 n}[1+o(1)]. \tag{37}$$

Let $q_1 = \frac{M_1}{M_1+M_2} \in (0,1)$. The claim is shown by using the identity $M = \min\{\frac{M_1}{q_1}, \frac{M_2}{1-q_1}\}$, applying the upper bounds (36) and (37), and optimizing over q_1, ϵ_1. □

Acknowledgment

This work was supported by DARPA under the ITMANET program and by NSF under grant CCF 12-19145.

References

[1] V. Strassen, "Asymptotische Abschätzungen in Shannon's Informationstheorie," *Trans. 3rd Prague Conf. Info. Theory*, pp. 689—723, 1962. English translation by M. Luthy available from `http://www.math.cornell.edu/~pmlut/strassen.pdf`.

[2] Y. Polyanskiy, H. V. Poor and S. Verdú, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE T-IT*, Vol. 56, pp. 2307—2359, 2010.

[3] M. Hayashi, "Information Spectrum Approach to Second-Order Coding Rate in Channel Coding," *IEEE T-IT*, Vol. 55, pp. 4947—4966, Nov. 2009.

[4] P. Moulin, "The Log-Volume of Optimal Codes for Memoryless Channels, Up to a Few Nats," *Proc. ITA Conference*, San Diego, CA, Feb. 2012.

[5] P. Moulin, "The Log-Volume of Optimal Codes for Memoryless Channels, Up to a Few Nats," to be posted on arxiv, June 2013.

[6] Y. Polyanskiy, "Saddle Point in the Minimax Converse for Channel Coding," *IEEE T-IT*, Vol. 59, pp. 2576—2595, May 2013.