Typicality of a Good Rate-Distortion Code

Angelos Kanlis Sanjeev Khudanpur Prakash Narayan *

This paper is dedicated to Mark Semenovich Pinsker
on the occasion of his seventieth birthday.

Abstract

We consider a good code for a discrete memoryless source with a specified distortion level to be one whose rate is close to the corresponding rate-distortion function and which, with large probability, reproduces the source within the allowed distortion level. We show that any good code must contain an exponentially large set of codewords, of effectively the same rate, which are all typical with respect to the output distribution induced by the rate-distortion achieving channel. Furthermore, the output distribution induced by a good code is asymptotically singular with respect to the i.i.d. output distribution induced by the rate-distortion achieving channel. However, the normalized (Kullback-Leibler) divergence between these output distributions converges to the conditional entropy of the output under the rate-distortion achieving channel.

1 Introduction

A good code for a discrete memoryless source (DMS) with a specified distortion level is one whose rate is close to the corresponding rate-distortion function and which, with large probability, reproduces the source within the allowed distortion level. The Covering Lemma of Rate-Distortion Theory (cf.,*e.g.*, [1], Lemma 4.1, p. 150) asserts the existence of a good code, all of whose codewords are typical with respect to the *optimal* distribution on the reproduction alphabet, namely that induced by the rate-distortion achieving channel. We show that *any* good code must contain an exponentially large set

*The authors are with the Department of Electrical Engineering and the Institute for Systems Research at the University of Maryland, College Park, MD 20742, U.S.A. This work of was supported by the Institute for Systems Research at the University of Maryland under NSF Grant OIR-85-00108. The work of A. Kanlis was additionally supported by an IBM Graduate Fellowship.

of codewords which are *all* typical (in the sense as above), and is of effectively the same rate. Next, this source code gives rise to a deterministic channel between the spaces of all the input and output sequences, which in turn induces an output distribution on the latter. Since the output distribution corresponding to this code restricts its mass to the set of codewords, it differs significantly from the i.i.d. (independent and identically distributed) measure induced by the optimal output distribution. It is shown that these two measures are asymptotically singular except in trivial cases. Furthermore, under an additional assumption of the "minimality" of a good code, the normalized (Kullback-Leibler) divergence between these output distributions converges to the conditional entropy of the output under the rate-distortion achieving channel. As is to be expected, this behavior of a good source code is in contrast with that of a good channel code for which, as shown by Han-Verdú [2], the normalized divergence between the corresponding output distribution and the optimal output distribution (induced by the capacity achieving input distribution) vanishes asymptotically.

2 Preliminaries and Main Results

We have adopted the terminology of Csiszár-Körner [1]. In particular, all logarithms and exponentials are taken with respect to the base 2.

Let $\{X_t\}_{t=1}^{\infty}$ be a discrete memoryless source (DMS) with finite alphabet $\mathcal{X}$, *i.e.*, an independent and identically distributed (i.i.d.) process, with common probability mass function (pmf) P, where $P(x) > 0,\ x \in \mathcal{X}$. Let $\mathcal{Y}$ be a finite reproduction alphabet. Let $d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ be a nonnegative-valued mapping with $\min_{x\in\mathcal{X},\ y\in\mathcal{Y}} d(x,y) = 0$ and $d_{max} = \max_{x\in\mathcal{X},\ y\in\mathcal{Y}} d(x,y) < \infty$. This mapping induces a distortion measure on $\mathcal{X}^n \times \mathcal{Y}^n$ according to

$$d(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{t=1}^{n} d(x_t, y_t), \quad \mathbf{x} \in \mathcal{X}^n,\ \mathbf{y} \in \mathcal{Y}^n.$$

A n-length block code consists of two mappings: An encoder $f_n : \mathcal{X}^n \to \{1,\ldots,M_n\}$ and a decoder $\phi_n : \{1,\ldots,M_n\} \to \mathcal{Y}^n$. The rate of this code is $R_n = \frac{1}{n}\log M_n$.

For $\Delta > 0$, the rate distortion function, $R(P,\Delta)$, characterizing the minimum achievable rate for a distortion Δ, is well known and given by

$$R(P,\Delta) = \min_{W:d(P,W)\le\Delta} I(P,W) \tag{1}$$

where W ranges over all stochastic matrices $W : \mathcal{X} \to \mathcal{Y}$, and

$$d(P,W) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x)W(y|x)d(x,y). \tag{2}$$

Here, $I(P,W)$ denotes the mutual information between the random variables (X,Y) on $\mathcal{X}\times\mathcal{Y}$ with pmf $P_{XY}(x,y)=P(x)W(y|x),\ x\in\mathcal{X},\ y\in\mathcal{Y}$.

It will be assumed throughout that the minimizing W in (1)*, denoted W^*, is unique.*

We shall hereafter be interested only in a "good" code, *i.e.*, a code which with large probability reproduces sequences from $\mathcal{X}^n$ with distortion no greater than a specified $\Delta>0$ and has a rate arbitrarily close to $R(P,\Delta)$. Precisely, a *good code* is defined as a sequence of n-length block codes (f_n,ϕ_n) with the following two properties: For every $\epsilon>0$, $\alpha>0$, and n sufficiently large (depending on ϵ,α),

$$P^n\left(\mathbf{x}\in\mathcal{X}^n\ :\ d(\mathbf{x},\phi_n(f_n(\mathbf{x})))\le\Delta\right)\ge 1-\epsilon \tag{3}$$

where

$$P^n(\mathbf{x})=\prod_{t=1}^{n}P(x_t),\quad \mathbf{x}\in\mathcal{X}^n, \tag{4}$$

and

$$R(P,\Delta)<R_n\le R(P,\Delta)+\alpha. \tag{5}$$

We shall use $\mathcal{C}_n=\{\mathbf{y}_i\in\mathcal{Y}^n:\mathbf{y}_i=\phi_n(i),\ 1,\ldots,M_n\}$ to denote the codewords of the n-length code (f_n,ϕ_n). These codewords induce a partition $\{\mathcal{A}_i\}_{i=1}^{M_n}$ of $\mathcal{X}^n$, where

$$\mathcal{A}_i=\{\mathbf{x}\in\mathcal{X}^n\ :\ \phi_n(f_n(\mathbf{x}))=\mathbf{y}_i\},\quad i=1,\ldots,M_n. \tag{6}$$

Of particular interest is the family of (disjoint) subsets $\{\mathcal{B}_i\}_{i=1}^{M_n}$ of $\mathcal{X}^n$ defined by

$$\mathcal{B}_i=\{\mathbf{x}\in\mathcal{A}_i\ :\ d(\mathbf{x},\mathbf{y}_i)\le\Delta\},\quad i=1,\ldots,M_n. \tag{7}$$

By virtue of property (3) of a good code, and the fact that $\{\mathcal{A}_i\}_{i=1}^{M_n}$ is a partition of $\mathcal{X}^n$, note that for every $\epsilon>0$,

$$P^n\left(\bigcup_{i=1}^{M_n}\mathcal{B}_i\right)\ge 1-\epsilon \tag{8}$$

for all n sufficiently large (depending on ϵ).

We recall from [1] that the *type* of $\mathbf{x}\in\mathcal{X}^n$ is a pmf $Q_{\mathbf{x}}$ on $\mathcal{X}$ where $Q_{\mathbf{x}}(x)$ is the relative frequency of the symbol x in $\mathbf{x}$. For any type Q on $\mathcal{X}$, let $T_Q^{(n)}$ denote the set of all sequences $\mathbf{x}\in\mathcal{X}^n$ with $Q_{\mathbf{x}}=Q$. (A similar notation will be used for types on $\mathcal{Y}$.) For any pmf Q on $\mathcal{X}$, pmf S on $\mathcal{Y}$, and $\Delta>0$, define

$$\tilde{R}(Q,S,\Delta)=\min_{W:d(Q,W)\le\Delta,\ Q\cdot W=S}I(Q,W) \tag{9}$$

where W ranges over stochastic matrices $W:\mathcal{X}\to\mathcal{Y}$ and $Q\cdot W$ denotes a pmf on $\mathcal{Y}$ defined by

$$Q\cdot W(y)=\sum_{x\in\mathcal{X}}Q(x)W(y|x),\quad y\in\mathcal{Y}. \tag{10}$$

Note that $\tilde{R}(Q,S,\Delta)$ is well-defined since $I(Q,W)$ is a convex function of W and $\{W : d(Q,W) \le \Delta,\ Q \cdot W = S\}$ is a convex compact set.

We present below a technical result needed to prove our main result in the subsequent Proposition 1. A refined version of this result can be found in the earlier work of Zhang-Yang-Wei [3, Lemma 3]. Let $|\cdot|$ denote cardinality of sets and $H(\cdot)$ denote entropy.

Lemma: Let Q be any type on $\mathcal{X}$. Then, for $i = 1, \ldots, M_n$, it holds that

$$\left|\mathcal{B}_i \bigcap T_Q^{(n)}\right| \le (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left[n\left(H(Q) - \tilde{R}(Q, Q_{\mathbf{y}_i}, \Delta)\right)\right]. \tag{11}$$

Proof: For each conditional type V on $\mathcal{X}$ given $\mathbf{y}_i \in \mathcal{Y}^n$, let $T_V^{(n)}(\mathbf{y}_i)$ denote the V-shell of $\mathbf{y}_i$ (cf., *e.g.*, [1]). Then for $i = 1, \ldots, M_n$,

$$\mathcal{B}_i \bigcap T_Q^{(n)} = \bigcup_{V:\ d(Q_{\mathbf{y}_i},V)\le\Delta,\ Q_{\mathbf{y}_i}\cdot V = Q} \left(\mathcal{B}_i \bigcap T_V^{(n)}(\mathbf{y}_i)\right)$$

where $d(Q_{\mathbf{y}_i}, V) = \sum_{x\in\mathcal{X},\ y\in\mathcal{Y}} Q_{\mathbf{y}_i}(y)V(x|y)d(x,y)$ and $Q_{\mathbf{y}_i} \cdot V$ denotes a pmf on $\mathcal{X}$ given by

$$Q_{\mathbf{y}_i} \cdot V(x) = \sum_{y\in\mathcal{Y}} Q_{\mathbf{y}_i}(y)V(x|y), \quad x \in \mathcal{X}. \tag{12}$$

Consequently, using standard bounds on types (cf., *e.g.*, [1]) we obtain

$$\begin{aligned}
&\left|\mathcal{B}_i \bigcap T_Q^{(n)}\right| \\
&= \sum_{V:\ d(Q_{\mathbf{y}_i},V)\le\Delta,\ Q_{\mathbf{y}_i}\cdot V = Q} \left|\mathcal{B}_i \bigcap T_V^{(n)}(\mathbf{y}_i)\right| \\
&\le \sum_{V:\ d(Q_{\mathbf{y}_i},V)\le\Delta,\ Q_{\mathbf{y}_i}\cdot V = Q} \left|T_V^{(n)}(\mathbf{y}_i)\right| \\
&\le (n+1)^{|\mathcal{X}||\mathcal{Y}|} \max_{V:\ d(Q_{\mathbf{y}_i},V)\le\Delta,\ Q_{\mathbf{y}_i}\cdot V = Q} \left|T_V^{(n)}(\mathbf{y}_i)\right| \\
&\le (n+1)^{|\mathcal{X}||\mathcal{Y}|} \max_{V:\ d(Q_{\mathbf{y}_i},V)\le\Delta,\ Q_{\mathbf{y}_i}\cdot V = Q} \exp\left[nH(V|Q_{\mathbf{y}_i})\right] \\
&= (n+1)^{|\mathcal{X}||\mathcal{Y}|} \max_{V:\ d(Q_{\mathbf{y}_i},V)\le\Delta,\ Q_{\mathbf{y}_i}\cdot V = Q} \exp\left[n\left(H(Q_{\mathbf{y}_i} \cdot V) - I(Q_{\mathbf{y}_i}, V)\right)\right] \\
&= (n+1)^{|\mathcal{X}||\mathcal{Y}|} \max_{W:\ d(Q,W)\le\Delta,\ Q\cdot W = Q_{\mathbf{y}_i}} \exp\left[n\left(H(Q) - I(Q,W)\right)\right]
\end{aligned}$$

from which the assertion of the Lemma follows. ■

At this juncture it is useful to recall two distance measures for pmf's on $\mathcal{X}$ (or $\mathcal{Y}$). The *Kullback-Leibler divergence* between the pmf's P and Q on $\mathcal{X}$ is defined as:

$$D(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

where the usual convention applies that $0 \log 0 = 0$, $0 \log \left(\frac{0}{0}\right) = 0$ and $t \log \left(\frac{t}{0}\right) = +\infty$, $t > 0$. The *variational distance* between P and Q is defined as

$$\|P - Q\| = \sum_{x \in \mathcal{X}} |P(x) - Q(x)|.$$

Proposition 1 below states that of the roughly $\exp[nR(P, \Delta)]$ codewords of a good code, an exponentially large subset, of effectively the same rate, has codewords whose types asymptotically approach $P \cdot W^*$ in variational distance. Clearly, not all codewords of a good code need possess this property.

Let $\mathcal{C}_n^T$ denote the subset of codewords which are $P \cdot W^*$-typical. Precisely, given $\gamma > 0$, define $\mathcal{C}_n^T = \mathcal{C}_n^T(\gamma)$ by

$$\mathcal{C}_n^T = \{\mathbf{y} \in \mathcal{C}_n : \|Q_{\mathbf{y}} - P \cdot W^*\| \leq \gamma\}. \tag{13}$$

Proposition 1: For every $\gamma > 0$

$$\lim_{n \to \infty} \frac{1}{n} \log \left| \mathcal{C}_n^T(\gamma) \right| = R(P, \Delta). \tag{14}$$

Proof: For any $\delta > 0$ and $\zeta > 0$, it holds for all n sufficiently large (depending on δ, ζ) that

$$P^n \left(T_{[P]}^{(n)}\right) \geq 1 - \zeta \tag{15}$$

where

$$T_{[P]}^{(n)} = \bigcup_{Q : \|P - Q\| \leq \delta} T_Q^{(n)} \tag{16}$$

with Q denoting types on $\mathcal{X}$. A suitable choice of δ and ζ will be made later.

In conjunction with (8), and using standard bounds for types, we get for all n sufficiently large (depending on ϵ, δ, ζ) that

$$\begin{aligned}
1 - \epsilon - \zeta &\leq P^n \left(T_{[P]}^{(n)} \bigcap \left(\bigcup_{i=1}^{M_n} \mathcal{B}_i\right)\right) \\
&= \sum_{i=1}^{M_n} P^n \left(T_{[P]}^{(n)} \bigcap \mathcal{B}_i\right) \\
&= \sum_{i=1}^{M_n} \sum_{Q : \|P - Q\| \leq \delta} P^n \left(T_Q^{(n)} \bigcap \mathcal{B}_i\right) \\
&= \sum_{i=1}^{M_n} \sum_{Q : \|P - Q\| \leq \delta} \left|T_Q^{(n)} \bigcap \mathcal{B}_i\right| \times \exp\left[-n \left(H(Q) + D(Q\|P)\right)\right] \\
&\leq \sum_{i=1}^{M_n} \sum_{Q : \|P - Q\| \leq \delta} (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left[n \left(H(Q) - \tilde{R}(Q, Q_{\mathbf{y}_i}, \Delta)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
& \times \exp\left[-n\left(H(Q)\right)\right], \quad \text{by (11)} \\
= & \sum_{i=1}^{M_n} \sum_{Q:\|P-Q\|\leq\delta} (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left[-n\left(\tilde{R}(Q, Q_{\mathbf{y}_i}, \Delta)\right)\right] \\
\leq & \sum_{i=1}^{M_n} (n+1)^{|\mathcal{X}|(|\mathcal{Y}|+1)} \exp\left[-n\left(\min_{Q:\|P-Q\|\leq\delta} \tilde{R}(Q, Q_{\mathbf{y}_i}, \Delta)\right)\right]. \qquad (17)
\end{aligned}
$$

Given any $\eta > 0$, observe by the continuity of $\tilde{R}(Q, Q_{\mathbf{y}_i}, \Delta)$ in Q that $\delta > 0$ can be chosen to satisfy

$$
\min_{Q:\|P-Q\|\leq\delta(\eta)} \tilde{R}(Q, Q_{\mathbf{y}_i}, \Delta) \geq \tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) - \eta. \qquad (18)
$$

Continuing the bounding in (17), we obtain

$$
1 - \epsilon - \zeta \leq (n+1)^{|\mathcal{X}|(|\mathcal{Y}|+1)} \sum_{i=1}^{M_n} \exp\left[-n\left(\tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) - \eta\right)\right]. \qquad (19)
$$

Consider first the sum in (19) over those indices i for which $\mathbf{y}_i \in \mathcal{C}_n^T$. Note that $\tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) \geq R(P, \Delta)$ and therefore

$$
\sum_{i:\mathbf{y}_i\in\mathcal{C}_n^T} \exp\left[-n\left(\tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) - \eta\right)\right] \leq |\ \mathcal{C}_n^T\ | \exp\left[-n\left(R(P, \Delta) - \eta\right)\right]. \qquad (20)
$$

Next, the sum in (19) over those indices i for which $\mathbf{y}_i \notin \mathcal{C}_n^T$ can be bounded above as follows:

$$
\begin{aligned}
& \sum_{i:\mathbf{y}_i\notin\mathcal{C}_n^T} \exp\left[-n\left(\tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) - \eta\right)\right] \\
& \leq \exp\left[n\left(R(P, \Delta) + \alpha\right)\right] \exp\left[-n \min_{i:\mathbf{y}_i\notin\mathcal{C}_n^T} \left(\tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) - \eta\right)\right], \quad \text{by (5)} \\
& = \exp\left[-n\left(\min_{i:\mathbf{y}_i\notin\mathcal{C}_n^T} \tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) - R(P, \Delta) - \alpha - \eta\right)\right] \qquad (21)
\end{aligned}
$$

Next, observe that $\tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) \geq R(P, \Delta)$ with equality iff $Q_{\mathbf{y}_i} = P \cdot W^*$. Together with the condition $\|Q_{\mathbf{y}_i} - P \cdot W^*\| > \gamma$ and the assumed uniqueness of W^*, there exists $\theta = \theta(\gamma) > 0$ such that

$$
\min_{i:\mathbf{y}_i\notin\mathcal{C}_n^T} \tilde{R}(P, Q_{\mathbf{y}_i}, \Delta) - R(P, \Delta) > \theta. \qquad (22)
$$

Therefore, given $\gamma > 0$, we can pick $\alpha = \alpha(\gamma)$, $\eta(\gamma)$ such that $\alpha + \eta < \frac{\theta}{3}$. Then the right side in (21) is bounded above by $\exp\left[-n\frac{2\theta}{3}\right]$ for all n sufficiently large (depending on

γ). Using this in (19), we obtain

$$1-\epsilon-\zeta \;\leq\; (n+1)^{|\mathcal{X}|(|\mathcal{Y}|+1)}\left|\mathcal{C}_n^T\right|\exp\left[-n\left(R(P,\Delta)-\eta\right)\right]+\exp\left[-n\frac{\theta}{3}\right]$$

which can be rearranged as

$$R(P,\Delta)-\xi \leq \frac{1}{n}\log\left|\;\mathcal{C}_n^T\;\right| \tag{23}$$

for all n sufficiently large (depending on γ), where

$$\xi = |\mathcal{X}|(|\mathcal{Y}|+1)\frac{\log(n+1)}{n}+\eta+\frac{1}{n}\log\left(1-\epsilon-\zeta-\exp\left(-n\frac{\theta}{3}\right)\right).$$

Observe that ξ can be made arbitrarily small, for all n sufficiently large (depending on ξ,γ,ζ) by choosing η to be, say, the smaller of $\frac{\xi}{2}$ and $\frac{\theta(\gamma)}{4}$. Combining this with (5) yields the assertion of the Proposition. ■

The sequence of n-length block codes (f_n,ϕ_n) constituting a good code induces a (deterministic) channel $\{\tilde{W}^{(n)}:\mathcal{X}^n\to\mathcal{Y}^n\}_{n=1}^{\infty}$, given by

$$\tilde{W}^{(n)}(\mathbf{y}|\mathbf{x}) = 1(\mathbf{y}=\phi_n(f_n(\mathbf{x}))),\quad \mathbf{x}\in\mathcal{X}^n,\ \mathbf{y}\in\mathcal{Y}^n, \tag{24}$$

where $1(\cdot)$ denotes the indicator function. As a consequence of Proposition 1, we shall characterize how "distant" the sequence of pmf's $P^n\cdot\tilde{W}^{(n)}$ on $\mathcal{Y}^n$, induced by a good code, can be from the sequence of pmf's $(P\cdot W^*)^n$ on $\mathcal{Y}^n$, induced by the rate distortion achieving stochastic matrix W^*. We note that the pmf's under consideration are given by

$$\begin{aligned} P^n\cdot\tilde{W}^{(n)}(\mathbf{y}) &= \sum_{\mathbf{x}\in\mathcal{X}^n}P^n(\mathbf{x})\tilde{W}^{(n)}(\mathbf{y}|\mathbf{x}) \\ &= \begin{cases} P^n(\mathcal{A}_i) & \text{if } \mathbf{y}=\mathbf{y}_i \text{ for some } i\in\{1,\ldots,M_n\} \\ 0 & \text{otherwise}\end{cases} \end{aligned} \tag{25}$$

and

$$(P\cdot W^*)^n(\mathbf{y}) = \prod_{t=1}^{n} P\cdot W^*(y_t). \tag{26}$$

Clearly, if W^* is $\{0,1\}$-valued, then a good code can be chosen accordingly, *i.e.*, $\tilde{W}^{(n)}=(W^*)^n$ for all n. On the other hand, if W^* is not $\{0,1\}$-valued, then the sequence of pmf's $P^n\cdot\tilde{W}^{(n)}$ and $(P\cdot W^*)^n$ on $\mathcal{Y}^n$ asymptotically become mutually singular. To see this, note that

$$P^n\cdot\tilde{W}^{(n)}(\mathcal{C}_n)=1.$$

Also,

$$(P\cdot W^*)^n(\mathcal{C}_n) = (P\cdot W^*)^n(\mathcal{C}_n^T)+(P\cdot W^*)^n(\mathcal{C}_n\setminus\mathcal{C}_n^T).$$

The first term, by (13), is bounded above by a quantity that is of the order of

$$|\mathcal{C}_n| \exp[-nH(P \cdot W^*)] \approx \exp[-nH(W^*|P)],$$

which decays to zero iff $H(W^*|P) > 0$. The second term is the $(P \cdot W^*)^n$-probability of codewords which are not $P \cdot W^*$-typical and, hence, can be made arbitrarily small by choosing n sufficiently large. Consequently,

$$\lim_{n\to\infty} \|P^n \cdot \tilde{W}^{(n)} - (P \cdot W^*)^n\| = 2 \tag{27}$$

and

$$\lim_{n\to\infty} D(P^n \cdot \tilde{W}^{(n)} \| (P \cdot W^*)^n) = \infty. \tag{28}$$

In the next section, we investigate the behavior of the normalized version of the divergence in (28).

At this juncture it is interesting to recall the following result of Han-Verdú ([2], Theorem 5) for the channel coding problem: The normalized divergence between the output distribution induced by a good channel code (*i.e.*, a code with rate close to channel capacity and arbitrarily small probability of error) and the optimal output distribution (induced by the capacity achieving input distribution) vanishes asymptotically. Not surprisingly, a similar result does not hold for a good source code, as will be seen in Proposition 2 below.

3 A Minimal Good Code

The assertion of Proposition 1 can be rephrased as follows: Given $\xi > 0$, $\gamma > 0$, and for all n sufficiently large (depending on ξ, γ), the set $\mathcal{C}_n^T$, as defined in (13), satisfies

$$\left| \mathcal{C}_n^T \right| \geq \exp\left[n \left(R(P, \Delta) - \xi\right)\right], \tag{29}$$

i.e., $\mathcal{C}_n^T$ is an exponentially large subset of $\mathcal{C}_n$, of effectively the same rate, whose codewords are $P \cdot W^*$-typical. This does not, however, exclude the possibility of $\mathcal{C}_n \setminus \mathcal{C}_n^T$ being exponentially large with the same rate. But we know, from the Covering Lemma (cf.,*e.g.*, [1], Lemma 4.1, p. 150) that there exists a good code, all of whose codewords are $P \cdot W^*$-typical. This motivates us to define a *minimal good code* as a good code with the following additional property: There exists a fixed $\omega > 0$ such that

$$\left|\mathcal{C}_n \setminus \mathcal{C}_n^T(\gamma)\right| \leq \exp\left[n \left(R(P, \Delta) - \omega\right)\right] \tag{30}$$

for all n sufficiently large (depending on ω, γ). Consequently, given any $\kappa > 0$, a minimal good code satisfies

$$P^n \left(\bigcup_{i \,:\, \mathbf{y}_i \in \mathcal{C}_n \setminus \mathcal{C}_n^T} \mathcal{A}_i \right) \leq \kappa \tag{31}$$

for all n sufficiently large (depending on κ, ω). The proof of this fact is relegated to the Appendix. In the following proposition, we show for a minimal good code that the normalized (Kullback-Leibler) divergence has a finite limit.

Proposition 2: For any minimal good code, it holds that

$$\liminf_{n\to\infty} \frac{1}{n} D(P^n \cdot \tilde{W}^{(n)} \| (P \cdot W^*)^n) \geq H(W^*|P). \tag{32}$$

Furthermore, if W^* is such that for every $y \in \mathcal{Y}$, there is an $x \in \mathcal{X}$ with $W^*(y|x) > 0$, then

$$\lim_{n\to\infty} \frac{1}{n} D(P^n \cdot \tilde{W}^{(n)} \| (P \cdot W^*)^n) = H(W^*|P). \tag{33}$$

Proof: First observe that the normalized divergence may be written as

$$\begin{aligned}
&\frac{1}{n} D(P^n \cdot \tilde{W}^{(n)} \| (P \cdot W^*)^n) \\
&\quad = \quad -\frac{1}{n} H(P^n \cdot \tilde{W}^{(n)}) \\
&\qquad +\frac{1}{n} \sum_{i:\mathbf{y}_i \in \mathcal{C}_n^T} P^n(\mathcal{A}_i) \log \frac{1}{(P \cdot W^*)^n(\mathbf{y}_i)} \\
&\qquad +\frac{1}{n} \sum_{i:\mathbf{y}_i \in \mathcal{C}_n \setminus \mathcal{C}_n^T} P^n(\mathcal{A}_i) \log \frac{1}{(P \cdot W^*)^n(\mathbf{y}_i)}.
\end{aligned} \tag{34}$$

Since $P^n \cdot \tilde{W}^{(n)}$ assigns full mass to a set of M_n codewords and since, by (5), $\frac{1}{n} \log M_n \leq R(P, \Delta) + \alpha$, it follows that

$$H(P^n \cdot \tilde{W}^{(n)}) \leq n(R(P, \Delta) + \alpha).$$

Also, the third term on the right hand side of (34) is nonnegative, so that

$$\begin{aligned}
&\frac{1}{n} D(P^n \cdot \tilde{W}^{(n)} \| (P \cdot W^*)^n) \\
&\quad \geq \quad -R(P, \Delta) - \alpha + \frac{1}{n} \sum_{i:\mathbf{y}_i \in \mathcal{C}_n^T} P^n(\mathcal{A}_i) \log \frac{1}{\exp\left[-n\left(H(P \cdot W^*) - \lambda(\gamma)\right)\right]} \\
&\quad = \quad -R(P, \Delta) - \alpha + \left(H(P \cdot W^*) - \lambda(\gamma)\right) \sum_{i:\mathbf{y}_i \in \mathcal{C}_n^T} P^n(\mathcal{A}_i) \\
&\quad \geq \quad -R(P, \Delta) - \alpha + \left(H(P \cdot W^*) - \lambda(\gamma)\right)(1 - \kappa), \quad \text{by (31)} \\
&\quad \geq \quad H(W^*|P) - \mu
\end{aligned} \tag{35}$$

for every $\mu > 0$ and for all n sufficiently large (depending on μ, ω). This establishes the inequality in (32).

To show the second part of the Proposition, note first that

$$d(P^n, \tilde{W}^{(n)}) \quad = \quad \sum_{i=1}^{M_n} \sum_{\mathbf{x} \in \mathcal{B}_i} P^n(\mathbf{x}) d(\mathbf{x}, \mathbf{y}_i) \quad + \quad \sum_{\mathbf{x} \notin \left(\bigcup_{i=1}^{M_n} \mathcal{B}_i\right)} P^n(\mathbf{x}) d(\mathbf{x}, \phi_n(f_n(\mathbf{x})))$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{M_n} \sum_{\mathbf{x}\in\mathcal{B}_i} P^n(\mathbf{x})\Delta \ + \sum_{\mathbf{x}\notin\left(\bigcup_{i=1}^{M_n}\mathcal{B}_i\right)} P^n(\mathbf{x}) d_{max} \\
&= P^n\left(\bigcup_{i=1}^{M_n}\mathcal{B}_i\right)\Delta \ + \ \left[1 - \left(\bigcup_{i=1}^{M_n}\mathcal{B}_i\right)\right] d_{max} \\
&\leq \Delta + \epsilon d_{max}.
\end{aligned}
$$

Hence

$$
\begin{aligned}
H(P^n \cdot \tilde{W}^{(n)}) &= I(P^n, \tilde{W}^{(n)}) \\
&\geq \min_{W^{(n)}: d(P^n, W^{(n)}) \leq \Delta + \epsilon d_{max}} I(P^n, W^{(n)}) \\
&= nR(P, \Delta + \epsilon d_{max}) \\
&\geq n\left(R(P,\Delta) - \beta\right) \qquad (36)
\end{aligned}
$$

for all n sufficiently large (depending on β), where the last inequality uses the continuity of $R(P,\Delta)$ in Δ. Therefore, the first term in (34) is bounded above by $-R(P,\Delta)+\beta$. The second term can easily be seen to be bounded above by $H(P\cdot W^*)+\lambda(\gamma)$ similarly as above. Finally, since

$$
(P\cdot W^*)^n(\mathbf{y}) \geq \left(\min_{y\in\mathcal{Y}} P\cdot W^*(y)\right)^n > 0
$$

the third term on the right hand side of (34) is bounded above by

$$
\kappa \log\left(\frac{1}{\min_{y\in\mathcal{Y}} P\cdot W^*(y)}\right)
$$

and hence for any $\mu > 0$,

$$
\frac{1}{n} D(P^n \cdot \tilde{W}^{(n)} \| (P\cdot W^*)^n) \leq H(W^*|P) + \mu
$$

for all n sufficiently large (depending on μ, ω). Hence,

$$
\lim_{n\to\infty} \sup \frac{1}{n} D(P^n \cdot \tilde{W}^{(n)} \| (P\cdot W^*)^n) \leq H(W^*|P),
$$

completing the proof of the Proposition. ■

We remark that for a good code which is not minimal in the sense of (30), it is not difficult to see that the limiting value of the normalized divergence in (34) can be bounded away from $H(W^*|P)$.

4 Appendix

Proof of inequality (31): Since the pmf $P^n \cdot \tilde{W}^{(n)}$ on $\mathcal{Y}^n$ assigns full mass to $\mathcal{C}_n$ and has entropy arbitrarily close to the rate of $\mathcal{C}_n$ (see (36)), it follows from the continuity of the entropy function that $\mathcal{C}_n$ contains a subset, $\mathcal{C}_n^U$, of large $P^n \cdot \tilde{W}^{(n)}$-probability with roughly equiprobable codewords. Precisely, any $\upsilon > 0$, $\nu > 0$, there exists a subset $\mathcal{C}_n^U \subseteq \mathcal{C}_n$ given by

$$\mathcal{C}_n^U = \left\{ \mathbf{y} \in \mathcal{C}_n : \exp\left[-n\left(R(P,\Delta)+\nu\right)\right] \leq P^n \cdot \tilde{W}^{(n)}(\mathbf{y}) \leq \exp\left[-n\left(R(P,\Delta)-\nu\right)\right] \right\} \tag{37}$$

with

$$P^n \cdot \tilde{W}_{(n)}(\mathcal{C}_n^U) > 1 - \upsilon \tag{38}$$

for all n sufficiently large (depending on υ, ν). Consequently,

$$\begin{aligned}
&\sum_{i\,:\,\mathbf{y}_i \in \mathcal{C}_n \setminus \mathcal{C}_n^T} P^n(\mathcal{A}_i) \\
&= \sum_{\mathbf{y}_i \in \mathcal{C}_n \setminus \mathcal{C}_n^T} P^n \cdot \tilde{W}^{(n)}(\mathbf{y}_i) \\
&= \sum_{\mathbf{y}_i \in (\mathcal{C}_n \setminus \mathcal{C}_n^T) \bigcap \mathcal{C}_n^U} P^n \cdot \tilde{W}^{(n)}(\mathbf{y}_i) + \sum_{\mathbf{y}_i \in (\mathcal{C}_n \setminus \mathcal{C}_n^T) \bigcap (\mathcal{C}_n \setminus \mathcal{C}_n^U)} P^n \cdot \tilde{W}^{(n)}(\mathbf{y}_i) \\
&\leq \mid \mathcal{C}_n \setminus \mathcal{C}_n^T \mid \times \exp\left[-n\left(R(P,\Delta)-\nu\right)\right] + P^n \cdot \tilde{W}^{(n)}(\mathcal{C}_n \setminus \mathcal{C}_n^U) \\
&\leq \exp\left[n\left(R(P,\Delta)-\omega\right)\right] \exp\left[-n\left(R(P,\Delta)-\nu\right)\right] + \upsilon \\
&= \exp\left[-n\left(\omega-\nu\right)\right] + \upsilon
\end{aligned} \tag{39}$$

which can be made smaller than any $\kappa > 0$ for all n sufficiently large (depending on κ, ω), since $\omega > 0$, by choosing, say, $\nu = \frac{\omega}{2}$ and $\upsilon = \frac{\kappa}{2}$.

References

[1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Akadémiai Kiadó, Budapest, 1986.

[2] T. S. Han and S. Verdú, "Approximation Theory of Output Statistics," *IEEE Transactions on Information Theory*, vol. IT-39, no. 3, pp. 752-772, May 1993.

[3] Z. Zhang, E. Yang and V. W. Wei, "The Redundancy of Source Coding with a Fidelity Criterion," *preprint*, June 1994.