Approximation of test channels in source coding

S. Sandeep Pradhan [1]
EECS Department
Univ. of Michigan, Ann Arbor, MI
{pradhanv}@eecs.umich.edu

Abstract — **We consider the asymptotic properties of source code sequences which approach the optimal rate-distortion bound. In this paper we show that for any arbitrary source code sequence which approach the optimal rate-distortion function $R(D)$ of a discrete memoryless source, the empirical conditional distribution of the n-length source sequence given the n length reconstruction sequence is close to the n-product of the unique minimum-mutual-information test channel conditional distribution. This closeness is given by the convergence of the normalized conditional divergence. One of the implications of this result is that it is possible to approximate arbitrary discrete memoryless channels as test channels in source coding. Though our results are presented for stationary discrete memoryless sources, these can be generalized to sources with memory.**

I. Introduction

Source coding [1, 8, 9] deals with efficient representations of information sources into index sets and efficient reconstructions of the sources from those index sets under some fidelity criteria. Ever since Shannon provided an information-theoretic characterization of the rate-distortion trade-off in source coding, the properties of efficient source codes have been studied in the literature.

In this paper, we explore one such set of properties of good source codes. Although, Shannon's rate-distortion function for a discrete memoryless source is given in terms of per-letter distributions, the actual codes which approach such bounds need to be constructed over sufficiently large block-lengths. The empirical properties of sequences of good source codes which approach the optimal rate-distortion bound have not been studied in the literature, even though, the asymptotic properties of many quantizers have been studied in detail [2, 3].

An outcome of Shannon's formulation of rate-distortion trade-off in source coding is the remarkable notion of test channels, in which the reconstruction random variable denoted by $\hat{X}$ is related to the source X via a forward channel $V_{\hat{X}|X}$, and the test channel characterized by $W_{X|\hat{X}}$. This is illustrated in Fig. 1.

The main results of this paper are summarized as follows. The following results are with respect to discrete memoryless sources with a bounded distortion measure. It is possible to extend these results to more general sources.

- The empirical conditional distribution of the n-length source vector X^n given the n-length reconstruction vector $\hat{X}^n$, i.e., $\hat{W}_{X^n|\hat{X}^n}$ of any sequence of good source codes is asymptotically close to the n-product of $\bar{W}_{X|\hat{X}}$, where $\bar{W}_{X|\hat{X}}$ is the unique test channel conditional distribution (uniqueness is also proved in this paper) induced from the information-theoretic minimization of the Shannon mutual information subject to the fidelity criterion. A sequence of source codes is said to be good if they approach the optimal rate-distortion function asymptotically. That is, although the encoder may be acting on blocks of source samples at a time, the resulting conditional distribution of the source block given the corresponding block of reconstruction samples approaches that of a discrete memoryless channel. In other words, it is possible to simulate discrete memoryless channels via test channels using a sequence of good source codes. The closeness of approximation is given by the convergence of the normalized divergence.
- Although a similar result does not hold true for the case of the empirical distribution $\hat{V}_{\hat{X}^n|X^n}$ of the n-length reconstruction vector $\hat{X}^n$ given the source vector X^n, and the empirical distribution $\hat{U}_{\hat{X}^n}$ of the n-length reconstruction vector, the first order empirical distribution of the above are close to the corresponding minimum-mutual-information distributions.

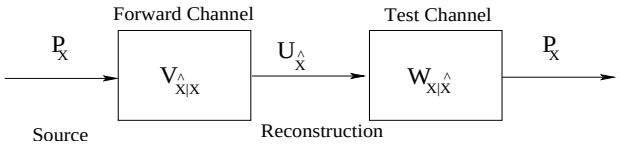

Figure 1: Source coding: the forward channel is given by $V_{\hat{X}|X}$ and the test channel is given by $W_{X|\hat{X}}$.

The precise statements of these results are given in the sequel. It is worth noting that similar results have been obtained for the case of channel coding in [4, 5, 6]. In this paper our analysis follows that of [6]. This leads us to the following interesting observation. Consider a pair of dual [7] source and channel coding problems, where the rate-distortion function $R(D)$ is equal to the capacity cost function $C(W)$, and a pair of good source code sequence and a good channel code sequence which approach $R(D) = C(W)$. Given a pair of vectors $\mathbf{x} \in \mathcal{X}^n$ and $\hat{\mathbf{x}} \in \hat{\mathcal{X}}^n$, for an observer, as n increases, it will become increasingly difficult to detect whether this pair of vectors came from a source coding operation or a channel coding operation. In other words, loosely speaking, asymptotically, the only way one could differentiate between source coding and channel coding operation is the causal relation re-

[1]This work was supported by NSF ITR grant CCR-0219735.

lation between the source/channel-output vector $\mathbf{x}$ and the reconstruction/channel-input vector $\hat{\mathbf{x}}$.

II. Notation and main results

Consider a stationary discrete memoryless source (DMS) X characterized by a distribution P_X with a finite alphabet $\mathcal{X}$, a reconstruction alphabet $\hat{\mathcal{X}}$, a bounded distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$. The distortion measure on vectors is defined by the average distortion of its samples:

$$d(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$$

$\forall \mathbf{x} \in \mathcal{X}^n$, $\hat{\mathbf{x}} \in \hat{\mathcal{X}}^n$ where x_i and $\hat{x}_i$ denote the ith samples of $\mathbf{x}$ and $\hat{\mathbf{x}}$ respectively, and n denotes block-length. We use the letters P, Q and U to denote distributions, and V and W to denote conditional distributions.

A source code $\{\mathbb{C}, f\}$ with parameters (n, M, Δ) consists of (1) $\mathbb{C} \subset \hat{\mathcal{X}}^n$, $|\mathbb{C}| = M$, and (2) a mapping $f : \mathcal{X}^n \to \mathbb{C}$; such that $Ed(\mathbf{X}, f(\mathbf{X})) = \Delta$. A sequence of source codes $\{\{\mathbb{C}_n, f_n\}\}_{n=1}^{\infty}$ (where the code with index n has parameters (n, M_n, Δ_n)) is said to be good with respect to a triple (P_X, d, D) of source distribution P_X, distortion measure d and a number $D \in \mathbb{R}^+$ if

$$\lim_{n\to\infty} \frac{1}{n} \log M_n = R(D) \text{ and } \lim_{n\to\infty} \Delta_n = D,$$

where $R(D)$ is the optimal Shannon rate-distortion function:

$$R(D) = \inf_{V_{\hat{X}|X} : Ed(X,\hat{X}) \le D} I(P_X, V_{\hat{X}|X}),$$

where for a pair of distribution P_X and conditional distribution $V_{\hat{X}|X}$, $I(P_X, V_{\hat{X}|X})$ denotes mutual information [8, 9]. For ease of notation, let us denote $\hat{X}^n = f_n(X^n)$ for a source encoder mapping f_n. Let $U_{\hat{X}^n}$, $V_{\hat{X}^n|X^n}$ and $W_{X^n|\hat{X}^n}$ denote the distribution of $\hat{X}^n$, encoder (quantizer) transformation and the conditional distribution of the source vector given the reconstruction vector associated with the source code $\{\mathbb{C}_n, f_n\}$. For any triple $(P_X, V_{\hat{X}|X}, V'_{\hat{X}|X})$ consisting of a distribution P_X and conditional distributions $V_{\hat{X}|X}$, and $V'_{\hat{X}|X}$ define

$$D(P_X|V_{\hat{X}|X}\|V'_{\hat{X}|X}) = \sum_{x\in\mathcal{X}} \sum_{\hat{x}\in\hat{\mathcal{X}}} P_X(x) V_{\hat{X}|X}(\hat{x}|x) \log \frac{V_{\hat{X}|X}(\hat{x}|x)}{V'_{\hat{X}|X}(\hat{x}|x)}.$$

It can be noted that $D(P_X|V_{\hat{X}|X}\|V'_{\hat{X}|X}) \ge 0$. Equality holds if and only if $\forall x \in \mathcal{X}$ such that $P_X(x) > 0$, and $\forall \hat{x} \in \hat{\mathcal{X}}$, the following equality holds: $V_{\hat{X}|X}(\hat{x}|x) = V'_{\hat{X}|X}(\hat{x}|x)$. Further, since we have asummed the distortion measure to be bounded, it follows that in the information-theoretic optimization of $I(P_X, V_{\hat{X}|X})$, $\forall\, V_{\hat{X}|X}$ that achieve the optimality, the following is true: $\forall \hat{x} \in \hat{\mathcal{X}}$, the induced $U_{\hat{X}}(\hat{x}) > 0$. This can be seen from the following arguments. Using Lemma 4 of [11], if for some $\hat{x} \in \hat{\mathcal{X}}$, and some $V'_{\hat{X}|X}$ that achieves the minimization, the induced distribution $U'_{\hat{X}}(\hat{x}) = 0$, then the distortion measure has to satisfy the following relation: $\forall x \in \mathcal{X}$,

$$d(x, \hat{x}) \ge -c_2 \log W'_{X|\hat{X}}(x|\hat{x}) + d_0(x) \tag{1}$$

for some $c_2 > 0$ and arbitrary $d_0(x)$, where $W'_{X|\hat{X}}$ is the induced test channel distribution. This is because $\forall x \in \mathcal{X}$, $V'_{\hat{X}|X}(\hat{x}|x) = 0$. Since $d_0(x)$ has to finite, this implies that $\forall x \in \mathcal{X}$, $d(x, \hat{x}) = \infty$ which is a contradiction of the assumption that the distortion measure is bounded.

First we have the following lemma which asserts the obvious statement that the mutual information per sample between the n-length source vector X^n and the reconstruction vector $\hat{X}^n$ is asymptotically equal to the rate-distortion function.

Lemma 1: For any source code sequence which is good with respect to a triple (P_X, d, D), we have

$$\lim_{n\to\infty} \frac{1}{n} I(P_X^n, V_{\hat{X}^n|X^n}) = R(D).$$

Proof: By the definition of $R(D)$ we have

$$\lim_{n\to\infty} \frac{1}{n} I(P_X^n, V_{\hat{X}^n|X^n}) \ge R(D). \tag{2}$$

Using the property of good codes,

$$\frac{1}{n} I(P_X^n, V_{\hat{X}^n|X^n}) = \frac{1}{n} H(\hat{X}^n) \le \frac{1}{n} \log M_n. \tag{3}$$

From this the result follows. □

Next we shall note the important properties of the test channel distribution induced from the information theoretic per-letter characterization of the rate-distortion function. The most important of them states that even though, the conditional distribution $V_{\hat{X}|X}$ that minimizes $I(P_X, V_{\hat{X}|X})$ such that $Ed \le D$ may not be unique, the resulting test channel conditional distribution $W_{X|\hat{X}}$ is unique. The proofs of the following theorems are given in the Appendix.

Theorem 1: For a stationary DMS P_X, and $\forall W_{X|\hat{X}} \in \mathcal{D}'(D)$, we have

(a) $W_{X|\hat{X}} \ll \bar{W}_{X|\hat{X}}$,

(b) $I(P_X, V_{\hat{X}|X}) - I(P_X, \bar{V}_{\hat{X}|X}) \ge D(U_{\hat{X}}|W_{X|\hat{X}}\|\bar{W}_{X|\hat{X}})$,

(c) If $V_{\hat{X}|X} \ll \bar{V}_{\hat{X}|X}$, then the above holds with equality,

(d) If $V_{X|\hat{X}}$ achieves $I(P_X, V_{\hat{X}|X}) = I(P_X, \bar{V}_{\hat{X}|X})$, then $W_{X|\hat{X}} = \bar{W}_{X|\hat{X}}$.

where

- $\bar{V}_{\hat{X}|X} = \underset{V_{\hat{X}|X} \,:\, Ed \le D}{\operatorname{argmin}} I(P_X, V_{\hat{X}|X})$, $\bar{W}_{X|\hat{X}}(x|\hat{x}) = \frac{P_X(x)\bar{V}_{\hat{X}|X}(\hat{x}|x)}{\sum_{a\in\mathcal{X}} P_X(a)\bar{V}_{\hat{X}|X}(\hat{x}|a)}$, $\forall x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}$,
- $\mathcal{D}(D) = \{V_{\hat{X}|X} : Ed \le D\}$,
- $$\mathcal{D}'(D) = \left\{ W_{X|\hat{X}} : W_{X|\hat{X}}(x|\hat{x}) = \frac{P_X(x) V_{\hat{X}|X}(\hat{x}|x)}{\sum_{a\in\mathcal{X}} P_X(a) V_{\hat{X}|X}(\hat{x}|a)} \right.$$
$$\left. \forall x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}} \text{ for some } V_{\hat{X}|X} \in \mathcal{D}(D) \right\}.$$

The above theorem leads us to the next result which says that for a source code sequence good with respect to (P_X, d, D), the empirical conditional distribution of the n-length source vector given the n-length reconstruction vector is asymptotically close to the n-product of the unique test channel conditional distribution.

Theorem 2: For a source code sequence good with respect to (P_X, d, D), we have

$$\lim_{n\to\infty} \frac{1}{n} D(U_{\hat{X}^n}|\hat{W}_{X^n|\hat{X}^n}\|\bar{W}_{X|\hat{X}}^n) = 0,$$

where $\bar{W}^n_{X|\hat{X}}$ is the n-product of the unique test channel distribution $\bar{W}_{X|\hat{X}}$ that achieves the optimal rate-distortion bound.

One would expect to get a similar result for the quantizer conditional distribution and the reconstruction distribution. But a more careful look at these distributions reveal that it is far from the corresponding n-product distributions induced from the per-letter optimization of $I(P_X, V_{\hat{X}|X})$. Still, we will show in the sequel that the above approximation is true for the first order empirical distributions, and is elucidated in Theorem 3, 4 and 5. It should be noted that similar results can be obtained for the corresponding kth order empirical distributions for fixed k, similar to those considered in [6]. Before we do this, let us consider the following definitions. For any source code sequence $\{\{\mathbb{C}_n, f_n\}\}_{n=1}^{\infty}$, define the encoder conditional distributions as

$$V_{\hat{X}^n|X^n}(\hat{\mathbf{x}}|\mathbf{x}) = \begin{cases} 1 & \text{if} \quad f_n(\mathbf{x}) = \hat{\mathbf{x}} \\ 0 & \text{else} \end{cases},$$

$\forall \mathbf{x} \in \mathcal{X}^n, \hat{\mathbf{x}} \in \hat{\mathcal{X}}^n$. Define $\forall a \in \mathcal{X}, b \in \hat{\mathcal{X}}$, the per-letter joint distribution of the ith sample of the source and the reconstruction induced from the encoder mapping as

$$\hat{Q}_{X_i,\hat{X}_i}(a,b) = \sum_{\mathbf{x}\in\mathcal{X}^n : x_i = a} \sum_{\hat{\mathbf{x}}\in\hat{\mathcal{X}}^n : \hat{x}_i = b} P_X^n(\mathbf{x}) V_{\hat{X}^n|X^n}(\hat{\mathbf{x}}|\mathbf{x}),$$

the first order empirical conditional distribution of the quantizer as

$$\hat{V}^{(1)}_{\hat{X}^n|X^n}(b|a) = \frac{1}{P_X(a)} \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_{X_i,\hat{X}_i}(a,b)$$

the first order empirical reconstruction distribution as

$$\hat{U}^{(1)}_{\hat{X}^n}(b) = \sum_{x\in\mathcal{X}} \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_{X_i,\hat{X}_i}(x,b)$$

and the first order empirical conditional distribution of source vector given the reconstruction vector as

$$\hat{W}^{(1)}_{X^n|\hat{X}^n}(a|b) = \frac{1}{\hat{U}^{(1)}_{\hat{X}^n}(b)} \frac{1}{n} \sum_{i=1}^{n} \hat{Q}_{X_i,\hat{X}_i}(a,b)$$

Theorem 3: For a source code sequence good with respect to (P_X, d, D), the first order empirical conditional distribution of the source vector given the reconstruction vector is close to the unique minimum-mutual-information conditional distribution $\bar{W}_{X|\hat{X}}$:

$$\lim_{n\to\infty} D(\hat{U}^{(1)}_{\hat{X}^n} | \hat{W}^{(1)}_{X^n|\hat{X}^n} \| \bar{W}_{X|\hat{X}}) = 0.$$

The convergence results for the empirical conditional distribution of the source vector given the reconstruction vector can not be easily extended to the empirical distribution of the reconstruction vector given the source vector and the empirical distribution of the reconstruction vector. But for a class of source codes, called regular codes, such convergence of these distributions take place. An (n, M, Δ) source code is said to be regular if the conditional distribution $V_{\hat{X}^n|X^n} \ll \bar{V}^n_{\hat{X}|X}$ for some $\bar{V}^n_{\hat{X}|X}$ such that $Ed \le D$, and

$$I(P_X^n, \bar{V}^n_{\hat{X}|X}) = nR(D). \tag{4}$$

In the following theorems we will present these convergence results.

Theorem 4: For a regular source code sequence good with respect to (P_X, d, D), the first order empirical conditional distribution associated with the encoder transformation is close to one of the minimum-mutual-information conditional distributions:

$$\lim_{n\to\infty} \min_{V_{\hat{X}|X} : I(P_X, V_{\hat{X}|X}) = R(D)} D(P_X | \hat{V}^{(1)}_{\hat{X}^n|X^n} \| V_{\hat{X}|X}) = 0$$

Theorem 5: For a regular source code sequence good with respect to (P_X, d, D), the first order empirical distribution of the reconstruction vector is close to the one of the minimum-mutual-information reconstruction distributions:

$$\lim_{n\to\infty} \min_{U_{\hat{X}} \in \mathcal{B}(D)} D(U^{(1)}_{\hat{X}^n} \| U_{\hat{X}}) = 0,$$

where

$$\mathcal{B}(D) = \Big\{ U_{\hat{X}} : \exists V_{\hat{X}|X} \text{ such that } \forall \hat{x} \in \hat{\mathcal{X}},$$

$$U_{\hat{X}}(\hat{x}) = \sum_{x\in\mathcal{X}} P_X(x) V_{\hat{X}|X}(\hat{x}|x), \text{ and } I(P_X, V_{\hat{X}|X}) = R(D) \Big\}.$$

III. Appendix

Proof of Theorem 1: Let us prove statement (a) by contradiction. Suppose there exists $V'_{\hat{X}|X}$ which induces $W'_{X|\hat{X}}$ such that for some $a \in \mathcal{X}$ and $b \in \hat{\mathcal{X}}$ the following is true:

$$\bar{W}_{X|\hat{X}}(a|b) = 0 < W'_{X|\hat{X}}(a|b). \tag{5}$$

Consider $V^{(\alpha)}_{\hat{X}|X} = \alpha V'_{\hat{X}|X} + (1-\alpha)\bar{V}_{\hat{X}|X}$. Now

$$\begin{aligned}
I(P_X, V^{(\alpha)}_{\hat{X}|X}) &= (1-\alpha) \sum_{x,\hat{x}} P_X(x) \bar{V}_{\hat{X}|X}(\hat{x}|x) \log V^{(\alpha)}_{\hat{X}|X}(\hat{x}|x) \\
&\quad + \alpha \sum_{x,\hat{x}} P_X V'_{\hat{X}|X}(\hat{x}|x) \log V^{(\alpha)}_{\hat{X}|X} + H(U^{(\alpha)}_{\hat{X}}) \\
&= (\alpha-1) D(P_X | \bar{V}_{\hat{X}|X} \| V^{(\alpha)}_{\hat{X}|X}) \\
&\quad + (\alpha-1) H(\bar{V}_{\hat{X}|X} | P_X) - \alpha D(P_X | V'_{\hat{X}|X} \| V^{(\alpha)}_{\hat{X}|X}) \\
&\quad - \alpha H(V'_{\hat{X}|X} | P_X) + H(U^{(\alpha)}_{\hat{X}}) \\
&= (\alpha-1) D(\bar{U}_{\hat{X}} | \bar{W}_{X|\hat{X}} \| W^{(\alpha)}_{X|\hat{X}}) \\
&\quad + (\alpha-1) D(\bar{U}_{\hat{X}} \| U^{(\alpha)}_{\hat{X}}) + (\alpha-1) H(\bar{V}_{\hat{X}|X} | P_X) \\
&\quad - \alpha D(U'_{\hat{X}} | W'_{X|\hat{X}} \| W^{(\alpha)}_{X|\hat{X}}) - \alpha D(U'_{\hat{X}} \| U^{(\alpha)}_{\hat{X}}) \\
&\quad - \alpha H(V'_{\hat{X}|X} | P_X) + H(U^{(\alpha)}_{\hat{X}}) \\
&= (\alpha-1) D(\bar{U}_{\hat{X}} | \bar{W}_{X|\hat{X}} \| W^{(\alpha)}_{X|\hat{X}}) \\
&\quad - \alpha D(U'_{\hat{X}} | W'_{X|\hat{X}} \| W^{(\alpha)}_{X|\hat{X}}) + \alpha I(P_X, V'_{\hat{X}|X}) \\
&\quad + (1-\alpha) I(P_X, \bar{V}_{\hat{X}|X}) \\
&\le I(P_X, \bar{V}_{\hat{X}|X}) - \alpha \Big[I(P_X, \bar{V}_{\hat{X}|X}) \\
&\quad - I(P_X, V'_{\hat{X}|X}) + D(U'_{\hat{X}} | W'_{X|\hat{X}} \| W^{(\alpha)}_{X|\hat{X}}) \Big]
\end{aligned}$$

Now note that

$$\begin{aligned}
W^{(\alpha)}(a|b) &= \frac{(1-\alpha)\bar{W}_{X|\hat{X}}(a|b)\bar{U}_{\hat{X}}(b) + \alpha W'_{X|\hat{X}}(a|b) U'_{\hat{X}}(b)}{U^{(\alpha)}_{\hat{X}}(b)} \\
&= \frac{\alpha W'_{X|\hat{X}}(a|b) U'_{\hat{X}}(b)}{U^{(\alpha)}_{\hat{X}}(b)},
\end{aligned}$$

and

$$D(U'_{\hat{X}}|W'_{X|\hat{X}}\|W^{(\alpha)}_{X|\hat{X}}) \geq U'_{\hat{X}}(b)W'_{X|\hat{X}}(a|b)\log\frac{W'_{X|\hat{X}}(a|b)}{W^{(\alpha)}_{X|\hat{X}}(a|b)} \quad (6)$$

$$+U'_{\hat{X}}(b)(1-W'_{X|\hat{X}}(a|b))\log\frac{(1-W'_{X|\hat{X}}(a|b))}{(1-W^{(\alpha)}_{X|\hat{X}}(a|b))}.$$

Clearly $U'_{\hat{X}}(b) > 0$, and $\bar{U}_{\hat{X}}(b) > 0$. In the limit $\alpha \downarrow 0$, the RHS of the above equation goes to ∞. Hence getting a contradiction.

We will now prove statement (b) by contradiction. Let there exist $W'_{X|\hat{X}} \in \mathcal{D}'(D)$ (which of course implies the existence of $V'(\hat{X}|X) \in \mathcal{D}(D)$), such that

$$I(P_X, V'_{\hat{X}|X}) - I(P_X, \bar{V}_{\hat{X}|X}) < D(U'_{\hat{X}}|W'_{X|\hat{X}}\|\bar{W}_{X|\hat{X}}). \quad (7)$$

Let $V^{(\alpha)}_{\hat{X}|X} = (1-\alpha)\bar{V}_{\hat{X}|X} + \alpha V'_{\hat{X}|X}$. Now

$$\begin{aligned} I(P_X, V^{(\alpha)}_{\hat{X}|X}) &= \sum_{x,\hat{x}} P_X(x)V^{(\alpha)}_{\hat{X}|X}(\hat{x}|x)\log\frac{V^{(\alpha)}_{\hat{X}|X}(\hat{x}|x)}{U^{(\alpha)}_{\hat{X}}(\hat{x})} \\ &= \alpha I(P_X, V'_{\hat{X}|X}) + (1-\alpha)I(P_X, \bar{V}_{\hat{X}|X}) \\ &\quad -\alpha D(U'_{\hat{X}}|W'_{X|\hat{X}}\|\bar{W}_{X|\hat{X}}) \\ &\quad +D(U^{(\alpha)}_X|W^{(\alpha)}_{X|\hat{X}}\|\bar{W}_{X|\hat{X}}) \end{aligned}$$

Now let us take the derivative of the above equation with respect to α to get

$$\frac{\partial I(P_X, V^{(\alpha)}_{\hat{X}|X})}{\partial\alpha} = I(P_X, V'_{\hat{X}|X}) - I(P_X, \bar{V}_{\hat{X}|X}) \quad (8)$$

$$-D(U'_{\hat{X}}|W'_{X|\hat{X}}\|\bar{W}_{X|\hat{X}}) + \frac{\partial D(U^{(\alpha)}_{\hat{X}}|W^{(\alpha)}_{X|\hat{X}}\|\bar{W}_{X|\hat{X}})}{\partial\alpha}.$$

Since $W^{(\alpha)}_{X|\hat{X}} \ll \bar{W}_{X|\hat{X}}$ (using (a)), we have

$$\frac{\partial}{\partial\alpha}D(U^{(\alpha)}_{\hat{X}}|W^{(\alpha)}_{X|\hat{X}}\|\bar{W}_{X|\hat{X}})\Big|_{\alpha=0} = 0 \quad (9)$$

which results in

$$\frac{\partial I(P_X, V^{(\alpha)}_{\hat{X}|X})}{\partial\alpha}\Big|_{\alpha=0} < 0. \quad (10)$$

So for small α, $I(P_X, V^{(\alpha)}_{\hat{X}|X})$ will be surely smaller than $I(P_X, \bar{V}_{\hat{X}|X})$ thus contradicting the optimality of $\bar{V}_{\hat{X}|X}$.

Now let us prove (c). Note that

$$I(P_X, V_{\hat{X}|X}) - I(P_X, \bar{V}_{\hat{X}|X}) - D(U_{\hat{X}}|W_{X|\hat{X}}\|\bar{W}_{X|\hat{X}}) \quad (11)$$

$$= E_{P_X V_{\hat{X}|X}}\left[\log\frac{\bar{W}_{X|\hat{X}}(X|\hat{X})}{P_X(X)}\right] - E_{P_X \bar{V}_{\hat{X}|X}}\left[\log\frac{\bar{W}_{X|\hat{X}}(X|\hat{X})}{P_X(X)}\right].$$

Let $g(x,\hat{x}) = \log\frac{\bar{W}_{X|\hat{X}}(x|\hat{x})}{P_X(x)}$ for all $x \in \mathcal{X}$ and $\hat{x} \in \hat{\mathcal{X}}$. The above equation and (b) show that $E_{P_X \bar{V}_{\hat{X}|X}} g(X,\hat{X}) \leq E_{P_X V_{\hat{X}|X}} g(X,\hat{X})$ for all $V_{\hat{X}|X}$. Now $\forall x \in \mathcal{X}$, let

$$B(x) = \max_{\hat{x}\in\hat{\mathcal{X}}} g(x,\hat{x}), \text{ and } B'(x) = \{\hat{x} \in \hat{\mathcal{X}} : g(x,\hat{x}) = B(x)\} \quad (12)$$

Thus we can conclude that $\forall x \in \mathcal{X}$, $\bar{V}_{\hat{X}|X}(\cdot|x)$ puts mass on a subset of $B'(x)$. Hence for any $V_{\hat{X}|X} \ll \bar{V}_{\hat{X}|X}$, the same must be true, thus proving the required result.

(d) follows directly from (b). An independent alternate proof of (d) can be obtained, and is given in the folioing using the techniques of [10]. Let $V'_{\hat{X}|X}$ achieve $I(P_X, V'_{\hat{X}|X}) = I(P_X, \bar{V}_{\hat{X}|X})$. Let $W'_{X|\hat{X}}$ and $\bar{W}_{X|\hat{X}}$ be the corresponding test channels respectively. With the source fixed at P_X, define a binary random variable Z with $P_Z(Z=1) = \alpha = 1-P_Z(Z=0)$. for some $0 < \alpha < 1$. Let $\hat{X}'$ be a new random variable defined on $\hat{\mathcal{X}}$ such that $P_{\hat{X}'|X} = \alpha V'_{\hat{X}|X} + (1-\alpha)\bar{V}_{\hat{X}|X}$, $P_{X|\hat{X}',Z}(X = x|\hat{X}' = \hat{x}, Z = 0) = W'_{X|\hat{X}}(x|\hat{x})$ and $P_{X|\hat{X}',Z}(X = x|\hat{X}' = \hat{x}, Z = 1) = \bar{W}_{X|\hat{X}}(x|\hat{x})$ for all $x \in \mathcal{X}$ and $\hat{x} \in \hat{\mathcal{X}}$, and $\hat{X}'$ and Z are independent. This also implies that $P_{X|\hat{X}',Z}(x|\hat{x},0) = W'_{X|\hat{X}}(x|\hat{x})$ and $P_{X|\hat{X}',Z}(x|\hat{x},1) = \bar{W}_{X|\hat{X}}(x|\hat{x})$ $\forall x \in \mathcal{X}$ and $\hat{x} \in \hat{\mathcal{X}}$. Thus the joint distribution of X, $\hat{X}'$ and Z are determined. Since $I(P_X, V_{\hat{X}|X})$ is convex in $V_{\hat{X}|X}$, and $\bar{V}_{\hat{X}|X}, V'_{\hat{X}|X}$ achieve the minimization of $I(P_X, V_{\hat{X}|X})$, we have

$$I(P_X, P_{\hat{X}'|X}) = I(P_X, \alpha V'_{\hat{X}|X} + (1-\alpha)\bar{V}_{\hat{X}|X}) = \alpha I(P_X, V'_{\hat{X}|X})$$

$$+(1-\alpha)I(P_X, \bar{V}_{\hat{X}|X}) = I(P_{X|Z}, P_{\hat{X}'|X,Z}|P_Z). \quad (13)$$

Using the chain rule of mutual information [8], it can be seen that

$$\begin{aligned} I(P_{X|\hat{X}'}, P_{Z|\hat{X}',X}|P_{\hat{X}'}) &= I(P_{X|Z}, P_{\hat{X}'|X,Z}|P_Z) - I(P_X, P_{\hat{X}'|X}) \\ &= 0 \end{aligned} \quad (14)$$

Hence

$$P_{X|\hat{X}',Z} = P_{X|\hat{X}'} \quad (15)$$

which implies that

$$W'_{X|\hat{X}}(x|\hat{x}) = P_{X|\hat{X}',Z}(x|\hat{x},0) = P_{X|\hat{X}',Z}(x|\hat{x},1) = \bar{W}_{X|\hat{X}}(x|\hat{x}) \quad (16)$$

$\forall x \in \mathcal{X}$ and $\hat{x} \in \hat{\mathcal{X}}$.

□

Proof of Theorem 2: Follows directly from Lemma 1 and Theorem 1(d).

□

Proof of Theorem 3: Note that

$$\begin{aligned} D(\hat{U}_{\hat{X}^n}|\hat{W}_{X^n|\hat{X}^n}\|\bar{W}^n_{X|\hat{X}}) &= \sum_{\mathbf{x},\hat{\mathbf{x}}} \hat{U}_{\hat{X}^n}(\hat{\mathbf{x}})\hat{W}_{X^n|\hat{X}^n}(\mathbf{x}|\hat{\mathbf{x}}) \\ &\quad \log\frac{\prod_{i=1}^n \hat{W}_{X_i|\hat{X}_i}(x_i|\hat{x}_i)}{\prod_{i=1}^n \bar{W}_{X|\hat{X}}(x_i|\hat{x}_i)} \\ &\quad +D(\hat{U}_{\hat{X}^n}|\hat{W}_{X^n|\hat{X}^n}\|\prod_{i=1}^n \hat{W}_{X_i|\hat{X}_i}) \\ &\geq \sum_{i=1}^n D(\hat{U}_{\hat{X}_i}|\hat{W}_{X_i|\hat{X}_i}|\bar{W}_{X|\hat{X}}). \end{aligned}$$

The RHS is equal to

$$\left[\sum_{\hat{x}\in\hat{\mathcal{X}}}\left\{\sum_{j=1}^n \hat{U}_{\hat{X}_j}(\hat{x})\right\}\left\{\sum_{i=1}^n \frac{\hat{U}_{\hat{X}_i}(\hat{x})}{\sum_{j=1}^n \hat{U}_{\hat{X}_j}(\hat{x})}\right.\right.$$

$$\left.\left. D(\hat{W}_{X_i|\hat{X}_i}(\cdot|\hat{x})\|\bar{W}_{X|\hat{X}}(\cdot|\hat{x}))\right\}\right]$$

$$\geq n \left[\sum_{\hat{x} \in \hat{\mathcal{X}}} \left\{ \frac{1}{n} \sum_{j=1}^{n} \hat{U}_{\hat{X}_j}(\hat{x}) \right\} \right.$$
$$\left. D\left(\frac{\sum_{i=1}^{n} \hat{U}_{\hat{X}_i}(\hat{x}) \hat{W}_{X_i|\hat{X}_i}(\cdot|\hat{x})}{\sum_{j=1}^{n} \hat{U}_{\hat{X}_j}(\hat{x})} \| \bar{W}_{X|\hat{X}}(\cdot|\hat{x}) \right) \right]$$
$$= nD(\hat{U}^{(1)}_{\hat{X}^n} | W^{(1)}_{X^n|\hat{X}^n} \| \bar{W}_{X|\hat{X}}),$$

where we have used the non-negativity and the convexity of information divergence. Now using Theorem 2, we get the desired result.

□

Proof of Theorem 4: To prove this theorem, let us define the following functions:

$$r(V_{\hat{X}|X}) = \min_{V'_{\hat{X}|X} \in \mathcal{V}} D(P_X | V_{\hat{X}|X} \| V'_{\hat{X}|X}) \tag{17}$$

$$w(\delta) = \max_{V_{\hat{X}|X}: h(V_{\hat{X}|X}) \leq \delta, Ed \leq D} r(V_{\hat{X}|X}), \tag{18}$$

where in the above equation $\forall a \in \mathcal{X}$ and $b \in \hat{\mathcal{X}}$, the following are true: $U_{\hat{X}}(b) = \sum_{a \in \mathcal{X}} P_X(a) V_{\hat{X}|X}(b|a)$ and $W_{X|\hat{X}}(a|b) = P_X(a) V_{\hat{X}|X}(b|a) / U_{\hat{X}}(b)$,

$$\mathcal{V} = \{V_{\hat{X}|X} : Ed \leq D, I(P_X, V_{\hat{X}|X}) = R(D)\},$$

and

$$h(V_{\hat{X}|X}) = D\left(P_X | \frac{P_X V_{\hat{X}|X}}{\sum_{a \in \mathcal{X}} P_X(a) V_{\hat{X}|X}(\cdot|a)} \| \bar{W}_{X|\hat{X}} \right). \tag{19}$$

Note that $\forall a \in \mathcal{X}$ and $b \in \hat{\mathcal{X}}$

$$W^{(1)}_{X^n|\hat{X}^n}(a|b) = \frac{P_X(a) V^{(1)}_{\hat{X}^n|X^n}(b|a)}{\sum_{x \in \mathcal{X}} P_X(x) V^{(1)}_{\hat{X}^n|X^n}(b|x)}. \tag{20}$$

Hence for a fixed source distribution, if $W^{(1)}_{X^n|\hat{X}^n}$ is close to $\bar{W}_{X|\hat{X}}$, then $V^{(1)}_{\hat{X}|X}$ must be close to $\bar{V}_{\hat{X}|X}$.

Since the code is regular, using Theorem 1(b) and 1(c), it can be noted that $w(0) = 0$. Consider the function $r(\cdot)$ of any two conditional distributions $V^*_{\hat{X}|X}$ and $V^\dagger_{\hat{X}|X}$ as follows:

$$\begin{aligned}
r(\alpha V^*_{\hat{X}|X} + (1-\alpha) V^\dagger_{\hat{X}|X}) &= \min_{V'_{\hat{X}|X} \in \mathcal{V}} D\left(P_X | \right. \\
&\qquad \left. (\alpha V^*_{\hat{X}|X} + (1-\alpha) V^\dagger_{\hat{X}|X}) \| V'_{\hat{X}|X} \right) \\
&= \min_{V'_{\hat{X}|X} \in \mathcal{V}, V''_{\hat{X}|X} \in \mathcal{V}} D\left(P_X | \right. \\
&\qquad (\alpha V^*_{\hat{X}|X} + (1-\alpha) V^\dagger_{\hat{X}|X}) \\
&\qquad \left. \| \alpha V'_{\hat{X}|X} + (1-\alpha) V''_{\hat{X}|X} \right) \\
&\leq \min_{V'_{\hat{X}|X} \in \mathcal{V}, V''_{\hat{X}|X} \in \mathcal{V}} \sum_{x \in \mathcal{X}} P_X(x) \\
&\qquad \Big[\alpha D(V^*_{\hat{X}|X}(\cdot|x) \| V'_{\hat{X}|X}(\cdot|x)) \\
&\quad + (1-\alpha) D(V^\dagger_{\hat{X}|X}(\cdot|x) \| V''_{\hat{X}|X}(\cdot|x)) \Big] \\
&= \alpha r(V^*_{\hat{X}|X}) + (1-\alpha) r(V^\dagger_{\hat{X}|X})
\end{aligned}$$

where we have used the following fact

$$\mathcal{V} = \{V_{\hat{X}|X} = \alpha V'_{\hat{X}|X} + (1-\alpha) V''_{\hat{X}|X} : V'_{\hat{X}|X} \in \mathcal{V}, V''_{\hat{X}|X} \in \mathcal{V}\} \tag{21}$$

i.e., $\mathcal{V}$ is convex using the convexity of mutual information in $V_{\hat{X}|X}$, and the convexity of information divergence. Hence r is convex. Clearly h is continuous and convex in $V_{\hat{X}|X}$. Hence the set $\mathcal{K}(\delta) = \{V_{\hat{X}|X} : h(V_{\hat{X}|X}) \leq \delta, Ed \leq D\}$ is compact, and thus

$$w(\delta) = r(V_{\hat{X}|X,\delta}) \tag{22}$$

for some $V_{\hat{X}|X,\delta} \in \mathcal{K}(\delta)$. Consider an arbitrary convergent decreasing sequence $\delta_n \to 0$. There must exist $V_{\hat{X}|X,\delta_n} \in \mathcal{K}(\delta_n)$ for all n. Since $\mathcal{K}$, which is the set of all conditional distributions $V_{\hat{X}|X}$, such that $Ed \leq D$ is compact, there must exists a subsequence $V_{\hat{X}|X,\delta_{n_k}} \to V^{**}_{\hat{X}|X}$, for some $V^{**}_{\hat{X}|X} \in \mathcal{K}$. Using the property of r and h,

$$\lim_{k \to \infty} w(\delta_{n_k}) = \lim_{k \to \infty} r(V_{\hat{X}|X,\delta_{n_k}}) = r(V^{**}_{\hat{X}|X}), \tag{23}$$

and

$$h(V^{**}_{\hat{X}|X}) = \lim_{k \to \infty} h(V_{\hat{X}|X,\delta_{n_k}}) \leq \lim_{k \to \infty} \delta_{n_k} = 0. \tag{24}$$

Hence $V^{**}_{\hat{X}|X} \in \mathcal{K}(0)$. This implies that

$$w(0) \geq r(V^{**}_{\hat{X}|X}) = \lim_{k \to \infty} w(\delta_{n_k}). \tag{25}$$

Since w is monotone nondecreasing, the result follows. □

Proof of Theorem 5: Define the set $\mathcal{V}$ as before, and the following two functions

$$w(\delta) = \max_{U_{\hat{X}} \in \mathcal{A}(\delta)} r(U_{\hat{X}}), \text{ and } r(U_{\hat{X}}) = \min_{\bar{U}_{\hat{X}} \in \mathcal{B}(D)} D(U_{\hat{X}} | \bar{U}_{\hat{X}}) \tag{26}$$

where

$$\mathcal{A}(\delta) = \{U_{\hat{X}} : g^*(U_{\hat{X}}) \leq \delta\}, \tag{27}$$

where

$$g^*(U_{\hat{X}}) = \min_{\{V_{\hat{X}|X} : \sum_x P_X(x) V_{\hat{X}|X}(\cdot|x) = U_{\hat{X}}(\cdot)\}} \tag{28}$$
$$\min_{\{V'_{\hat{X}|X} \in \mathcal{V}\}} D(P_X | V_{\hat{X}|X} \| V'_{\hat{X}|X}).$$

First note that $\mathcal{V}$ is compact using the continuity of mutual information. Also note that for a fixed P_X, $D(P_X | V_{\hat{X}|X} \| V'_{\hat{X}|X})$ is convex in the pair $(V_{\hat{X}|X}, V'_{\hat{X}|X})$. We can also note that

$$g^*(U_{\hat{X}}) = \min_{\{V_{\hat{X}|X} : \sum_x P_X(x) V_{\hat{X}|X}(\cdot|x) = U_{\hat{X}}(\cdot)\}} h^*(V_{\hat{X}|X}) \tag{29}$$

where h^* is given by

$$h^*(V_{\hat{X}|X}) = \min_{\{V'_{\hat{X}|X} \in \mathcal{V}\}} D(P_X | V_{\hat{X}|X} \| V'_{\hat{X}|X}). \tag{30}$$

We have the following Lemmas:

Lemma A: For a function $d'(y,z)$ which is convex and continuous in the pair (y,z) over compact and convex sets $\mathcal{Z}$ and $\mathcal{Y}$, the following function h' is convex and continuous over $\mathcal{Y}$

$$h'(y) = \min_{z \in \mathcal{Z}} d'(y,z). \tag{31}$$

Lemma B: For a function $h^*(y)$ which is convex and continuous in y over a convex and compact set $\mathcal{Y}$, a linear function

l, and a compact set $\mathcal{W} = \{w : \exists y \in \mathcal{Y} \text{ such that } l(y) = w\}$, the following function is convex and continuous in w over $\mathcal{W}$,

$$g'(w) = \min_{y \in l^{-1}(w)} h^*(y). \tag{32}$$

Using the above two lemmas, we can conclude that g^* is convex and continuous in $U_{\hat{X}}$. Since r is convex and continuous, and using the arguments of the proof of Theorem 4, the desired result follows.

□

Proof of Lemma A: Let $y = \alpha y_1 + (1-\alpha)y_2$ where y, y_1 and $y_2 \in \mathcal{Y}$. Then

$$\begin{aligned} h'(\alpha y_1 + (1-\alpha)y_2) &= \min_{z \in \mathcal{Z}} d'(y, z) \\ &\leq \min_{z_1, z_2 \in \mathcal{Z}} d'(y, \alpha z_1 + (1-\alpha)z_2) \\ &\leq \min_{z_1, z_2 \in \mathcal{Z}} \alpha d'(y_1, z_1) + (1-\alpha) d'(y_2.z_2) \\ &= \alpha h'(y_1) + (1-\alpha) h'(y_2). \end{aligned}$$

Hence we have shown that h' is convex. Now consider an arbitrary convergent sequence $y_n \to y_0$. $\forall n$, we have $h'(y_n) = d'(y_n, z_n^*)$ for some $z_n^* \in \mathcal{Z}$. $\exists$ a convergent subsequence z_{n_k} such that $\lim_{k\to\infty} z_{n_k} = z_0^*$ for some $z_0^* \in \mathcal{Z}$. Now

$$h'(y_0) = \min_{z \in \mathcal{Z}} d'(y_0, z) \leq d'(y_0, z_0^*) = \lim_{k\to\infty} d'(y_{n_k}, z_{n_k}^*) \tag{33}$$

$$= \lim_{k\to\infty} h'(y_{n_k}).$$

Now using the convexity of h' it can be concluded that h' is continuous over $\mathcal{Y}$.

□

Proof of Lemma B: Let $w = \alpha w_1 + (1-\alpha)w_2$, with $w, w_1, w_2 \in \mathcal{W}$. Clearly $\forall w \in \mathcal{W}$, $l^{-1}(w)$ is compact and convex. Now

$$\begin{aligned} g'(w) &= \min_{y \in l^{-1}(w)} h^*(y) \\ &\leq \min_{y_1 \in l^{-1}(w_1), y_2 \in l^{-1}(w_2)} h^*(\alpha y_1 + (1-\alpha)y_2) \\ &\leq \min_{y_1 \in l^{-1}(w_1), y_2 \in l^{-1}(w_2)} \alpha h^*(y_1) + (1-\alpha) h^*(y_2) \\ &= \alpha g'(w_1) + (1-\alpha) g'(w_2). \end{aligned}$$

Hence we have shown that g' is convex. Using arguments similar to the proof of Lemma A, we can show that for any convergent sequence $w_n \to w_0$ in $\mathcal{W}$, there exists a subsequence w_{n_k} such that $g'(w_0) \leq \lim_{k\to\infty} g(w_{n_k})$. Hence the desired result follows.

□

Acknowledgments

The author would like to thank Professor Kannan Ramchandran and Dr. Prakash Ishwar of the University of California at Berkeley, and Professor David Neuhoff of University of Michigan for inspiring discussions on duality.

References

[1] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *In IRE Nat. Conv. Rec.*, vol. Part 4, pp. 142–163, 1959.

[2] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. on Information Theory*, vol. 44, pp. 2325–2383, October 1998.

[3] D. Hui and D. L. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," *IEEE Trans. on Information Theory*, vol. 47, pp. 957–977, March 2001.

[4] D. L. Neuhoff and P. C. Shields, "Channel entropy and primitive approximation," *Ann. Probab.*, vol. 10, pp. 188–198, Feb 1982.

[5] T. S. Han and S. Verdu, "Approximation theory of output statistics," *IEEE Trans. Inform. Theory*, vol. 39, pp. 752–772, May 1993.

[6] S. Shamai and S. Verdu, "The empirical distribution of good codes," *IEEE Trans. Inform. Theory*, vol. 43, pp. 836–846, May 1997.

[7] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. on Information Theory*, vol. 49, pp. 1181–1203, May 2003.

[8] T. M. Cover and J. A. Thomas, *Elements of information theory.* New York: Wiley, 1991.

[9] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete memoryless sources.* Academic Press, New York, 1981.

[10] R. G. Gallager, *Information Theory and Reliable Communication.* New York: John Wiley and Sons, 1968.

[11] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code or not to code: Lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, pp. 1147–1158, May 2003.