Because the last expression is independent of k or τ^*, its negative logarithm is a lower bound on $I(\mu_m)$, that is

$$\begin{aligned} I(\mu_m) &\geq -E_{\tau^*} \log E_\tau e^{-n/4 \cdot H^2(f_\tau, f_{\tau^*})} \\ &= -E_{\tau^*} \log E e^{-n/4 \cdot C_m W} \\ &= -\log\left(2^{-m}(1+e^{-t_m})^m\right) \\ &= m \log 2 - m \log\left(1+e^{-t_m}\right) \\ &= m \log 2 - m \log\left(1+e^{-\frac{n}{4}C_m}\right) \\ &\geq m \log 2 - m \log\left(1+e^{-\frac{n}{4}B_g m^{-2s-1}}\right). \end{aligned}$$

Choosing $m = An^{1/(2s+1)}$ $(A > 0)$ to maximize the rate in the above lower bound, we get the following theorem.

Theorem:

$$\min_{q_n} \max_{f \in LIP(s,C)} E_{f^n} \log\left(f^n/q_n\right) \geq I(\mu_m) \geq A_{g,s} n^{1/(2s+1)}$$

where

$$\begin{aligned} A_{g,s} &= A(\log 2 - \log(1+e^{-B_g/(4A^{2s+1})})) \\ &= A(\log 2 - \log(1+e^{-a_g c^2/(4A^{2s+1})})) > 0. \end{aligned}$$

Taking $k = 0$, $\nu = 1$ therefore $s = 1$ in the theorem, we obtain the optimal rate lower bound in [13], as shown in the corollary below.

Corollary:

$$\min_{q_n} \max_{f \in LIP(1,C)} E_{f^n} \log\left(f^n/q_n\right) \geq O(n^{1/3}).$$

Remarks:

1) In general, we can consider the $LIP(s, C)$ classes on $[0, 1]^d$ $(d \geq 1)$. Minimax lower bounds on redundancy of rates $O(n^{d/(2s+d)})$ can be obtained. These rates are believed to be optimal in the sense that universal codes can be constructed to achieve these rates. In the case of $LIP(1, C)$ the rate $n^{1/3}$ has been shown to be optimal in [13].
2) The proof for the minimax lower bound $\frac{k}{2} \log n$ in the parametric case follows from the asymptotic expansion of $I(\mu)$ in [1] or [8] for smooth priors. Superficially, this approach has a continuous flavor since μ needs to have nice smoothness properties on the whole parameter space, whereas the proof in the nonparametric case as we just saw has a discrete flavor because of the hypercube subclass over which $I(\mu_m)$ is estimated. Heuristically, however, the continuous prior can be made discrete. Under regularity conditions, we believe that $I(\mu)$ should give the same lower bound $\frac{k}{2} \log n$ even for a discrete uniform prior μ on a grid subset of the parametric space, as long as the grid size is of the order or smaller than $n^{-1/2}$. Note that the nearest neighbors on the hypercube for the optimal choice $m = n^{1/(2s+1)}$ also have Hellinger distances of order $n^{-1/2}$, the rate at which n i.i.d. data points can possibly distinguish two probability densities. In other words, what seems essential to both the parametric and the nonparametric case is to find a subclass of densities whose closest elements are $n^{-1/2}$ apart from each other in terms of Hellinger distance.

ACKNOWLEDGMENT

The author wishes to thank Dr. J. Rissanen, an anonymous referee, the Associate Editor, and J. Huang for their very helpful comments. Special thanks are due to Dr. N. Hengartner for spotting an inexcusable error in an earlier version of the correspondence and for pointing to the Haussler paper.

REFERENCES

[1] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory,* vol. 36, pp. 453–471, May 1990.

[2] I. Csiszár, "Information theoretical methods in statistics," class notes, University of Maryland, College Park, Spring 1990.

[3] L. Davisson, "Universal noiseless coding," *IEEE Trans. Inform. Theory,* vol. IT-19, pp. 783–795, Nov. 1973.

[4] L. Devroye, *A Course in Density Estimation.* Basel, Switzerland: Birkhäuser-Verlag, 1987.

[5] R. G. Gallager, "Source coding with side information and universal coding," unpublished manuscript, Sept. 1979.

[6] D. Haussler, "A general minimax result for relative entropy," preprint.

[7] ——, "General bounds on the mutual information between a parameter and n conditionally independent observations," preprint.

[8] I. A. Ibragimov and R. Z. Has'minsky, "On the information in a sample about a parameter," in *Proc. 2nd Int. Symp. on Information Theory,* B. N. Petrov and F. Csáki, Eds. Budapest, Hungary: Akademiai Kiado, 1973.

[9] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory,* vol. 41, pp. 714–722, May 1995.

[10] J. Rissanen, "Stochastic complexity and modeling," *Annals Statist.,* vol. 14, pp. 1080–1100, 1986.

[11] J. Rissanen, T. Speed, and B. Yu, "Density estimation by stochastic complexity," *IEEE Trans. Inform. Theory,* vol. 38, pp. 315–323, May 1993.

[12] B. Yu, "Assouad, Fano, and Le Cam," to appear in *Festschrift in Honor of L. Le Cam on His 70th Birthday,* 1995.

[13] B. Yu and T. Speed, "Data compression and histograms," *Probab. Theory Related Fields,* vol. 92, pp. 195–229, 1992.

Error Exponents for Successive Refinement by Partitioning

Angelos Kanlis and Prakash Narayan

***Abstract*—Given a discrete memoryless source (DMS) with probability mass function P, we seek first an asymptotically optimal description of the source with distortion not exceeding Δ_1, followed by an asymptotically optimal refined description with distortion not exceeding $\Delta_2 < \Delta_1$. The rate-distortion function for successive refinement by partitioning, denoted $R(P, \Delta_1, \Delta_2)$, is the overall optimal rate of these descriptions obtained via a two-step coding process. We determine the error exponents for this two-step coding process, namely, the negative normalized asymptotic log likelihoods of the event that the distortion in either step exceeds its prespecified acceptable value, and of the conditional event that the distortion in the second step exceeds the prespecified value given the rate and distortion of the code for the first step. We show that even when the rate-distortion functions for one- and two-step coding coincide, the error exponent in the former case may exceed those in the latter.**

***Index Terms*—Covering Lemma, error exponent, rate-distortion function, successive refinement by partitioning.**

Manuscript received May 17, 1994; revised August 16, 1995. This work of A. Kanlis was supported by the Bodossaki Foundation, Greece, and the Institute for Systems Research at the University of Maryland, College Park, MD, under NSF Grant OIR-85-00108. The work of P. Narayan was supported by the Institute for Systems Research at the University of Maryland, College Park, MD, under NSF Grant OIR-85-00108. The material in this paper was proesented at the 1995 IEEE International Symposium on Information Theory, Whistler, BC, Canada, September 17–22, 1995.

The authors are with the Electrical Engineering Department, Institute for Systems Research, University of Maryland, College Park, MD 20742 USA.

Publisher Item Identifier S 0018-9448(96)00034-X.

I. Introduction

The problem of successive refinement of information by partitioning, also referred to variously as "hierarchical lossy data compression" and "sequential aproximations," has received much attention over the years (Koshélev [9], [10], Equitz-Cover [4], Yamamoto [19], Rimoldi [15]). Related problems include those of "multiple descriptions" (cf. [1], [3], [14], [17], [18], [20]) and that of determining the achievable rate region for cascade communication systems [19]. Given a discrete memoryless source (DMS) with probability mass function (pmf) P, and a suitable distortion measure, suppose that we first seek to describe the source with distortion not exceeding Δ_1. The (asymptotically) minimum rate of coding is, of course, given by the rate distortion function $R(P, \Delta_1)$. Subsequently, if a better (finer) description is required, say with distortion $\Delta_2 < \Delta_1$, additional information at rate ΔR can be provided, so that the overall augmented rate is $R(P, \Delta_1) + \Delta R$. Clearly, $R(P, \Delta_1) + \Delta R \geq R(P, \Delta_2)$. It is of interest to determine the rate-distortion function $R(P, \Delta_1, \Delta_2)$ for this two-step coding process, and find conditions under which it coincides with $R(P, \Delta_2)$.

The condition under which these two rate-distortion functions coincide was determined independently by Koshélev [9] and Equitz-Cover [4], and subsequently given a geometrical interpretation by Rimoldi [15]. This condition requires that the source random variable (rv) and the two reproduction rv's satisfy a Markov property. Rimoldi [15] also provided a complete characterization of the achievable rate region for two-step coding.

In this correspondence, we determine the error exponents for the two-step coding process. It is then shown that even under the Markov condition, when the two rate-distortion functions coincide, the performance of the two-step coding process—as measured by its error exponents—may be inferior to that of one-step coding.

II. Preliminaries

Let $\mathcal{X}$ be a finite set. Let $\{X_t\}_{t=1}^{\infty}$ be a $\mathcal{X}$-valued discrete memoryless source (DMS), i.e., an independent and identically distributed (i.i.d.) process, with (common) probability mass function (pmf) P. Let $\mathcal{Y}_1$ be a finite reproduction alphabet. Let $d_1\colon \mathcal{X} \times \mathcal{Y}_1 \to \boldsymbol{R}^+$ be a nonnegative-valued mapping with $\min_{x\in\mathcal{X}, y\in\mathcal{Y}_1} d_1(x,y) = 0$. This mapping induces a distortion measure on $\mathcal{X}^n \times \mathcal{Y}_1^n$ according to

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum_{t=1}^{n} d_1(x_t, y_t), \quad \boldsymbol{x} \in \mathcal{X}^n,\ \boldsymbol{y} \in \mathcal{Y}_1^n.$$

An n-length block code consists of two mappings: An encoder

$$f_1^{(n)}\colon \mathcal{X}^n \to \mathcal{M}_1 = \{1, \cdots, M_1\}$$

and a decoder

$$\phi_1^{(n)}\colon \mathcal{M}_1 \to \mathcal{Y}_1^n.$$

The rate of this code is $R_1 = \frac{1}{n}\log M_1$. All logarithms and exponentials are with respect to the base 2.

For $R_1 > 0, \Delta_1 > 0$, we say that the pair (R_1, Δ_1) is achievable if for every $\epsilon > 0, \delta > 0$ and n sufficiently large, there exists an n-length block code $(f_1^{(n)}, \phi_1^{(n)})$ of rate not exceeding $R_1 + \delta$ such that

$$\Pr\left\{d_1\left(X^n, \phi_1^{(n)}\left(f_1^{(n)}(X^n)\right)\right) \leq \Delta_1\right\} \geq 1 - \epsilon.$$

The corresponding rate-distortion function, $R(P, \Delta_1)$, characterizing the minimum achievable rate for a distortion Δ_1, is well known and given by

$$R(P, \Delta_1) = \inf_{P_X = P,\ \boldsymbol{E}d_1(X,Y_1)\leq \Delta_1} I(X \wedge Y_1)$$

where $\boldsymbol{E}$ denotes expectation.

Let $\mathcal{Y}_2$ be a (refining) finite reproduction alphabet. A refined description of the source $\{X_t\}_{t=1}^{\infty}$ can be provided by means of a n-length *refining block code* $(f_2^{(n)}, \phi_2^{(n)})$, specified by an encoder

$$f_2^{(n)}\colon \mathcal{X}^n \to \mathcal{M}_2 = \{1, \cdots, M_2\}$$

and a decoder

$$\phi_2^{(n)}\colon \mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{Y}_2^n.$$

The rate of the refining code $(f_2^{(n)}, \phi_2^{(n)})$ is defined as

$$R_2 = \frac{1}{n}\log M_1 M_2.$$

Let $d_2\colon \mathcal{X} \times \mathcal{Y}_2 \to \boldsymbol{R}^+$ be a nonnegative-valued mapping with $\min_{x\in\mathcal{X}, y\in\mathcal{Y}_2} d_2(x,y) = 0$, which induces a distortion measure on $\mathcal{X}^n \times \mathcal{Y}_2^n$ according to

$$d_2(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n}\sum_{t=1}^{n} d_2(x_t, y_t), \quad \boldsymbol{x} \in \mathcal{X}^n,\ \boldsymbol{y} \in \mathcal{Y}_2^n.$$

Definition 1 (Rimoldi [15]): For the DMS $\{X_t\}_{t=1}^{\infty}$ with pmf P and distortion measures d_1, d_2, the quadruple $(R_1, R_2, \Delta_1, \Delta_2)$, $R_1 > 0, R_2 > 0, \Delta_1 > 0, \Delta_2 > 0$, is achievable if for every $\epsilon > 0, \delta > 0$, and n sufficiently large, there exist

- an n-length code $(f_1^{(n)}, \phi_1^{(n)})$ such that

$$\frac{1}{n}\log M_1 \leq R_1 + \delta$$

and

$$\Pr\{d_1(X^n, \phi_1^{(n)}(f_1^{(n)}(X^n))) \leq \Delta_1\} \geq 1 - \epsilon$$

- an n-length refining code $(f_2^{(n)}, \phi_2^{(n)})$ such that

$$\frac{1}{n}\log M_1 M_2 \leq R_2 + \delta$$

and

$$\begin{aligned}\Pr\{&d_1(X^n, \phi_1^{(n)}(f_1^{(n)}(X^n))) \leq \Delta_1,\\ &d_2(X^n, \phi_2^{(n)}(f_1^{(n)}(X^n), f_2^{(n)}(X^n))) \leq \Delta_2\} \geq 1 - \epsilon.\end{aligned}$$

□

A characterization of the set of achievable quadruples $(R_1, R_2, \Delta_1, \Delta_2)$ has been provided independently by Koshélev [10] and Rimoldi [15] and also follows as a combination of the results of El Gamal and Cover [3] and Yamamoto [19].

Theorem 1 ([3], [10], [15], [19]): Consider a DMS $\{X_t\}_{t=1}^{\infty}$ with pmf P and distortion measures d_1, d_2. The quadruple $(R_1, R_2, \Delta_1, \Delta_2)$ is achievable iff there exists a pmf $P_{XY_1Y_2}$ on $\mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2$ with marginal P on $\mathcal{X}$ satisfying the following inequalities:

$$\begin{aligned} I(X \wedge Y_1) &\leq R_1, & I(X \wedge Y_1 Y_2) &\leq R_2\\ \boldsymbol{E}d_1(X, Y_1) &\leq \Delta_1, & \boldsymbol{E}d_2(X, Y_2) &\leq \Delta_2\end{aligned}$$

where (X, Y_1, Y_2) is a $\mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2$-valued rv with pmf $P_{XY_1Y_2}$□.

Suppose that the coding in the first step were done at rate $R_1 > R(P, \Delta_1)$. If the DMS is now to be described in a second step at distortion $\Delta_2 < \Delta_1$, let $R(P, R_1, \Delta_1, \Delta_2)$ be the minimum rate R_2 of the refining code such that $(R_1, R_2, \Delta_1, \Delta_2)$ is achievable. This minimal rate $R(P, R_1, \Delta_1, \Delta_2)$ is characterized by the following Corollary to Theorem 1.

Corollary 1: For $\Delta_1 > 0$, $\Delta_2 > 0$, and $R_1 > R(P, \Delta_1)$, we have

$$R(P, R_1, \Delta_1, \Delta_2) = \inf_{\substack{P_X = P \\ Ed_1(X,Y_1) \le \Delta_1 \\ Ed_2(X,Y_2) \le \Delta_2 \\ I(X \wedge Y_1) \le R_1}} I(X \wedge Y_1 Y_2). \quad (1)$$

□

Suppose next that the coding in the first step at distortion Δ_1 were done optimally, i.e., $R_1 \simeq R(P, \Delta_1)$. If the DMS is now to be described in a second step at distortion $\Delta_2 < \Delta_1$, what is the smallest possible amount of additional information required for this purpose?

The following definition is a special case of Definition 1 for the situation wherein the first step is optimal.

Definition 2: For the DMS $\{X_t\}_{t=1}^{\infty}$ with pmf P, the rate R_2 is (Δ_1, Δ_2)-refinement-achievable if for every $\epsilon > 0, \delta > 0$ and n sufficiently large, there exist

- an n-length code $(f_1^{(n)}, \phi_1^{(n)})$ such that

$$\frac{1}{n} \log M_1 \le R(P, \Delta_1) + \delta$$

and

$$\Pr\left\{d_1(X^n, \phi_1^{(n)}(f_1^{(n)}(X^n))) \le \Delta_1\right\} \ge 1 - \epsilon$$

- an n-length refining code $(f_2^{(n)}, \phi_2^{(n)})$ such that

$$\frac{1}{n} \log M_1 M_2 \le R_2 + \delta$$

and

$$\Pr\{d_1(X^n, \phi_1^{(n)}(f_1^{(n)}(X^n))) \le \Delta_1, \\ d_2(X^n, \phi_2^{(n)}(f_1^{(n)}(X^n), f_2^{(n)}(X^n))) \le \Delta_2\} \ge 1 - \epsilon.$$

□

Let $R(P, \Delta_1, \Delta_2)$ denote the infimum of the set of (Δ_1, Δ_2) refinement-achievable rates. It constitutes the rate-distortion function for the refining code and is given by the following Corollary to Theorem 1.

Corollary 2: For $\Delta_1 > 0, \Delta_2 > 0$, we have

$$\begin{aligned} R(P, \Delta_1, \Delta_2) &= R(P, R(P, \Delta_1), \Delta_1, \Delta_2) \\ &= \inf_{\substack{P_X = P \\ Ed_1(X,Y_1) \le \Delta_1 \\ Ed_2(X,Y_2) \le \Delta_2 \\ I(X \wedge Y_1) = R(P, \Delta_1)}} I(X \wedge Y_1 Y_2). \end{aligned} \quad (2)$$

□

Remarks:

1) For $d_1 = d_2$, $\Delta_1 = \Delta_2$, and $\mathcal{Y}_1 = \mathcal{Y}_2$, we have that $R(P, \Delta_1, \Delta_2) = R(P, \Delta_1)$, the minimum achievable rate for one-step coding.
2) For two-step coding with $\Delta_2 < \Delta_1$, and $R(P, \Delta_2) > R_1 > R(P, \Delta_1)$, clearly

$$R(P, \Delta_2) \le R(P, R_1, \Delta_1, \Delta_2) \le R(P, \Delta_1, \Delta_2).$$

Koshélev [9] has provided a sufficient condition for the inequalities above to hold as equalities. Cover and Equitz [4] have independently shown this condition to be both necessary and sufficient (see Theorem 2 below).

3) It follows from the observation of Equitz and Cover [4, p. 271] in the context of Gray's work [7] on conditional rate-distortion function that *if* $\{Y_{1,t}\}_{t=1}^{\infty}$ is an i.i.d. process, then

$$R(P, \Delta_1, \Delta_2) = R(P, \Delta_1) + R_{X|Y_1}(\Delta_2)$$

where the conditional rate-distortion function $R_{X|Y_1}(\Delta_2)$ [7] is given by

$$R_{X|Y_1}(\Delta_2) = \inf_{P_{Y_2|XY_1}: \sum_{x,y_1,y_2} P_{Y_2|XY_1}(y_2|x,y_1) P_{XY_1}(x,y_1) d_2(x,y_2) \le \Delta_2} I(X \wedge Y_2 \mid Y).$$

However, $R_{X|Y_1}(\Delta_2)$ is defined only when $\{Y_{1,t}\}_{t=1}^{\infty}$ is an i.i.d. process, which, of course, does not hold in general.

Theorem 2 (Cover and Equitz [4]): For the DMS $\{X_t\}_{t=1}^{\infty}$ with pmf P, given distortion measures $d_1 = d_2 = d$, reproduction alphabet $\mathcal{Y}_1 = \mathcal{Y}_2 = \mathcal{Y}$, and for $R(P, \Delta_2) > R_1 > R(P, \Delta_1)$, we have

$$R(P, \Delta_1, \Delta_2) = R(P, R_1, \Delta_1, \Delta_2) = R(P, \Delta_2)$$

iff there exists a pmf $P_{XY_1Y_2}$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$ with

$$\begin{aligned} P_X &= P, \\ R(P, \Delta_1) &= I(X \wedge Y_1), \quad Ed(X, Y_1) \le \Delta_1 \\ R(P, \Delta_2) &= I(X \wedge Y_2), \quad Ed(X, Y_2) \le \Delta_2 \end{aligned}$$

where (X, Y_1, Y_2) is a $\mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$-valued rv with pmf $P_{XY_1Y_2}$. Furthermore, $P_{XY_1Y_2}$ must satisfy a Markov condition, namely

$$P_{XY_1Y_2}(x, y_1, y_2) = P_X(x) P_{Y_2|X}(y_2 \mid x) P_{Y_1|Y_2}(y_1 \mid y_2), \\ x \in \mathcal{X},\ y_1, y_2 \in \mathcal{Y} \quad (3)$$

i.e., (X, Y_2, Y_1) form a Markov chain. □

III. The Error Exponents

In this section, we shall characterize the error exponents for successive refinement by partitioning. Corollaries 2 and 1 imply that for $\Delta_2 \le \Delta_1$ and for numbers R_1, R_2 with $R_1 < R_2$ and $R_1 > R(P, \Delta_1)$, $R_2 > R(P, R_1, \Delta_1, \Delta_2)$, there exists a sequence of n-length block codes $(f_1^{(n)}, \phi_1^{(n)})$, $(f_2^{(n)}, \phi_2^{(n)})$, such that

$$\lim_n \frac{1}{n} \log \|f_1^{(n)}\| = R_1 \quad (4)$$

$$\lim_n \frac{1}{n} \log \|f_1^{(n)}\| \|f_2^{(n)}\| = R_2 \quad (5)$$

and

$$\lim_n \Pr(d_1(X^n, \phi_1(f_1(X^n))) > \Delta_1 \text{ or } \\ d_2(X^n, \phi_2(f_1(X^n), f_2(X^n))) > \Delta_2) = 0$$

where $\|f_1^{(n)}\|$ (resp., $\|f_2^{(n)}\|$) denotes the cardinality of the domain of $f_1^{(n)}$ (resp., $f_2^{(n)}$). Our objective is to characterize the rate of convergence to zero of the previous probability.

Definition 3: For given distortion measures d_1, d_2, positive numbers $\Delta_1 > \Delta_2 > 0$, n-length block code (f_1, ϕ_1), and n-length refining block code (f_2, ϕ_2), we define the error function $e(P, (f_1, \phi_1), (f_2, \phi_2), \Delta_1, \Delta_2)$ as the probability that the source sequence X^n in $\mathcal{X}^n$ of the DMS with (common) pmf P is not reproduced within distortion Δ_1 in the first step of coding, or within distortion Δ_2 by the refining code. Thus the error function for two-step coding is defined as

$$e(P, (f_1, \phi_1), (f_2, \phi_2), \Delta_1, \Delta_2) \\ \triangleq \Pr(d_1(X^n, \phi_1(f_1(X^n))) > \Delta_1 \text{ or } \\ d_2(X^n, \phi_2(f_1(X^n), f_2(X^n))) > \Delta_2). \quad (6)$$

□

Definition 4: The conditional error function $e(P,(f_1,\phi_1),(f_2,\phi_2),\Delta_2 \mid R_1,\Delta_1)$ is the probability that the source sequence X^n in $\mathcal{X}^n$ is not reproduced within distortion Δ_2 by the refining code, given that (f_1,ϕ_1), i.e., the code for the first step, has rate that does not exceed R_1 and distortion that does not exceed Δ_1. Thus

$$e(P,(f_1,\phi_1),(f_2,\phi_2),\Delta_2 \mid R_1,\Delta_1) \triangleq \Pr(d_2(X^n,\phi_2(f_1(X^n),f_2(X^n))) > \Delta_2). \tag{7}$$

□

We show below for suitable two-step n-length block codes with rates converging to R_1 and R_2, that $e(P,(f_1^{(n)},\phi_1^{(n)}),(f_2^{(n)},\phi_2^{(n)}),\Delta_1,\Delta_2)$ converges to zero exponentially rapidly with rate given by the error exponent

$$F(P,R_1,R_2,\Delta_1,\Delta_2) \triangleq \inf_{\substack{Q:\ R(Q,\Delta_1)>R_1 \\ \text{or } R(Q,R_1,\Delta_1,\Delta_2)>R_2}} D(Q\,\|\,P) \tag{8}$$

provided $R_1 > R(P,\Delta_1)$ and $R_2 > R(P,R_1,\Delta_1,\Delta_2)$. This provides an extension of the result of Marton [11] (cf. also Csiszár and Körner [2]) on the error exponent for one-step coding. Our approach is along the lines of Csiszár and Körner [2, sec. 2.4].

It follows from [6] that upon setting $d_1 = d_2, \Delta_1 = \Delta_2, \mathcal{Y}_1 = \mathcal{Y}_2$ and $\phi_2(m_1,m_2) = \phi_1(m_1)$, $m_1 \in \mathcal{M}_1, m_2 \in \mathcal{M}_2$ that

$$e(P,(f_1,\phi_1),(f_2,\phi_2),\Delta_1,\Delta_2) = e(P,(f_1,\phi_1),\Delta_1)$$

where the term on the right side above corresponds to the error function for one-step coding which converges to zero exponentially rapidly with rate given by the error exponent

$$F(P,R_1,\Delta_1) \triangleq \inf_{Q:\ R(Q,\Delta_1)>R_1} D(Q\,\|\,P) \tag{9}$$

provided $R_1 > R(P,\Delta_1)$ (cf. Marton [11], Csiszár and Körner [2]).

We further show that for suitable two-step n-length block codes with rates converging to R_1 and R_2, that $e(P,(f_1,\phi_1),(f_2,\phi_2),\Delta_2 \mid R_1,\Delta_1)$, i.e., the conditional error function converges to zero exponentially rapidly with rate given by the error exponent

$$F_c(P,R_2,\Delta_2 \mid R_1,\Delta_1) \triangleq \inf_{Q:\ R(Q,R_1,\Delta_1,\Delta_2)>R_2} D(Q\,\|\,P).$$

Theorem 3 (Two-Step Coding Error Exponent): Let $\{X_t\}_{t=1}^\infty$ be a DMS with pmf P. For every $R_1 < R_2 < \log|\mathcal{X}|$ and distortion measures d_1 on $\mathcal{X}\times\mathcal{Y}_1$ and d_2 on $\mathcal{X}\times\mathcal{Y}_2$, there exists a sequence of n-length block codes for two-step coding such that

- $$\lim_n \frac{1}{n}\log\|f_1^{(n)}\| = R_1$$

  and

  $$\lim_n \frac{1}{n}\log\|f_1^{(n)}\|\,\|f_2^{(n)}\| = R_2 \tag{10}$$

- for every pmf P on $\mathcal{X}$, $\Delta_1 > \Delta_2 \geq 0$, and $\delta_1 > 0, \delta_2 > 0, \delta_3 > 0$

  $$\frac{1}{n}\log e\big(P,(f_1^{(n)},\phi_1^{(n)}),\Delta_1\big) \leq -F(P,R_1,\Delta_1) + \delta_1$$

  $$\frac{1}{n}\log e\big(P,(f_1^{(n)},\phi_1^{(n)}),(f_2^{(n)},\phi_2^{(n)}),\Delta_1,\Delta_2\big) \leq -F(P,R_1,R_2,\Delta_1,\Delta_2) + \delta_2$$

  and

  $$\frac{1}{n}\log e\big(P,(f_1^{(n)},\phi_1^{(n)}),(f_2^{(n)},\phi_2^{(n)}),\Delta_2 \mid R_1,\Delta_1\big) \leq -F_c(P,R_2,\Delta_2 \mid R_1,\Delta_1) + \delta_3$$

  whenever $n \geq N(|\mathcal{X}|,d_1,d_2,\delta_1,\delta_2,\delta_3)$.

Further, for every sequence of codes satisfying (10) and every distribution P on $\mathcal{X}$, the following hold:

$$\liminf_{n\to\infty}\frac{1}{n}\log e\big(P,(f_1^{(n)},\phi_1^{(n)}),\Delta_1\big) \geq -F(P,R_1,\Delta_1)$$

$$\liminf_{n\to\infty}\frac{1}{n}\log e\big(P,(f_1^{(n)},\phi_1^{(n)}),(f_2^{(n)},\phi_2^{(n)}),\Delta_1,\Delta_2\big) \geq -F(P,R_1,R_2,\Delta_1,\Delta_2)$$

and

$$\liminf_{n\to\infty}\frac{1}{n}\log e\big(P,(f_1^{(n)},\phi_1^{(n)}),(f_2^{(n)},\phi_2^{(n)}),\Delta_2 \mid R_1,\Delta_1\big) \geq -F_c(P,R_2,\Delta_2 \mid R_1,\Delta_1).$$

□

Remark: Note that the two-step code in the statement of the forward part of Theorem 3 does not rely on a knowledge of the pmf P of the DMS. Hence, this code is *universal* in that it is applicable to any $\mathcal{X}$-valued DMS.

As is to be expected, the error exponent for two-step coding cannot exceed that for one-step coding. This is obvious from (8) by observing for $d_1 = d_2 = d$ and $\mathcal{Y}_1 = \mathcal{Y}_2 = \mathcal{Y}$, that

$$F(P,R_1,R_2,\Delta_1,\Delta_2) = \min\{F(P,R_1,\Delta_1), F_c(P,R_2,\Delta_2 \mid R_1,\Delta_1)\}. \tag{11}$$

Even with the Markov condition in effect so that the rate-distortion functions for one- and two-step coding coincide, i.e.

$$R(P,R_1,\Delta_1,\Delta_2) = R(P,\Delta_2), \quad \Delta_2 < \Delta_1$$

note that if

$$F(P,R_1,\Delta_1) < F_c(P,R_2,\Delta_2 \mid R_1,\Delta_1) \tag{12}$$

then

$$F(P,R_1,R_2,\Delta_1,\Delta_2) < F(P,R_2,\Delta_2). \tag{13}$$

This is illustrated in the following example.

Example: Let $\{X_t\}_{t=1}^\infty$ be a DMS with pmf P, where $P(x) > 0$, $x \in \mathcal{X}$. Let $\mathcal{Y}_1 = \mathcal{Y}_2 = \mathcal{Y}$, and $d_1 = d_2 = d$ where d denotes Hamming distortion measure. Thus

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

It was shown in [4] that this setup admits the Markov condition so that

$$R(P,R_1,\Delta_1,\Delta_2) = R(P,\Delta_2).$$

Next, note that

$$R(P,R_1,\Delta_1,\Delta_2) = R(P,\Delta_2) = H(P) - \Delta_2\log(|\mathcal{X}|-1) - h_b(\Delta_2), \quad \Delta_1 \geq \Delta_2$$

where $H(P)$ is the entropy of P. It then follows from (9) and (8) that the exponents for one- and two-step coding are

$$F(P,R_2,\Delta_2) = \min_{Q:\ H(Q)\geq c(R_2,\Delta_2)} D(Q\,\|\,P) \tag{14}$$

$$F(P,R_1,R_2,\Delta_1,\Delta_2) = \min_{Q:\ H(Q)\geq\max\{c(R_1,\Delta_1),c(R_2,\Delta_2)\}} D(Q\,\|\,P) \tag{15}$$

where

$$c(R,\Delta) \triangleq \Delta\log(|\mathcal{X}|-1) + h_b(\Delta) + R.$$

Let

$$\mathcal{Q}_\alpha \triangleq \{Q\colon H(Q) > \alpha\}, \quad \alpha \geq 0.$$

Now, the necessary and sufficient condition for (12) and (13) to hold can be expressed as $\mathcal{Q}_{c(R_1,\Delta_1)} \supset \mathcal{Q}_{c(R_2,\Delta_2)}$, or, equivalently

$$c(R_1,\Delta_1) < c(R_2,\Delta_2)$$

which is the same as

$$R(P,\Delta_1) < R_1 < R_2 - [R(P,\Delta_2) - R(P,\Delta_1)].$$

The condition above says that, with the Markov condition in effect, the error exponent (15) for two-step coding is worse than that for one-step coding iff

$$R_2 - R(P,\Delta_2) > R_1 - R(P,\Delta_1). \tag{16}$$

Note that this condition is given by (14) and (15) as $D(\mathcal{Q}_2 \,\|\, P) < D(\mathcal{Q}_1 \,\|\, P)$, where

$$D(\mathcal{Q}_i \,\|\, P) \triangleq \min_{Q:\ H(Q) \geq c(R_i,\Delta_i)} D(Q \,\|\, P), \quad i = 1,2.$$

This means that *in the case of the Hamming distortion measures* (divergence) distances in the space of pmf's on $\mathcal{X}$ correspond to distances (differences) between the actual rates and the corresponding values of the rate-distortion functions.

Clearly, if R_1, R_2 are chosen to violate inequality (16), e.g., with R_1 large enough, the two-step coding process will no longer suffer from the disadvantage of a smaller error exponent.

The proof of Theorem 3 relies on the following Covering Lemma for two-step coding, which is a straightforward extension of the corresponding lemma for one-step coding [2, Lemma 2.4.1].

Let $P = P^{(n)}$ be a *type* on $\mathcal{X}$ (cf. e.g., [2]), i.e., a pmf with rational probabilities with (common) denominator n. Let $\mathcal{T}_P^n$ denote the set of sequences in $\mathcal{X}^n$ of (common) type P.

Lemma 1: For distortion measures d_1 on $\mathcal{X} \times \mathcal{Y}_1$ and d_2 on $\mathcal{X} \times \mathcal{Y}_2$, type $P = P^{(n)}$ on $\mathcal{X}$ and numbers $\Delta_1 > \Delta_2 > 0$, $R_1 > R(P,\Delta_1)$, $\delta_1 > \delta_2 > 0$, there exist

- a set $\mathcal{B}_1 \subset \mathcal{Y}_1^n$, such that

$$\frac{1}{n}\log|\mathcal{B}_1| \leq R_1 + \delta_1 \tag{17}$$

and $\mathcal{B}_1$ covers $\mathcal{T}_P^n$

$$\bigcup_{\boldsymbol{y}_1 \in \mathcal{B}_1} \mathcal{N}_1(\boldsymbol{y}_1) = \mathcal{T}_P^n$$

where

$$\mathcal{N}_1(\boldsymbol{y}_1) = \{\boldsymbol{x} \in \mathcal{T}_P^n\colon d_1(\boldsymbol{x},\boldsymbol{y}_1) \leq \Delta_1\}, \quad \boldsymbol{y}_1 \in \mathcal{Y}_1^n;$$

- sets $\mathcal{B}_2(\boldsymbol{y}_1) \subset \mathcal{Y}_2^n$, $\boldsymbol{y}_1 \in \mathcal{B}_1$, such that

$$\frac{1}{n}\log\left(\sum_{\boldsymbol{y}_1 \in \mathcal{B}_1} |\mathcal{B}_2(\boldsymbol{y}_1)|\right) \leq R(P,R_1,\Delta_1,\Delta_2) + \delta_2 \tag{18}$$

and $\mathcal{B}_2(\boldsymbol{y}_1)$ covers $\mathcal{N}_1(\boldsymbol{y}_1)$, i.e.

$$\bigcup_{\boldsymbol{y}_2 \in \mathcal{B}_2(\boldsymbol{y}_1)} \mathcal{N}_2(\boldsymbol{y}_2) \supset \mathcal{N}_1(\boldsymbol{y}_1), \quad \boldsymbol{y}_1 \in \mathcal{B}_1,$$

where

$$\mathcal{N}_2(\boldsymbol{y}_2) = \{\boldsymbol{x} \in \mathcal{T}_P^n\colon d_2(\boldsymbol{x},\boldsymbol{y}_2) \leq \Delta_2\}, \quad \boldsymbol{y}_2 \in \mathcal{Y}_2^n$$

provided that $n \geq N(d_1,d_2,\delta_1,\delta_2)$.

Proof of Theorem 3: We commence with the existence part of the proof. It is convenient to define the following quantities: For a pmf Q on $\mathcal{X}$ and $R_2 > R_1 > 0$, let

$$\Delta(Q,R_1) \triangleq \inf_{P_X = Q,\, I(X \wedge Y_1) \leq R_1} \boldsymbol{E}d_1(X,Y_1)$$

and

$$\Delta(Q,R_1,\Delta_1,R_2) \triangleq \inf_{\substack{P_X = Q \\ I(X \wedge Y_1) \leq R_1 \\ I(X \wedge Y_1 Y_2) \leq R_2 \\ \boldsymbol{E}d_1(X,Y_1) \leq \Delta_1}} \boldsymbol{E}d_2(X,Y_2).$$

Consider the sets

$$\mathcal{U}_1^{(n)} \triangleq \bigcup_{Q:\ R(Q,\Delta_1) > R_1} \mathcal{T}_Q^n$$

and

$$\mathcal{U}_2^{(n)} \triangleq \mathcal{U}_1^{(n)} \bigcup_{Q:\ R(Q,R_1,\Delta_1,\Delta_2) > R_2} \mathcal{T}_Q^n.$$

Obviously

$$\begin{aligned} P^n\left(\mathcal{U}_1^{(n)}\right) &\leq (n+1)^{|\mathcal{X}|}\exp\{-nF(P,R_1,\Delta_1)\} \\ &\leq \exp\{-n[F(P,R_1,\Delta_1) - \delta_1]\} \end{aligned}$$

for all n large (cf. [2, Lemma 1.2.6]), and

$$\begin{aligned} P^n\left(\mathcal{U}_2^{(n)}\right) &\leq (n+1)^{|\mathcal{X}|}\exp\{-nF_c(P,R_2,\Delta_2 \mid R_1,\Delta_1)\} \\ &\leq \exp\{-n[F_c(P,R_2,\Delta_2 \mid R_1,\Delta_1) - \delta_2]\} \end{aligned}$$

for all n large. Next, by Lemma 1, there exist sequences $\epsilon_1^{(n)}$ and $\epsilon_2^{(n)}$ with $\lim_n \epsilon_1^{(n)} = 0$ and $\lim_n \epsilon_2^{(n)} = 0$, such that for every type Q of sequences in $\mathcal{X}^n$, there exist sets $\mathcal{B}_{Q,1} \subset \mathcal{Y}_1^n$ and $\mathcal{B}_{Q,2}(\boldsymbol{y}_1) \subset \mathcal{Y}_2^n, \boldsymbol{y}_1 \in \mathcal{B}_{Q,1}$, satisfying

$$\frac{1}{n}\log|\mathcal{B}_{Q,1}| \leq R_1 + \epsilon_1^{(n)}$$

$$\frac{1}{n}\log\left(\sum_{\boldsymbol{y}_1 \in \mathcal{B}_{Q,1}} |\mathcal{B}_{Q,2}(\boldsymbol{y}_1)|\right) \leq R_2 + \epsilon_2^{(n)}$$

and for every $\boldsymbol{x} \in \mathcal{T}_Q^k$

$$d_1(\boldsymbol{x},\mathcal{B}_{Q,1}) \triangleq \min_{\boldsymbol{y}_1 \in \mathcal{B}_{Q,1}} d_1(\boldsymbol{x},\boldsymbol{y}_1) \leq \Delta(Q,R_1)$$

and

$$\begin{aligned} &d_2(\boldsymbol{x},\mathcal{B}_{Q,2}(\boldsymbol{y}_1)) \\ &\quad \triangleq \min_{\boldsymbol{y}_2 \in \mathcal{B}_{Q,2}(\boldsymbol{y}_1)} d_2(\boldsymbol{x},\boldsymbol{y}_2) \leq \Delta(Q,R_1,\Delta_1,R_2),\ \boldsymbol{y}_1 \in \mathcal{B}_{Q,1} \end{aligned}$$

for $\boldsymbol{x}$ such that $d_1(\boldsymbol{x},\boldsymbol{y}_1) \leq \Delta_1$. Next, we set

$$\mathcal{B}_1 \triangleq \bigcup_Q \mathcal{B}_{Q,1}$$

$$\mathcal{B}_2(\boldsymbol{y}_1) \triangleq \mathcal{B}_{Q_{\boldsymbol{y}_1},2}(\boldsymbol{y}_1), \quad \boldsymbol{y}_1 \in \mathcal{B}_1,$$

where $Q_{\boldsymbol{y}_1}$ is the type of $\boldsymbol{y}_1$. Then, the Type Counting Lemma [2, Lemma 1.2.2] yields that

$$\frac{1}{n}\log|\mathcal{B}_1| \leq R_1 + \eta_1^{(n)}$$

$$\frac{1}{n}\log\left(\sum_{\boldsymbol{y}_1 \in \mathcal{B}_1} |\mathcal{B}_2(\boldsymbol{y}_1)|\right) \leq R_2 + \eta_2^{(n)}$$

where $\lim_n \eta_1^{(n)} = \lim_n \eta_2^{(n)} = 0$. Furthermore, $R(Q,\Delta_1) \leq R_1$ implies in a standard manner that

$$\Delta(Q,R_1) \leq \Delta_1 \tag{19}$$

and $R(Q, R_1, \Delta_1, \Delta_2) \leq R_2$ implies

$$\Delta(Q, R_1, \Delta_\perp, R_2) \leq \Delta_2 \tag{20}$$

which is seen as follows.

Let $P^*_{XY_1Y_2}$ achieve $R(Q, R_1, \Delta_1, \Delta_2) = I(X \wedge Y_1Y_2) \leq R_2$, where the rv (X, Y_1, Y_2) is distributed according to $P^*_{XY_1Y_2}$. Then

$$\boldsymbol{E}d_1(X, Y_1) \leq \Delta_1, \boldsymbol{E}d_2(X, Y_2) \leq \Delta_2, I(X \wedge Y_1) \leq R_1.$$

Hence, $\Delta(Q, R_1, \Delta_\perp, R_2) \leq \boldsymbol{E}d(X, Y_2) \leq \Delta_2$, since

$$\begin{aligned} I(X \wedge Y_1) \leq R_1, I(X \wedge Y_1Y_2) \leq R_2 \\ \boldsymbol{E}d_1(X, Y_1) \leq \Delta_1 \leq \Delta(Q, R_1). \end{aligned}$$

As a consequence of (19) and (20), we have

$$\begin{aligned} d_1(\boldsymbol{x}, \mathcal{B}_1) &\leq \Delta_1, \quad \boldsymbol{x} \in \mathcal{X}^n \setminus \mathcal{U}_1^{(n)} \\ d_2(\boldsymbol{x}, \mathcal{B}_2(\boldsymbol{y}_1)) &\leq \Delta_2, \quad \boldsymbol{y}_1 \in \mathcal{B}_1, \boldsymbol{x} \in \mathcal{X}^n \setminus \mathcal{U}_2^{(n)}. \end{aligned}$$

These inequalities establish the direct (existence) part of the proof.

Turning to the converse part, consider any pmf Q on $\mathcal{X}$ with the property that

$$R(Q, R_1, \Delta_1, \Delta_2) > R_2 + \delta_2. \tag{21}$$

The remainder of the proof relies on the following Claim which constitutes a strong converse to Theorem 1.

Claim 1: For any sequence of codes $(f_1^{(n)}, \phi_1^{(n)})$, $(f_2^{(n)}, \phi_2^{(n)})$ with

$$\begin{aligned} \frac{1}{n} \log \|f_1^{(n)}\| &\leq R_1 + \delta_1 \\ \frac{1}{n} \log \|f_1^{(n)}\| \|f_2^{(n)}\| &\leq R_2 + \delta_2 \end{aligned}$$

for all n large, (21) implies

$$e\big(Q, (f_1^{(n)}, \phi_1^{(n)}), (f_2^{(n)}, \phi_2^{(n)}), \Delta_2 \mid R_1, \Delta_1\big) \geq \frac{1}{2}$$

whenever $n \geq N(d_1, d_2, \delta_1, \delta_2)$.

The proof of the Claim is obtained by mimicking the proof of [2, Theorem 2.2.3], and is, therefore, omitted.

Then, by [2. Corollary 1.1.2] it holds for all n large enough that

$$\begin{aligned} e\big(P, (f_1^{(n)}, \phi_1^{(n)}), (f_2^{(n)}, \phi_2^{(n)}), \Delta_2 \mid R_1, \Delta_1\big) \\ \geq \exp\{-n[D(Q \parallel P) + \delta_1]\}. \end{aligned}$$

Note that the case $R(Q, \Delta_1) > R_1 + \delta_1$ alone was shown ([2, Theorem 2.4.5]) to imply that

$$e\big(P, (f_1^{(n)}, \phi_1^{(n)}), \Delta_1\big) \geq \exp\{-n[D(Q \parallel P) + \delta_1]\}.$$

Since δ_1, δ_2 were arbitrary as was Q subject to (21), the desired converse follows.

It is obvious that, since

$$\begin{aligned} \max\{e\big(P, (f_1^{(n)}, \phi_1^{(n)}), \Delta_1\big), e\big(P, (f_1^{(n)}, \phi_1^{(n)}), \\ (f_2^{(n)}, \phi_2^{(n)}), \Delta_2 \mid R_1, \Delta_1\big)\} \\ \leq e\big(P, (f_1^{(n)}, \phi_1^{(n)}), (f_2^{(n)}, \phi_2^{(n)}), \Delta_2 \mid R_1, \Delta_1\big) \\ \leq e\big(P, (f_1^{(n)}, \phi_1^{(n)}), \Delta_1\big) + e\big(P, (f_1^{(n)}, \phi_1^{(n)}), \\ (f_2^{(n)}, \phi_2^{(n)}), \Delta_2 \mid R_1, \Delta_1\big), \end{aligned}$$

the overall error exponent will be the greater of the two. Hence, (8) follows. □

IV. Discussion

We have determined the error exponent for the problem of successive refinement by partitioning for a DMS. As expected, it is generally smaller than the error exponent for one-step coding. It is interesting to note that even when the Markov condition (3) holds so that the rate-distortion functions coincide for one-step and two-step coding, it may hold that the error exponents for the latter are strictly smaller than that for the former.

For a DMS with Hamming distance distortion measure, a simple necessary and sufficient condition for the error exponents to differ in the presence of Markov condition, can be expressed in terms of the coding rates R_1, R_2 and the one-step rate-distortion function. An extension of this result to arbitrary distortion measures remains unresolved.

Finally, it can be shown that the error function for two-step coding (cf. (6)) goes to one exponentially fast, whenever $R_2 < R(P, \Delta_1, \Delta_2)$. To this end, we conclude from Rimoldi ([15, Theorem 1, converse part]) that if $R_2 < (P, R_1, \Delta_1, \Delta_2)$, then for every sequence of n-length block codes satisfying (4) and (5), we have

$$\lim_n e\big(P, (f_1^{(n)}, \phi_1^{(n)}), (f_2^{(n)}, \phi_2^{(n)}), \Delta_1, \Delta_2\big) = 1.$$

Indeed, in analogy with one-step coding (cf. Csiszár and Körner [2, sec. 2.4]), this convergence occurs exponentially fast. To see this, define

$$\begin{aligned} e^{(n)}&(P, R_1, R_2, \Delta_1, \Delta_2) \\ &\triangleq \min e\big(P, (f_1^{(n)}, \phi_1^{(n)}), (f_2^{(n)}, \phi_2^{(n)}), \Delta_1, \Delta_2\big) \end{aligned}$$

where the minimum is taken over all codes satisfying (4) with $R_1 > R(P, \Delta_1)$ and

$$\frac{1}{n} \log \|f_1^{(n)}\| \|f_2^{(n)}\| \leq R_2$$

where $0 < R_2 < R(P, \Delta_1, \Delta_2)$. Then, it holds that

$$\begin{aligned} \lim_n \left\{ -\frac{1}{n} \log \left[1 - e^{(n)}(P, R_1, R_2, \Delta_1, \Delta_2)\right] \right\} \\ = G(R_1, R_2, \Delta_1, \Delta_2) \end{aligned}$$

where

$$G(R_1, R_2, \Delta_1, \Delta_2) = \inf_Q \big(D(Q \parallel P) + |R(Q, R_1 \Delta_1, \Delta_2) - R_2|^+\big)$$

with $|x|^+$ denoting $\max\{x, 0\}$. The proof is identical to that for one-step coding (cf. eg., Csiszár and Körner [2, problem 2.4.6, pp. 158–159]) and relies on the strong converse (Rimoldi, [15, Theorem 1, converse part]).

Appendix
Proof of Lemma 1

Let $P = P^{(n)}$ be a fixed but arbitrary type of sequences in $\mathcal{X}^n$. Let $\mathcal{U}(\mathcal{B}_1)$ and $\mathcal{U}(\mathcal{B}_2(\boldsymbol{y}_1))$, respectively, denote the subsets of $\mathcal{T}_P^n$ not covered by $\mathcal{B}_1$ and $\mathcal{B}_2(\boldsymbol{y}_1)$, i.e.

$$\mathcal{U}(\mathcal{B}_1) = \mathcal{T}_P^n \setminus \bigcup_{\boldsymbol{y}_1 \in \mathcal{B}_1} \mathcal{N}_1(\boldsymbol{y}_1) = \mathcal{T}_P^n \bigcap \left(\bigcap_{\boldsymbol{y}_1 \in \mathcal{B}_1} \mathcal{N}_1(\boldsymbol{y}_1)^c \right)$$

and

$$\begin{aligned} \mathcal{U}(\mathcal{B}_2(\boldsymbol{y}_1)) &= \mathcal{N}_1(\boldsymbol{y}_1) \setminus \bigcup_{\boldsymbol{y}_2 \in \mathcal{B}_2(\boldsymbol{y}_1)} \mathcal{N}_2(\boldsymbol{y}_2) \\ &= \mathcal{N}_1(\boldsymbol{y}_1) \bigcap \left(\bigcap_{\boldsymbol{y}_2 \in \mathcal{B}_2(\boldsymbol{y}_1)} \mathcal{N}_2(\boldsymbol{y}_2)^c \right), \quad \boldsymbol{y}_1 \in \mathcal{B}_1. \end{aligned}$$

Consider a $\mathcal{X} \times \mathcal{Y}_1 \times \mathcal{Y}_2$-valued RV (X, Y_1, Y_2) with $P_X = P$ and

$$\begin{aligned} \boldsymbol{E}d_1(X, Y_1) &= |\Delta_1 - \eta_1|^+ \\ \boldsymbol{E}d_2(X, Y_2) &\le |\Delta_2 - \eta_2|^+ \end{aligned} \tag{22}$$

where $\eta_1 > 0, \eta_2 > 0$ will be specified later.

Let $\mathcal{T}^n_{[Y_1]}$ (resp., $\mathcal{T}^n_{[Y_2|Y_1]}(\boldsymbol{y}_1)$) denote the set of Y_1-typical sequences $\boldsymbol{y}_1 \in \mathcal{Y}_1^n$ (resp., $Y_2 \mid Y_1$-typical sequences $\boldsymbol{y}_2 \in \mathcal{Y}_2^n$ with respect to $\boldsymbol{y}_1 \in \mathcal{Y}_1^n$) [2, Definitions 1.2.8, 1.2.9, pp. 33–34]. Let $\mathcal{F}_m$ denote the set of all (not necessarily distinct) collections of m elements of $\mathcal{T}^n_{[Y_1]}$. Similarly, for $i = 1, \cdots, m$, let $\mathcal{G}_{n_i}(\boldsymbol{y}_1)$ denote the set of all (not necessarily distinct) collections of n_i elements of $\mathcal{T}^n_{[Y_2|Y_1]}(\boldsymbol{y}_1)$.

Let $Z^m = (Z_1, \cdots, Z_m)$ be an RV uniformly distributed on $\mathcal{F}_m$. Next, for $i = 1, \cdots, m$, given that $Z_i = z_i$, let $W_i^{n_i} = (W_{i1}, \cdots, W_{in_i})$, be an RV distributed (conditionally) uniformly on $\mathcal{G}_{n_i}(z_i)$. In other words, Z_i's are i.i.d. with

$$\Pr\{Z_i = z_i\} = \frac{\mathbf{1}(z_i \in \mathcal{T}^n_{[Y_1]})}{|\mathcal{T}^n_{[Y_1]}|}, \quad i = 1, \cdots, m$$

and the W_{ij}'s are conditionally i.i.d., with

$$\Pr\{W_{ij} = \boldsymbol{w}_{ij} \mid Z_i = z_i\} = \frac{\mathbf{1}(\boldsymbol{w}_{ij} \in \mathcal{T}^n_{[Y_2|Y_1]}(z_i))}{|\mathcal{T}^n_{[Y_2|Y_1]}(z_i)|}, \quad j = 1, \cdots, n_i,\ i = 1, \cdots, m.$$

We must show the existence of sets $\mathcal{B}_1 \subset \mathcal{Y}_1^n$ and $\mathcal{B}_2(\boldsymbol{y}_1) \subset \mathcal{Y}_2^n, \boldsymbol{y}_1 \in \mathcal{B}_1$, such that

$$\left| \mathcal{U}(\mathcal{B}_1) \bigcup \left(\bigcup_{\boldsymbol{y}_1 \in \mathcal{B}_1} \mathcal{U}(\mathcal{B}_2(\boldsymbol{y}_1)) \right) \right| = 0.$$

To this end, it suffices to show that

$$\boldsymbol{E}\left| \mathcal{U}(Z^m) \bigcup \left(\bigcup_{i=1}^m \mathcal{U}(W_i^{n_i}) \right) \right| < 1.$$

Now

$$\begin{aligned} &\boldsymbol{E}\left| \mathcal{U}(Z^m) \bigcup \left(\bigcup_{i=1}^m \mathcal{U}(W_i^{n_i}) \right) \right| \\ &= \sum_{\boldsymbol{x} \in \mathcal{T}_P^n} \Pr\left\{ \boldsymbol{x} \in \mathcal{U}(Z^m) \bigcup \left(\bigcup_{i=1}^m \mathcal{U}(W_i^{n_i}) \right) \right\} \\ &\le \sum_{\boldsymbol{x} \in \mathcal{T}_P^n} \Pr\{\boldsymbol{x} \in \mathcal{U}(Z^m)\} + \sum_{\boldsymbol{x} \in \mathcal{T}_P^n} \sum_{i=1}^m \Pr\{\boldsymbol{x} \in \mathcal{U}(W_i^{n_i})\}. \& \end{aligned} \tag{23}$$

We consider first the first term on the right side of the inequality above. By choosing

$$\exp\left[n\left(I(X \wedge Y_1) + \frac{2}{3}\delta_1\right)\right] \le m \le \exp\left[n\left(I(X \wedge Y_1) + \frac{3}{4}\delta_1\right)\right]$$

we get that

$$\sum_{\boldsymbol{x} \in \mathcal{T}_P^n} \Pr\{\boldsymbol{x} \in \mathcal{U}(Z^m)\} \le \exp\left[n \log|\mathcal{X}| - \exp\frac{n\delta_1}{6}\right] < \frac{1}{2} \tag{24}$$

for all $n \ge N_1(\delta_1, d_1, \eta_1)$ (cf. e.g., [2, p. 152]). Next, to bound the second term on the right side of (23), for each $\boldsymbol{x} \in \mathcal{T}_P^n$, let

$$\mathcal{A}(\boldsymbol{x}) \triangleq \{z \in \mathcal{T}^n_{[Y_1]}\colon \boldsymbol{x} \in \mathcal{N}_1(z)\} = \{z \in \mathcal{T}^n_{[Y_1]}\colon d_1(\boldsymbol{x}, z) \le \Delta_1\}.$$

Assume for the time being, the following

Claim 2: For any $\delta' > 0$, we have

$$|\mathcal{A}(\boldsymbol{x})| \le \exp\left[n\left(H(Y_1) - R(P, \Delta_1) + \delta'\right)\right]$$

for all $n \ge N_2(d_1, \eta_1, \delta')$.

Then, the remainder of the proof of the Lemma is straightforward. The second term on the right side of (23) is

$$\begin{aligned} \Pr\{\boldsymbol{x} \in \mathcal{U}(W_i^{n_i})\} &= \Pr\left\{ \boldsymbol{x} \in \mathcal{N}_1(Z_i) \bigcap \left(\bigcap_{j=1}^{n_i} \mathcal{N}_2(W_{ij})^c \right) \right\} \\ &= \sum_{z \in \mathcal{T}^n_{[Y_1]}} \Pr\{Z_i = z\} \\ &\quad \cdot \Pr\left\{ \boldsymbol{x} \in \mathcal{N}_1(Z_i) \bigcap \left(\bigcap_{j=1}^{n_i} \mathcal{N}_2(W_{ij})^c \right) \middle| Z_i = z \right\} \\ &= \frac{1}{|\mathcal{T}^n_{[Y_1]}|} \cdot \sum_{z \in \mathcal{A}(\boldsymbol{x})} \Pr\left\{ \boldsymbol{x} \in \bigcap_{j=1}^{n_i} \mathcal{N}_2(W_{ij})^c \middle| Z_i = z \right\}. \end{aligned}$$

Further

$$\begin{aligned} \Pr&\left\{ \boldsymbol{x} \in \bigcap_{j=1}^{n_i} \mathcal{N}_2(W_{ij})^c \middle| Z_i = z \right\} \\ &= \prod_{j=1}^{n_i} \Pr\{\boldsymbol{x} \in \mathcal{N}_2(W_{ij})^c \mid Z_i = z\} \\ &= \prod_{j=1}^{n_i} (1 - \Pr\{\boldsymbol{x} \in \mathcal{N}_2(W_{ij}) \mid Z_i = z\}) \\ &\le \prod_{j=1}^{n_i} \left(1 - \Pr\{W_{ij} \in \mathcal{T}^n_{[Y_2/X]}(\boldsymbol{x}) \mid Z_i = z\}\right) \end{aligned}$$

where the previous inequality follows as in [2, p. 151]. We continue the bounding according to

$$\Pr\left\{ \boldsymbol{x} \in \bigcap_{j=1}^{n_i} \mathcal{N}_2(W_{ij})^c \middle| Z_i = z \right\} \le \prod_{j=1}^{n_i} (1 - \Pr\{W_{ij} \in \mathcal{T}^n_{[Y_2/X,Y_1]}(\boldsymbol{x}, z) \mid Z_i = z\})$$

for all $n \ge N_3(d_2, \eta_2)$ (cf. [2, p. 151]). Furthermore, the right side above is bounded above by

$$\left\{1 - \exp\left[-n\left(I(X \wedge Y_2/Y_1) + \frac{\delta_2'}{2}\right)\right]\right\}^{n_i}$$

with $\delta_2' \triangleq \delta_2 - \delta_1$, for all $n \ge N_4(d_2, \eta_2, \delta_1, \delta_2)$, which in turn does not exceed

$$\exp\left[-\exp\frac{n\delta_2'}{6}\right]$$

provided that

$$\exp\left\{n\left[I(X \wedge Y_2 \mid Y_1) + \frac{2}{3}\delta_2'\right]\right\} \le n_i \le \exp\left\{n\left[I(X \wedge Y_2 \mid Y_1) + \frac{3}{4}\delta_2'\right]\right\}.$$

Hence, we can finally bound the second term on the right side of (23) as

$$\sum_{\boldsymbol{x} \in \mathcal{T}_P^n} \sum_{i=1}^m \Pr\{\boldsymbol{x} \in \mathcal{U}(W_i^{n_i})\} \le m \exp\left\{n \log|\mathcal{X}| - \exp\frac{n\delta_2'}{6}\right\} \frac{1}{|\mathcal{T}^n_{[Y_1]}|} |\mathcal{A}(\boldsymbol{x})|.$$

Using the facts

$$m \le \exp\left\{n\left[I(X \wedge Y_1) + \frac{3}{4}\delta_1\right]\right\}$$
$$|\mathcal{T}^n_{[Y_1]}| \ge \exp\left\{n\left[H(Y_1) - \frac{\delta_1}{4}\right]\right\}$$

and Claim 2, the right side above is easily seen to be bounded by

$$\exp\left\{n[2\log|\mathcal{X}| + 2\delta_1] - \exp\frac{n\delta_2'}{6}\right\} < \frac{1}{2} \tag{25}$$

for all $n \ge N_5(\delta_1, \delta_2')$ (because $R_1 - R(P, \Delta_1) \le \log|\mathcal{X}|$). Finally, we obtain from (24) and (25) that

$$\boldsymbol{E}\left|\mathcal{U}(Z^m) \bigcup \left(\bigcup_{i=1}^m \mathcal{U}(W_i^{n_i})\right)\right| < 1$$

for all

$$n \ge N(d_1, d_2, \eta_1, \eta_2, \delta_1, \delta_2) = \max\{N_1, \cdots, N_5\}.$$

It then follows for all n suitably large that there exist sets $\mathcal{B}_1 \subset \mathcal{Y}_1^n$, $\mathcal{B}_2(\boldsymbol{y}_i) \subset \mathcal{Y}_2^n$, $\boldsymbol{y}_i \in \mathcal{B}_1$, such that

$$|\mathcal{B}_1| \le m, \quad |\mathcal{B}_2(\boldsymbol{y}_i)| \le n_i, \quad i = 1, \cdots, m$$

so that

$$\begin{aligned}\sum_{i=1}^m |\mathcal{B}_2(\boldsymbol{y}_i)| &\le \sum_{i=1}^m n_i \\ &\le m\exp\left\{n\left[I(X \wedge Y_2 \mid Y_1) + \frac{3}{4}\delta_2'\right]\right\} \\ &\le \exp\left\{n\left[I(X \wedge Y_1Y_2) + \frac{3}{4}\delta_2\right]\right\}.\end{aligned}$$

Equivalently

$$\begin{aligned}\frac{1}{n}\log|\mathcal{B}_1| &\le I(X \wedge Y_1) + \frac{3}{4}\delta_1 \\ \frac{1}{n}\log|\mathcal{B}_2| &\le I(X \wedge Y_1Y_2) + \frac{3}{4}\delta_2.\end{aligned} \tag{26}$$

Assume that $I(X \wedge Y_1) \le R_1$. Then by the uniform continuity of $R(P, R_1, \Delta_1, \Delta_2)$ (which follows analogously as in [2, Lemma 2.2.2, pp. 124–125]), the desired inequalities (17), (18) follow for η_1, η_2 sufficiently small (cf. (22)).

It remains to establish the Claim. To this end, observe that $z \in \mathcal{A}(\boldsymbol{x})$ iff

$$z \in \mathcal{T}^n_{[Y_1]} \quad \text{and} \quad d_1(\boldsymbol{x}, z) = \boldsymbol{E}_{P_{\boldsymbol{x}}V}[d_1(X, \tilde{Y})] \le \Delta_1$$

where $(X, \tilde{Y})$ is a $\mathcal{X} \times \mathcal{Y}_1$-valued r.v. with joint pmf equal to the joint type $P_{\boldsymbol{x}}V$ of $(\boldsymbol{x}, z)$. Further,

$$\begin{aligned}H(V \mid P_{\boldsymbol{x}}) &= H(P_{\boldsymbol{x}}) - I(X \wedge \tilde{Y}) \\ &\le H(Y_1) + \frac{\delta'}{3} - \min I(X \wedge \tilde{Y}) \\ &= H(Y_1) + \frac{\delta'}{3} - R(P, \Delta_1)\end{aligned} \tag{27}$$

where the inequality follows for all n large (depending on δ') from [2, Lemma 1.2.7] and the minimum is taken over all r.v.'s $(X, \tilde{Y})$ such that $P_X = P_{\boldsymbol{x}}$ and $\boldsymbol{E}d_1(X, \tilde{Y}) \le \Delta_1$.

Hence, finally

$$\mathcal{A}(\boldsymbol{x}) = \left(\bigcup_{V:\ \boldsymbol{E}_{P_{\boldsymbol{x}}V}[d_1(X,\tilde{Y})] \le \Delta_1} \mathcal{T}^n_V(\boldsymbol{x})\right) \bigcap \mathcal{T}^n_{[Y_1]}$$

so that

$$|\mathcal{A}(\boldsymbol{x})| \le (1+n)^{|\mathcal{X}||\mathcal{Y}_1|}\exp\left\{n\left[H(Y_1 \mid X) + \frac{2\delta'}{3}\right]\right\}$$

for all $n \ge N_2(d_1, \eta_1, \delta')$, where the inequality above is a consequence of (27).

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their comments which resulted in a significant modification of our results.

REFERENCES

[1] R. Ahlswede, "The rate-distortion region for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-31, no. 6, pp. 721–726, Nov. 1985.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic Press, 1981.

[3] A. El Gamal and T. M. Cover, "Achievable rates for multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 6, pp. 851–857, Nov. 1982.

[4] W. H. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.

[5] ——, "Addendum to 'Successive refinement of information'," *IEEE Trans. Inform. Theory*, vol. 39, no. 4, pp. 1465–1466, July 1993.

[6] R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 4, pp. 480–489, July 1973.

[7] ——, "Conditional rate-distortion theory," Tech. Rep. 6502-2, Stanford Electronics Lab., Stanford Univ., Stanford, CA, Oct. 1972.

[8] A. H. Kaspi and T. Berger, "Rate-distortion for correlated sources with partially separated encoders," *IEEE Trans. Inform. Theory*, vol. IT-28, no. 6, pp. 828–840, Nov. 1982.

[9] V. Koshélev, "Estimation of mean error for a discrete successive-approximation scheme," *Probl. Pered. Informat.*, vol. 17, no. 3, pp. 20–33, July–Sept., 1981.

[10] ——, "Hierarchical coding of discrete sources," *Probl. Pered. Informat.*, vol. 16, no. 3, pp. 31–49, 1980.

[11] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 197–199, 1974.

[12] J. K. Omura, "Coding theorem for discrete-time sources," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 4, pp. 490–497, July 1973.

[13] ——, "A lower bounding method for channel and source coding probabilities," *Inform. Contr.*, vol. 27, pp. 148–177, 1975.

[14] L. Ozarow, "On a source-coding problem with two channels and three receivers," *Bell Syst. Tech. J.*, vol. 59, no. 10, pp. 1909–1921, Dec. 1980.

[15] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inform. Theory*, vol 40, no. 1, pp. 253–259, Jan. 94.

[16] T. R. Rockafellar, *Convex Analysis.* Princeton, NJ: Princeton Univ. Press, 1970.

[17] H. S. Witsenhausen and A. D. Wyner, "Source coding for multiple descriptions II: A binary source," *Bell Syst. Tech. J.*, vol. 60, no. 10, pp. 2281–2292, Dec. 1981.

[18] J. K. Wolf, A. D. Wyner, and J. Ziv, "Source coding for multiple description," *Bell Syst. Tech. J.*, vol. 60, no. 10, pp. 2281–2292, Dec. 1981.

[19] H. Yamamoto, "Source coding theory for cascade and branching communication systems," *IEEE Trans. Inform. Theory*, vol. IT-27, no. 3, pp. 299–308, May 1981.

[20] Z. Zhang and T. Berger, "New results in binary multiple descriptions," *IEEE Trans. Inform. Theory*, vol. IT-33, no. 4, pp. 502–521, July 1987.