Capacity/Storage Tradeoff in High-Dimensional Identification Systems

Ertem Tuncel, *Member, IEEE*

***Abstract*—The asymptotic tradeoff between the number of distinguishable objects and the necessary storage space (or equivalently, the search complexity) in an identification system is investigated. In the discussed scenario, high-dimensional (and noisy) feature vectors extracted from objects are first compressed and then enrolled in the database. When the user submits a random query object, the extracted noisy feature vector is compared against the compressed entries, one of which is output as the identified object. The first result this paper presents is a complete single-letter characterization of achievable storage and identification rates (measured in bits per feature dimension) subject to vanishing probability of identification error as the dimensionality of feature vectors becomes very large. This single-letter characterization is then extended for a multistage system whereby depending on the number of entries, the identification is performed by utilizing part or all of the recorded bits in the database. Finally, it is shown that a necessary and sufficient condition for a two-stage system to achieve single-stage capacities at each stage is Markovity of the optimal test channels.**

***Index Terms*—Capacity, databases, identification systems, successive refinement.**

I. Introduction

In [14], Willems *et al.* investigated the capacity of an identification system, i.e., the maximum achievable exponential rate of the number of distinguishable objects in a database, where the feature vectors extracted from objects have constant and known statistics. They showed that $\approx 2^{nR}$ objects can be distinguished from each other if and only if $R < C$ as n, the dimensionality of the feature space, becomes very large, and presented a single-letter characterization for the capacity C.

It was assumed in [14] that the feature vectors (corrupted by observation noise) are stored in the database as is, i.e., without any preprocessing. In that setting, however, a major bottleneck in high-dimensional retrieval systems, namely, *search complexity*, is not considered. Since the required storage space for 2^{nR} feature vectors grows without bound with n, it is impractical to store all the database in a random access memory, and using a hard storage medium becomes inevitable. This, in turn, impedes the identification process as hard storage devices are notoriously slow.

As a remedy to this bottleneck, one must employ either an *indexing* scheme, where clever pruning methods eliminate most of the data before retrieval, or a *compression*-based scheme, where only partial information is retrieved from each data entry. It is noted in the database literature that compression-based schemes outperform indexing schemes for high-dimensional applications (e.g., see [11].) Motivated by this fact, we discuss here the performance of identification systems where feature vectors are compressed before storage, and the entire database is retrieved in the compressed form, thereby expediting the identification process.

Of course, the price to be paid is the inevitable degradation in the identification performance, as some objects distinguishable in the original scenario of [14] will now be mapped to the same quantization index. This creates a tradeoff between the achieved compression rate R^{c} and the identification rate R^{i} both of which are measured in bits per feature dimension. In other words, $\approx 2^{nR^{\mathrm{i}}}$ objects are reliably identified if and only if $R^{\mathrm{i}} < C(R^{\mathrm{c}})$, where $C(R^{\mathrm{c}})$ is the storage-constrained capacity of the system. We derive a single-letter information-theoretic characterization of all achievable rate pairs $(R^{\mathrm{c}}, R^{\mathrm{i}})$ subject to vanishing probability of identification error as the dimensionality of feature vectors grows without bound.[1] As expected, this characterization reduces to that in [14] when the feature vectors are compressed losslessly.

We then extend our single-letter characterization to identification systems based on multistage compression. The motivation for a multistage system is that the number of entries, $2^{nR^{\mathrm{i}}}$, may not be known beforehand, or may even be dynamically changing as new entries are made and some obsolete entries are deleted. During identification, if R^{i} is lower than the system capacity, the amount of bits read from disk will be unnecessarily large in a single-stage system described above. Conversely, if R^{i} is above the capacity, identification cannot be successful with high probability. A multistage system, on the other hand, can partially solve this problem by adapting to the number (or exponential rate, to be precise) of entries. Specifically, depending on the actual number of entries, the system can utilize part or all of the recorded bits, thereby processing the query in a reliable and more time-efficient manner.

A naturally arising question is whether and under what conditions a performance loss is not incurred on a multistage identification system compared with its single-stage counterpart. More formally, with two stages of compression operating at cumulative rates R_1^{c} and R_2^{c}, the question is for which system statis-

Manuscript received August 02, 2006. Current version published April 22, 2009. The material in this paper was presented in part at IEEE International Symposium on Information Theory (ISIT), Seattle, WA, July 2006, and at the IEEE International Symposium on Information Theory (ISIT), Nice, France, June 2007.

The author is with the Department of Electrical Engineering, University of California, Riverside, CA 92521 USA (e-mail: ertem@ee.ucr.edu).

Communicated by K. Kobayashi, Associate Editor for Shannon Theory.
Digital Object Identifier 10.1109/TIT.2009.2016057

[1] After the submission of this manuscript, the author was informed that the characterization of a similar problem in the context of sensory pattern recognition was independently given in [7], [12], [13].

0018-9448/$25.00 © 2009 IEEE

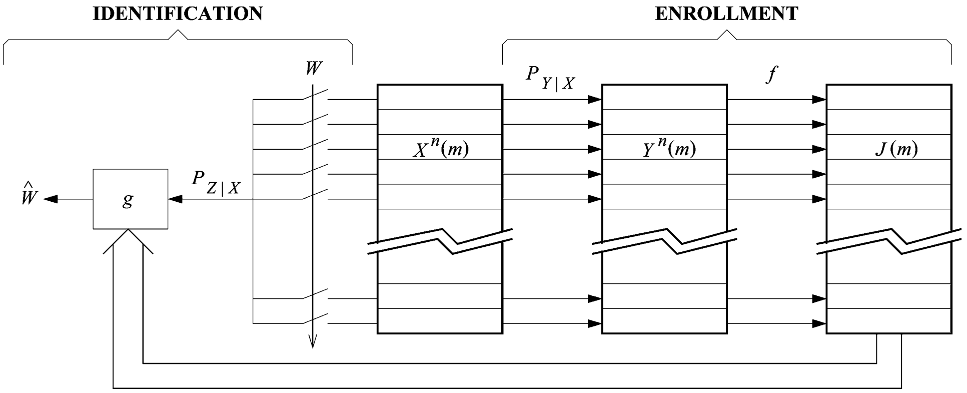

Fig. 1. The enrollment and the identification phases depicted together.

tics can we reach the identification rates $C(R_1^c)$ and $C(R_2^c)$ simultaneously? The usual Markovity condition in rate-distortion theory regarding distributions attaining optimum performance in various levels (c.f [3]–[5]) also becomes a necessary and sufficient condition in our scenario.

The scenario considered in this paper is not the same as, or a special case of, the author's previous work in [10], similarities notwithstanding. In particular, [10] discussed the fundamental performance of content-based retrieval *in a noise-free environment* in terms of search complexity, storage space, search quality, and reconstruction quality. In contrast, we assume perfect search quality with high probability where both the data and the query are noisy, and are not interested in reconstructing the data objects themselves. Our concern is the size of the database, whereas in [10], the database was considered to be of infinite size.

The rest of this paper is organized as follows. We begin by a formal definition of the problem in the next section. Section III presents a complete single-letter characterization of the compression/identification tradeoff. This characterization is then extended to multiple stages in Section IV. We present a summary and a few concluding remarks in Section V.

II. Notation and Formal Problem Definitions

We first discuss details of both the enrollment (i.e., storage) and the identification (i.e., query) phases in an information-theoretic framework, and then present the formal definition of the problem.

Assume that the feature vectors $\{X^n(m)\}_{m=1}^M$ are generated independently and according to

$$\Pr[X^n(m)=x^n]=\prod_{t=1}^n P_X(x_t)$$

where the feature alphabet $\mathcal{X}$ is finite. In the enrollment phase, noisy versions of $X^n(m)$, denoted by $Y^n(m)$, are observed and recorded. Let $Y^n(m)$ be modeled as the output of a discrete memoryless channel (DMC) governed by $P_{Y|X}$ with finite output alphabet $\mathcal{Y}$. That is

$$\Pr[Y^n(m)=y^n|X^n(m)=x^n]=\prod_{t=1}^n P_{Y|X}(y_t|x_t)$$

for $1 \leq m \leq M$. Unlike the original work of Willems *et al.* [14], we consider the scenario where $\{Y^n(m)\}_{m=1}^M$ are compressed before storage. For this purpose, a deterministic function

$$f:\mathcal{Y}^n \longrightarrow \mathcal{L}=\{1,2,\ldots,L\}$$

is applied on each $Y^n(m)$. We use the notation $J(m)=f(Y^n(m))$ for the compression indices of entries.

Let W be independent from $\{X^n(m),Y^n(m)\}_{m=1}^M$, uniformly distributed in $\mathcal{M}=\{1,2,\ldots,M\}$, and unknown to the user of the database. In the identification phase, the user observes Z^n, which is a noisy version of $X^n(W)$ corrupted by the DMC $P_{Z|X}$ with finite output alphabet $\mathcal{Z}$, i.e.,

$$\Pr[Z^n=z^n|X^n(W)=x^n]=\prod_{t=1}^n P_{Z|X}(z_t|x_t)$$

with $Z^n - X^n(W) - Y^n(W)$ forming a Markov chain, and desires to identify W using $\{J(m)\}_{m=1}^M$ and Z^n. A sequence of identification functions

$$g^{(M)}:\mathcal{L}^M \times \mathcal{Z}^n \longrightarrow \mathcal{M}$$

for $M=1,2,\ldots$ is to be designed for this purpose. Let

$$\hat{W}^{(M)}=g^{(M)}(J(1),\ldots,J(M),Z^n)$$

be the estimate of W. The identification process is considered successful with M entries if $\Pr[\hat{W}^{(M)} \neq W] \rightarrow 0$ as $n \rightarrow \infty$. The enrollment and the identification phases are illustrated together in Fig. 1.

Two conflicting qualities of this system are the compression rate $\frac{1}{n}\log L$ and the identification rate $\frac{1}{n}\log M$.[2] It is immediate that $\frac{1}{n}\log M \leq \frac{1}{n}\log L+\delta$ for arbitrarily small $\delta > 0$, because one cannot distinguish exponentially more entries than L while keeping the probability error arbitrarily small. Although it should also be clear that it is not possible to achieve $\frac{1}{n}\log M \approx \frac{1}{n}\log L$ due to both the enrollment noise and the identification noise, the precise relation between the two rates is not obvious. In this paper, we derive a single-letter characterization of all achievable rate pairs, where achievability is defined as follows.

[2] All logarithms in this paper are base 2.

Definition 1: $(R^{\mathrm{c}}, R^{\mathrm{i}})$ is an achievable *compression/identification* rate pair if for any $\delta > 0$ and large enough n, there exist a deterministic function f and a sequence of deterministic functions $g^{(1)}, g^{(2)}, \ldots$, such that

$$\frac{1}{n}\log L \le R^{\mathrm{c}} + \delta \tag{1}$$

and $\Pr[\hat{W}^{(M)} \ne W] \le \delta$ for all M satisfying

$$\frac{1}{n}\log M \le R^{\mathrm{i}} - \delta. \tag{2}$$

Remark 1: In the achievability definitions of capacity problems, one always encounters

$$\frac{1}{n}\log M \ge R^{\mathrm{i}} - \delta$$

instead of (2). The difference here is that our goal is not to design a code for the largest possible M only. In particular, we need a proper decoder for any M below capacity.

We denote by $\mathcal{R}$ the closure of all achievable $(R^{\mathrm{c}}, R^{\mathrm{i}})$. We also define the *capacity* of the system for a fixed compression rate as

$$C(R^{\mathrm{c}}) = \max\{R^{\mathrm{i}} : (R^{\mathrm{c}}, R^{\mathrm{i}}) \in \mathcal{R}\}.$$

Consider next an N-stage system, where multiresolution compression is performed on each feature vector, i.e., using N separate stage encoders

$$f_i : \mathcal{Y}^n \longrightarrow \mathcal{L}_i = \{1, 2, \ldots, L_i\}$$

for $1 \le i \le N$, and N sequences of identification functions

$$g_i^{(M)} : \mathcal{L}_1^M \times \cdots \times \mathcal{L}_i^M \times \mathcal{Z}^n \longrightarrow \mathcal{M}$$

for $1 \le i \le N$ and $M \ge 1$. We use the notation $J_i(m) = f_i(Y^n(m))$ for the ith-stage compression indices. We also denote by

$$\hat{W}_i^{(M)} = g_i^{(M)}(J_1(1), \ldots, J_1(M), \ldots, J_i(1), \ldots, J_i(M), Z^n)$$

the ith-stage estimates of W.

Definition 2: A $2N$-tuple $(R_1^{\mathrm{c}}, \ldots, R_N^{\mathrm{c}},\ R_1^{\mathrm{i}}, \ldots, R_N^{\mathrm{i}})$ with $R_1^{\mathrm{c}} \le R_2^{\mathrm{c}} \le \cdots \le R_N^{\mathrm{c}}$ and $R_1^{\mathrm{i}} \le R_2^{\mathrm{i}} \le \cdots \le R_N^{\mathrm{i}}$ is a *successively achievable* N-stage compression/identification rate vector if for any $\delta > 0$ and large enough n, there exist deterministic functions $f_i, g_i^{(1)}, g_i^{(2)}, \ldots$, for $1 \le i \le N$ such that

$$\sum_{j=1}^{i} \frac{1}{n}\log L_j \le R_i^{\mathrm{c}} + i\delta \tag{3}$$

and $\Pr[\hat{W}_i^{(M)} \ne W] \le \delta$ for all M satisfying

$$\frac{1}{n}\log M \le R_i^{\mathrm{i}} - \delta \tag{4}$$

for $1 \le i \le N$.

We denote by $\mathcal{R}_s$ the closure of all successively achievable $(R_1^{\mathrm{c}}, \ldots, R_N^{\mathrm{c}},\ R_1^{\mathrm{i}}, \ldots, R_N^{\mathrm{i}})$, and also derive a single-letter characterization of $\mathcal{R}_s$.

In proving the direct parts of our results, we employ strong typicality and use the notation of Csiszár and Körner [2]. More specifically, x^n is said to be strongly δ-typical with X if

$$\left|\frac{1}{n}N(a|x^n) - P_X(a)\right| \le \delta$$

and $P_X(a) = 0$ implies $N(a|x^n) = 0$ for all $a \in \mathcal{X}$, where $N(a|x^n)$ denotes the number of occurrences of a in x^n. The set of all x^n strongly δ-typical with X is denoted by $T_{[X]_\delta}^n$. This concept is easily generalized to collections of vectors and joint distributions. We refer the reader to [2] for a detailed discussion on strong typicality.

III. Characterization of Single-Stage Identification Performance

Definition 3: Let $\mathcal{R}^*$ be the set of all $(R^{\mathrm{c}}, R^{\mathrm{i}})$ such that there exists a joint distribution

$$P_{XYZU}(x,y,z,u) = P_X(x)P_{Z|X}(z|x)P_{Y|X}(y|x)P_{U|Y}(u|y)$$

i.e., a Markov chain $Z - X - Y - U$, satisfying

$$I(Y;U) \le R^{\mathrm{c}}$$
$$I(Z;U) \ge R^{\mathrm{i}}$$

where U is distributed over some discrete alphabet $\mathcal{U}$.

Lemma 1: $\mathcal{R}^*$ is convex. Moreover, in determining $\mathcal{R}^*$, it suffices to focus on $\mathcal{U}$ with $|\mathcal{U}| = |\mathcal{Y}| + 1$.

Proof: The proof is very similar to that of Theorem A2 in [16], which became folklore thereafter. We nevertheless include a sketch here for completeness. Note that P_X, $P_{Y|X}$, and $P_{Z|X}$ are fixed source statistics. Thus, $I(Y;U)$ and $I(Z;U)$ are solely determined by the choice of the *test channel* $P_{U|Y}$, or equivalently, by P_U and $P_{Y|U}$ subject to

$$\sum_{u\in\mathcal{U}} P_U(u)P_{Y|U}(y|u) = P_Y(y). \tag{5}$$

In light of this, define $\mathcal{I}$ as the set of all possible values the pair

$$\Big(I(Y;U), I(Z;U)\Big)$$

can assume as $P_{U|Y}$ changes. To prove convexity of $\mathcal{R}$, it suffices to prove that of $\mathcal{I}$. Towards that end, define for a generic distribution Q over $\mathcal{Y}$

$$\Gamma_y(Q) \triangleq Q(y)$$

for $1 \le y \le |\mathcal{Y}| - 1$, where it is assumed without loss of generality that $\mathcal{Y} = \{1, 2, \ldots, |\mathcal{Y}|\}$. Also define

$$\Gamma_{|\mathcal{Y}|}(Q) \triangleq H(Y) + \sum_y Q(y)\log Q(y)$$

$$\Gamma_{|\mathcal{Y}|+1}(Q) \triangleq H(Z) + \sum_{y,z} P_{Z|Y}(z|y)Q(y)\cdot$$
$$\log\left(\sum_{y'} P_{Z|Y}(z|y')Q(y')\right).$$

Now, the key observation is that for any P_U and $P_{Y|U}$ satisfying (5), it is true that

$$I(Y;U) = \sum_{u\in\mathcal{U}} P_U(u)\Gamma_{|\mathcal{Y}|}(P_{Y|U}(\cdot|u)) \tag{6}$$

$$I(Z;U) = \sum_{u\in\mathcal{U}} P_U(u)\Gamma_{|\mathcal{Y}|+1}(P_{Y|U}(\cdot|u)). \tag{7}$$

In other words, $\mathcal{I}$ coincides with the cross section of all convex combinations

$$\{\gamma_i\}_{i=1}^{|\mathcal{Y}|+1} = \left\{\sum_{u\in\mathcal{U}} \alpha_u \Gamma_i(P_{Y|U}(\cdot|u))\right\}_{i=1}^{|\mathcal{Y}|+1}$$

corresponding to $\gamma_i = P_Y(i)$ for $1 \le i \le |\mathcal{Y}| - 1$. Thus, $\mathcal{I}$ is convex by construction. Moreover, from the well-known Caratheodory's theorem [1], any point in the convex hull of a connected set in a k-dimensional Euclidean space can be expressed as the convex combination of at most k vectors in that set. Thus, $|\mathcal{U}| = |\mathcal{Y}| + 1$ is sufficient in characterizing $\mathcal{I}$ (and therefore $\mathcal{R}^*$). □

We are now ready to prove the main result of this paper.

Theorem 1: $\mathcal{R} = \mathcal{R}^*$.

Proof: We begin by proving $\mathcal{R} \subset \mathcal{R}^*$. Assume that $(R^{\mathrm{c}}, R^{\mathrm{i}}) \in \mathcal{R}$. Then, for any $\delta > 0$ and large enough n, there exist deterministic functions $f, g^{(1)}, g^{(2)}, \ldots$ such that (1) is satisfied, and $\Pr[\hat{W}^{(M)} \ne W] \le \delta$ for all M satisfying (2). For any such system and M

$$\begin{aligned}\log M &= H(W)\\ &= H(W|J(1),\ldots,J(M),Z^n)\\ &\quad + I(W;J(1),\ldots,J(M),Z^n)\\ &\overset{(a)}{\le} H(W|\hat{W}^{(M)})\\ &\quad + I(W;J(1),\ldots,J(M),Z^n)\\ &\overset{(b)}{\le} 1 + \Pr[\hat{W}^{(M)} \ne W]\log M\\ &\quad + I(W;J(1),\ldots,J(M),Z^n)\\ &\le 1 + \delta\log M + I(W;J(1),\ldots,J(M),Z^n)\end{aligned}$$

and thus

$$\begin{aligned}&(1-\delta)\log M - 1\\ &\quad\le I(W;J(1),\ldots,J(M))\\ &\qquad + I(W;Z^n|J(1),\ldots,J(M))\\ &\quad\overset{(c)}{=} I(W;Z^n|J(1),\ldots,J(M))\\ &\quad= H(Z^n|J(1),\ldots,J(M))\\ &\qquad - H(Z^n|J(1),\ldots,J(M),W)\\ &\quad\le H(Z^n) - H(Z^n|J(1),\ldots,J(M),W)\\ &\quad\overset{(d)}{=} H(Z^n) - H(Z^n|J(W))\end{aligned}$$

where (a) follows because $\hat{W}$ is a function of $J(1),\ldots,J(M)$, and Z^n, (b) follows from Fano's inequality, (c) is due to the fact that W is independent of $\{J(m)\}_{m=1}^M$, and (d) holds because Z^n is independent of all $J(m)$ with $m \ne W$. Defining

$$U_t = (Z_1^{t-1}, J(W))$$

and considering the largest M satisfying (2), we then have

$$\begin{aligned}R^{\mathrm{i}} - \delta' &\le \frac{1}{n}\Big[H(Z^n) - H(Z^n|J(W))\Big]\\ &= \frac{1}{n}\sum_{t=1}^n \Big[H(Z_t) - H(Z_t|Z_1^{t-1}, J(W))\Big]\\ &= \frac{1}{n}\sum_{t=1}^n \Big[H(Z_t) - H(Z_t|U_t)\Big]\\ &= \frac{1}{n}\sum_{t=1}^n I(Z_t;U_t)\end{aligned} \tag{8}$$

where $\delta' \to 0$ as $\delta \to 0$ and $n \to \infty$.

Regarding the compression rate the system achieves, we have the following chain of inequalities:

$$\begin{aligned}n(R^{\mathrm{c}} + \delta) &\ge \log L \ge H(J(W))\\ &= H(J(W)) - H(J(W)|Y^n(W))\\ &= I(J(W);Y^n(W))\\ &= H(Y^n(W)) - H(Y^n(W)|J(W))\\ &= \sum_{t=1}^n \Big[H(Y_t(W)) - H(Y_t(W)|J(W),Y_1^{t-1}(W))\Big]\\ &\overset{(e)}{=} \sum_{t=1}^n \Big[H(Y_t(W))\\ &\qquad - H(Y_t(W)|J(W),Y_1^{t-1}(W),Z_1^{t-1})\Big]\\ &\overset{(f)}{\ge} \sum_{t=1}^n \Big[H(Y_t(W)) - H(Y_t(W)|J(W),Z_1^{t-1})\Big]\\ &= \sum_{t=1}^n \Big[H(Y_t(W)) - H(Y_t(W)|U_t)\Big]\\ &= \sum_{t=1}^n I(Y_t(W);U_t)\end{aligned} \tag{9}$$

where (e) and (f), respectively, follow from the fact that $Y_t(W) - (J(W), Y_1^{t-1}(W)) - Z_1^{t-1}$ forms a Markov chain and because conditioning reduces entropy. This Markov chain is implied by

$$Z_1^{t-1} - Y_1^{t'}(W) - Y^n(W) - J(W)$$

for any $t' \ge t-1$, because

$$\begin{aligned}&I(Y_t(W);Z_1^{t-1}|J(W),Y_1^{t-1}(W))\\ &\quad= H(Z_1^{t-1}|J(W),Y_1^{t-1}(W)) - H(Z_1^{t-1}|J(W),Y_1^{t}(W))\\ &\quad= H(Z_1^{t-1}|Y_1^{t-1}(W)) - H(Z_1^{t-1}|Y_1^{t}(W))\\ &\quad= H(Z_1^{t-1}|Y_1^{t-1}(W)) - H(Z_1^{t-1}|Y_1^{t-1}(W))\\ &\quad= 0.\end{aligned}$$

We next show that $Z_t - X_t(W) - Y_t(W) - U_t$ also forms a Markov chain. This will finish the converse part of the theorem, because then

$$\Big(I(Y_t(W);U_t), I(Z_t;U_t)\Big) \in \mathcal{R}^*$$

for $1 \le t \le n$. Since $\mathcal{R}^*$ is convex, this, in turn, implies $(R^{\mathrm{c}} + \delta, R^{\mathrm{i}} - \delta') \in \mathcal{R}^*$ for any $\delta > 0$ and large enough n. Towards that end, it suffices to show $Z_t - (X_t(W), Y_t(W)) - U_t$, because $Z_t - X_t(W) - Y_t(W)$ and $X_t(W) - Y_t(W) - U_t$ are obvious, and $A - B - C$, $B - C - D$, and $A - (B, C) - D$ implies $A - B - C - D$. Now, observe

$$Z_t - (Z_1^{t-1}, X_t(W), Y_t(W)) - Y^n(W) - J(W)$$

and thus

$$\begin{aligned} I(Z_t; U_t | X_t(W), Y_t(W)) &= H(Z_1^{t-1}, J(W) | X_t(W), Y_t(W)) \\ &\quad - H(Z_1^{t-1}, J(W) | X_t(W), Y_t(W), Z_t) \\ &= H(Z_1^{t-1} | X_t(W), Y_t(W)) \\ &\quad + H(J(W) | X_t(W), Y_t(W), Z_1^{t-1}) \\ &\quad - H(Z_1^{t-1} | X_t(W), Y_t(W), Z_t) \\ &\quad - H(J(W) | X_t(W), Y_t(W), Z_1^t) \\ &= H(J(W) | X_t(W), Y_t(W), Z_1^{t-1}) \\ &\quad - H(J(W) | X_t(W), Y_t(W), Z_1^t) \\ &= 0. \end{aligned}$$

To prove the direct part, we use standard random coding techniques. Assume $(R^c, R^i) \in \mathcal{R}^*$, and generate codevectors $\{U^n(j)\}_{j=1}^L$ independent and identically distributed (i.i.d.) $\sim P_U$, where L is to be determined later, and P_U is the marginal distribution of U satisfying $Z - X - Y - U$, $I(Y; U) \leq R^c$, and $I(Z; U) \geq R^i$. Also fix $\delta > 0$.

Enrollment: For arbitrary $y^n \in \mathcal{Y}^n$, define $f(y^n)$ as the smallest j such that $(y^n, U^n(j)) \in T^n_{[YU]_\delta}$, and if no such j is found, let $f(y^n) = 1$. Set $J(m) = f(Y^n(m))$.

Identification: For arbitrary $z^n \in \mathcal{Z}^n$ and $j(1), \ldots, j(M)$, let $g^{(M)}(j(1), \ldots, j(M), z^n)$ be defined as the smallest m such that $(z^n, U^n(j(m))) \in T^n_{[ZU]_\delta}$. If no such m is found, let $g^{(M)}(j(1), \ldots, j(M), z^n) = 1$. Set

$$\hat{W}^{(M)} = g^{(M)}(J(1), J(2), \ldots, J(M), Z^n).$$

Probability of Error: Define the following events:

$$\begin{aligned} \mathcal{E}_1(m) &\equiv \left\{ (Y^n(m), U^n(J(m))) \notin T^n_{[YU]_\delta} \right\} \\ \mathcal{E}_2(m) &\equiv \left\{ (Z^n, U^n(J(m))) \notin T^n_{[ZU]_\delta} \right\}. \end{aligned}$$

The probability of error computed over the ensemble of codebooks can then be bounded as

$$\begin{aligned} \Pr[\hat{W}^{(M)} \neq W | W = w] &\leq \Pr[\mathcal{E}_1(w)] + \Pr[\mathcal{E}_2(w) | \mathcal{E}_1(w)^c] \\ &\quad + \sum_{m \neq w} \Pr[\mathcal{E}_2(m)^c]. \end{aligned}$$

It can be shown using standard techniques that $\Pr[\mathcal{E}_1(w)] \leq \delta/3$ for large enough n if

$$I(Y; U) + \frac{\delta}{2} \leq \frac{1}{n} \log L \leq R^c + \delta.$$

It also follows from the well-known Markov lemma [1] and $Z - Y - U$ that $\Pr[\mathcal{E}_2(w) | \mathcal{E}_1(w)^c] \leq \delta/3$. Finally

$$\sum_{m \neq w} \Pr[\mathcal{E}_2(m)^c] \leq M 2^{-n[I(Z;U) - \delta/2]} \leq \frac{\delta}{3}$$

for large enough n whenever

$$\frac{1}{n} \log M \leq I(Z; U) - \delta$$

and therefore, whenever

$$\frac{1}{n} \log M \leq R^i - \delta.$$

In summary, we created a random codebook satisfying $\frac{1}{n} \log L \leq R^c + \delta$ and $\Pr[\hat{W}^{(M)} \neq W] \leq \delta$ averaged over the whole ensemble whenever $\frac{1}{n} \log M \leq R^i - \delta$. There must then exist a deterministic codebook $\{u^n(j)\}_{j=1}^L$ satisfying the same properties. Thus, $(R^c, R^i) \in \mathcal{R}$, and therefore $\mathcal{R}^* \subset \mathcal{R}$. □

Corollary 1: The identification capacity subject to a storage constraint is given by

$$C(R^c) = \max_{\substack{Z-X-Y-U \\ I(Y;U) \leq R^c}} I(Z; U). \tag{10}$$

We denote by $P^*_{U|Y}(R^c)$ the test channel that achieves the point $(R^c, C(R^c))$.

A. Computation of $C(R^c)$

As was observed in [8], (10) coincides exactly with the *information bottleneck* function introduced in [9], thereby providing an operational meaning to it. When $R^c = H(Y)$, the constraint $I(Y; U) \leq R^c$ in (10) becomes vacuous, and we have

$$C(R^c) = I(Z; Y)$$

where the maximum is achieved by $U = Y$. This is the same result derived by Willems *et al.* [14]. At the other extreme, when $R^c = 0$, $I(Y; U) \leq R^c$ implies independence of Y and U and thus of Z and U. Therefore, $C(0) = 0$, as expected.

In principle, the intermediate values of $C(R^c)$ can be calculated using the Lagrangian maximization

$$L(\beta) = \max_{Z-X-Y-U} \left[I(Z; U) - \beta I(Y; U) \right]$$

and the dual minimization

$$C(R^c) = \min_{0 \leq \beta \leq 1} \left[L(\beta) + \beta R^c \right]$$

since by Lemma 1, $C(R^c)$ is concave in R^c. However, $I(Z; U) - \beta I(Y; U)$ is in general neither convex nor concave in $P_{U|Y}$.[3] In that sense, the Lagrangian approach proves difficult even with small alphabets. Although the treatment of the problem in [9] provides a solution in terms of *necessary conditions*, the nonconvexity of the problem was not addressed, and in general more than one $P_{U|Y}$ can satisfy the conditions.

As an alternative, we next adopt the convex combination approach of the proof of Lemma 1 for two examples with $|\mathcal{Y}| = 2$. The first example was also treated in [15] in a different context and using a different technique.

1) Enrollment and Identification Subject to Binary Symmetric Channel Noise: Let $P_X(x) = 1/2$ for $x \in \mathcal{X} = \{0, 1\}$, and let $P_{Y|X}$ and $P_{Z|X}$ be binary symmetric channels (BSC) with crossover probabilities p_1 and p_2, respectively. This implies that

[3] Observe that *both* $I(Y; U)$ and $I(Z; U)$ are convex in $P_{U|Y}$.

$P_Y(y) = 1/2$ for $y \in \{0,1\}$ and $P_{Z|Y}$ is a BSC with crossover probability $\gamma = p_1 \star p_2$, where

$$a \star b = a(1-b) + b(1-a).$$

We determine $\mathcal{R} = \mathcal{R}^*$ through computing the set $\mathcal{I}$. To that end, let $q = P_{Y|U}(0|u)$ for some fixed u, and note that $\mathcal{I}$ is the cross section of the convex hull of all

$$\Big(q, 1-\mathcal{H}(q), 1-\mathcal{H}(q \star \gamma)\Big)$$

corresponding to uniform P_Y, where $0 \le q \le 1$, and $\mathcal{H}$ denotes the binary entropy function. Since both $1-\mathcal{H}(q)$ and $1-\mathcal{H}(q\star\gamma)$ are symmetric around $q = 1/2$, the set $\mathcal{I}$ coincides with the convex hull of

$$\Big(1-\mathcal{H}(q), 1-\mathcal{H}(q \star \gamma)\Big) \tag{11}$$

for $0 \le q \le 1/2$. That is, because given $0 \le q_1, q_2 \le 1/2$ and $0 < \alpha < 1$, the point

$$\alpha\Big(1-\mathcal{H}(q_1), 1-\mathcal{H}(q_1 \star \gamma)\Big) \\ +(1-\alpha)\Big(1-\mathcal{H}(q_2), 1-\mathcal{H}(q_2 \star \gamma)\Big)$$

can be obtained by setting

$$P_U(\cdot) = \left\{\frac{\alpha}{2}, \frac{\alpha}{2}, \frac{1-\alpha}{2}, \frac{1-\alpha}{2}\right\}$$

and

$$P_{Y|U}(0|\cdot) = \{q_1, 1-q_1, q_2, 1-q_2\}.$$

On the other hand, as we prove in the Appendix, (11) yields a *concave* curve on the $(R^{\mathrm{c}}, R^{\mathrm{i}})$-plane for all $0 \le \gamma \le 1/2$. This, in turn, implies that $C(R^{\mathrm{c}})$ itself is parametrically characterized by (11). Alternatively

$$C(R^{\mathrm{c}}) = 1 - \mathcal{H}(\gamma \star \mathcal{H}^{-1}(1-R^{\mathrm{c}}))$$

for all $0 \le R^{\mathrm{c}} \le 1$, where $\mathcal{H}^{-1}$ returns values between 0 and 1/2. Also, the optimum $P_{Y|U}$ and P_U corresponding to points on $C(R^{\mathrm{c}})$ are given by

$$P_{Y|U}(\cdot|0) = \{q, 1-q\}$$
$$P_{Y|U}(\cdot|1) = \{1-q, q\}$$

and $P_U(\cdot) = \{1/2, 1/2\}$. In other words, $P^*_{U|Y}(R^{\mathrm{c}})$ is a binary symmetric channel with crossover probability q. Fig. 2 depicts the behavior of $C(R^{\mathrm{c}})$ when $p_1 = p_2 = 0.1$.

2) Noiseless Enrollment and Identification Subject to Erasure Noise: This example is motivated by the scenario where enrollment is offline (and thus very reliable), whereas identification must be performed in real time, and therefore is prone to erasure (some feature points may not be obtained).

For $P_X(x) = 1/2$ for $x \in \mathcal{X} = \{0,1\}$, let $Y = X$, and let $P_{Z|X}$ be the erasure channel with erasure probability ϵ. We use the same technique as in the previous example. Observe that for $q = P_{Y|U}(0|u)$

$$\Gamma_2(q) = 1 - \mathcal{H}(q)$$

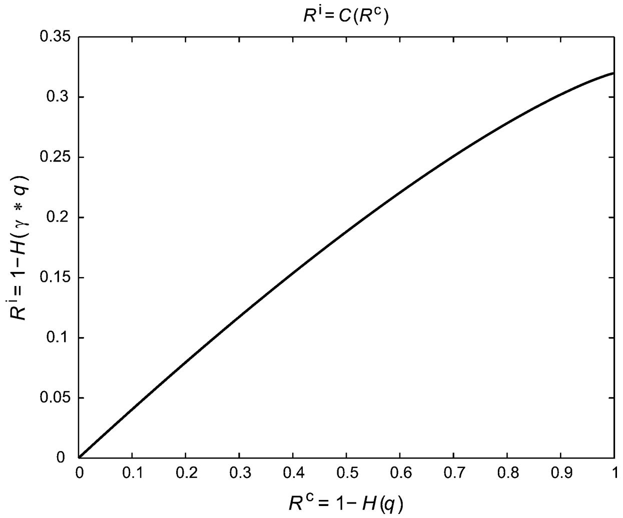

Fig. 2. $C(R^{\mathrm{c}})$ for an identification system with BSC noise when $p_1 = p_2 = 0.1$.

and

$$\begin{aligned}\Gamma_3(q) =& \mathcal{H}\left(\frac{1-\epsilon}{2}, \frac{1-\epsilon}{2}, \epsilon\right) \\ &- \mathcal{H}\left(\frac{q(1-\epsilon)}{2}, \frac{(1-q)(1-\epsilon)}{2}, \epsilon\right) \\ =& \Big[\mathcal{H}(\epsilon) + 1 - \epsilon\Big] - \Big[\mathcal{H}(\epsilon) + (1-\epsilon)\mathcal{H}(q)\Big] \\ =& (1-\epsilon)(1-\mathcal{H}(q))\end{aligned}$$

where $\mathcal{H}$ denotes the binary entropy as well as, with an abuse of notation, the entropy of an arbitrary distribution. Thus, $\mathcal{I}$ is parameterized by

$$\Big(1-\mathcal{H}(q), (1-\epsilon)(1-\mathcal{H}(q))\Big)$$

for $0 \le q \le 1/2$. In other words

$$C(R^{\mathrm{c}}) = (1-\epsilon)R^{\mathrm{c}}$$

for all $0 \le R^{\mathrm{c}} \le 1$. Also, as in the previous example, $P^*_{U|Y}(R^{\mathrm{c}})$ is a binary symmetric channel with crossover probability q.

IV. The Multistage Extension

A. Achievability Region for Two Stages

Definition 4: Let $\mathcal{R}^*_s$ be the set of all $(R^{\mathrm{c}}_1, R^{\mathrm{c}}_2, R^{\mathrm{i}}_1, R^{\mathrm{i}}_2)$ such that there exists a joint distribution

$$P_{XYZUV}(x,y,z,u,v) \\ = P_X(x)P_{Z|X}(z|x)P_{Y|X}(y|x)P_{UV|Y}(u,v|y)$$

i.e., a Markov chain $Z - X - Y - (U,V)$, satisfying

$$\begin{aligned}I(Y;U) &\le R^{\mathrm{c}}_1 \\ I(Y;U,V) &\le R^{\mathrm{c}}_2 \\ I(Z;U) &\ge R^{\mathrm{i}}_1 \\ I(Z;U,V) &\ge R^{\mathrm{i}}_2\end{aligned}$$

where U and V are auxiliary random variables distributed over discrete alphabets $\mathcal{U}$ and $\mathcal{V}$, respectively.

Lemma 2: $\mathcal{R}_s^*$ is convex. Moreover, in determining $\mathcal{R}_s^*$, it suffices to focus on $\mathcal{U}$ and $\mathcal{V}$ with

$$\begin{aligned}|\mathcal{U}| &= |\mathcal{Y}|+3\\ |\mathcal{V}| &= |\mathcal{U}|\cdot|\mathcal{Y}|+1.\end{aligned}$$

Proof: In proving convexity of $\mathcal{R}_s^*$ and that $|\mathcal{U}|=|\mathcal{Y}|+3$ is sufficient, we use the same approach as in the proof of Lemma 1. Define $\mathcal{J}$ as the set of all possible values the quadruple

$$\Big(I(Y;U), I(Y;U,V), I(Z;U), I(Z;U,V)\Big)$$

can assume as $P_{UV|Y}$ changes. To prove convexity of $\mathcal{R}_s^*$, it suffices to prove that $\mathcal{J}$ is convex. Towards that end, we first define $|\mathcal{Y}|+3$ continuous functionals of a generic distribution Q over $\mathcal{V}\times\mathcal{Y}$. Let

$$\Phi_y(Q) \triangleq \sum_v Q(v,y)$$

for $1 \le y \le |\mathcal{Y}|-1$, and

$$\begin{aligned}\Phi_{|\mathcal{Y}|}(Q) &\triangleq H(Y) + \sum_{v,y} Q(v,y)\log\left(\sum_{v'} Q(v',y)\right)\\ \Phi_{|\mathcal{Y}|+1}(Q) &\triangleq H(Y) + \sum_{v,y} Q(v,y)\log\left(\frac{Q(v,y)}{\sum_{y'} Q(v,y')}\right)\\ \Phi_{|\mathcal{Y}|+2}(Q) &\triangleq H(Z) + \sum_{v,y,z} P_{Z|Y}(z|y)Q(v,y)\cdot\\ &\qquad \log\left(\sum_{v',y'} P_{Z|Y}(z|y')Q(v',y')\right)\\ \Phi_{|\mathcal{Y}|+3}(Q) &\triangleq H(Z) + \sum_{v,y,z} P_{Z|Y}(z|y)Q(v,y)\cdot\\ &\qquad \log\left(\frac{\sum_{y'} P_{Z|Y}(z|y')Q(v,y')}{\sum_{y'} Q(v,y')}\right).\end{aligned}$$

Now, consider the cross section of all convex combinations

$$\{\phi_i\}_{i=1}^{|\mathcal{Y}|+3} = \left\{\sum_{u\in\mathcal{U}} \alpha_u \Phi_i(P_{VY|U}(\cdot,\cdot|u))\right\}_{i=1}^{|\mathcal{Y}|+3}$$

corresponding to $\phi_i = P_Y(i)$ for $1 \le i \le |\mathcal{Y}|-1$. This set, by construction, is convex and coincides with $\mathcal{J}$ where $P_U(u)$ plays the role of α_u. Moreover, from the Caratheodory's theorem, for any fixed discrete alphabet $\mathcal{V}$, $|\mathcal{U}|=|\mathcal{Y}|+3$ is sufficient in characterizing $\mathcal{J}$ (and therefore $\mathcal{R}_s^*$).

We now follow the methodology in [6, Sec. VI-A] and limit the size of $\mathcal{V}$. Another interpretation of what we showed so far is that given arbitrarily large discrete alphabets $\mathcal{U}$ and $\mathcal{V}$ and a distribution $P_U P_{VY|U}$ attaining the point $I \in \mathcal{J}$, there exists a distribution $P'_U P'_{VY|U}$ also attaining I, where $P'_Y = P_Y$ and the new U is confined to a reduced alphabet $\mathcal{U}'$ with $|\mathcal{U}'| = |\mathcal{Y}|+3$. To limit $|\mathcal{V}|$, we iterate this idea one step further. More specifically, rewriting $P'_U P'_{VY|U}$ as $P'_V P'_{UY|V}$, one can construct another distribution $P''_V P''_{UY|V}$ also attaining I, where $P''_{UY} = P'_{UY}$ and the new V is confined to an alphabet $\mathcal{V}''$ with $|\mathcal{V}''| = |\mathcal{U}'|\cdot|\mathcal{Y}|+1$. To see this, define for any $Q(u,y)$

$$\Psi_i(Q) \triangleq Q(i)$$

for $1 \le i \le |\mathcal{U}'|\cdot|\mathcal{Y}|-1$, where with abuse of notation, $Q(i)$ indicates the probability of the ith element of $\mathcal{U}'\times\mathcal{Y}$ with respect to some arbitrary order, and

$$\begin{aligned}\Psi_{|\mathcal{U}'|\cdot|\mathcal{Y}|}(Q) &\triangleq H(Y) + \sum_{u,y} Q(u,y)\log\left(\frac{Q(u,y)}{\sum_{y'} Q(u,y')}\right)\\ \Psi_{|\mathcal{U}'|\cdot|\mathcal{Y}|+1}(Q) &\triangleq H(Z) + \sum_{u,y,z} P_{Z|Y}(z|y)Q(u,y)\cdot\\ &\qquad \log\left(\frac{\sum_{y'} P_{Z|Y}(z|y')Q(u,y')}{\sum_{y',z'} P_{Z|Y}(z'|y')Q(u,y')}\right).\end{aligned}$$

Similarly to the previous construction, the set of all possible $\Big(I(Y;U,V), I(Z;U,V)\Big)$ coincides with the cross section of convex combinations of $\{\Psi_i\}_{i=1}^{|\mathcal{U}'|\cdot|\mathcal{Y}|+1}$ corresponding to a fixed P_{UY}. This completes the proof. □

We next extend Theorem 1 to two-stage systems.

Theorem 2: $\mathcal{R}_s = \mathcal{R}_s^*$.

Proof: We begin by proving $\mathcal{R}_s \subset \mathcal{R}_s^*$. Assume that $(R_1^{\mathrm{c}}, R_2^{\mathrm{c}}, R_1^{\mathrm{i}}, R_2^{\mathrm{i}}) \in \mathcal{R}_s$. Then, for any $\delta > 0$ and large enough n, there exist deterministic functions $f_1, g_1^{(1)}, g_1^{(2)}, \ldots$ and $f_2, g_2^{(1)}, g_2^{(2)}, \ldots$ such that (3) holds for $i = 1, 2$, and $\Pr[\hat{W}_i^{(M)} \neq W] \le \delta$ for all M satisfying (4) for $i = 1, 2$.

It can be shown by following the exact same steps as in (8) that

$$R_1^{\mathrm{i}} - \delta' \le \frac{1}{n}\sum_{t=1}^{n} I(Z_t; U_t) \tag{12}$$

where

$$U_t = (Z_1^{t-1}, J_1(W))$$

and $\delta' \longrightarrow 0$ as $\delta \longrightarrow 0$ and $n \longrightarrow \infty$. Similarly, following the steps in (9), one can show

$$R_1^{\mathrm{c}} + \delta \ge \frac{1}{n}\sum_{t=1}^{n} I(Y_t(W); U_t). \tag{13}$$

Further, replacing $J(1),\ldots,J(W)$ and $\hat{W}^{(M)}$ in the derivation of (8) with $J_1(1),\ldots,J_1(W), J_2(1),\ldots,J_2(W)$ and $\hat{W}_2^{(M)}$, respectively, it immediately follows that

$$\begin{aligned}R_2^{\mathrm{i}} - \delta' &\le \frac{1}{n}\sum_{t=1}^{n}\Big[H(Z_t) - H(Z_t|Z_1^{t-1}, J_1(W), J_2(W))\Big]\\ &= \frac{1}{n}\sum_{t=1}^{n}\Big[H(Z_t) - H(Z_t|U_t, V_t)\Big]\\ &= \frac{1}{n}\sum_{t=1}^{n} I(Z_t; U_t, V_t)\end{aligned} \tag{14}$$

where $V_t = J_2(W)$ for $1 \le t \le n$. Finally, we also have

$$\begin{aligned}n(R_2^{\mathrm{c}} + 2\delta) &\ge \log L_1 + \log L_2\\ &\ge H(J_1(W)) + H(J_2(W))\\ &\ge H(J_1(W), J_2(W))\\ &= H(J_1(W), J_2(W))\\ &\qquad - H(J_1(W), J_2(W)|Y^n(W))\\ &= I(J_1(W), J_2(W); Y^n(W))\end{aligned}$$

$$
\begin{aligned}
&= H(Y^n(W)) - H(Y^n(W)|J_1(W), J_2(W))\\
&= \sum_{t=1}^{n} \Big[H(Y_t(W))\\
&\qquad - H(Y_t(W)|J_1(W), J_2(W), Y_1^{t-1}(W))\Big]\\
&\overset{\text{(g)}}{=} \sum_{t=1}^{n} \Big[H(Y_t(W))\\
&\qquad - H(Y_t(W)|J_1(W), J_2(W), Y_1^{t-1}(W), Z_1^{t-1})\Big]\\
&\geq \sum_{t=1}^{n} \Big[H(Y_t(W))\\
&\qquad - H(Y_t(W)|J_1(W), J_2(W), Z_1^{t-1})\Big]\\
&= \sum_{t=1}^{n} \Big[H(Y_t(W)) - H(Y_t(W)|U_t, V_t)\Big]\\
&= \sum_{t=1}^{n} I(Y_t(W); U_t, V_t) \qquad (15)
\end{aligned}
$$

where (g) follows from the fact that $Y_t(W) - (J_1(W), J_2(W), Y_1^{t-1}(W)) - Z_1^{t-1}$ forms a Markov chain. Observe that $Z_t - X_t(W) - Y_t(W) - (U_t, V_t)$ also forms a Markov chain. The converse part is therefore complete since the convexity of $\mathcal{R}_s^*$ implies

$$(R_{c_1} + \delta, R_{c_2} + 2\delta, R_1^{\mathrm{i}} - \delta', R_2^{\mathrm{i}} - \delta') \in \mathcal{R}_s^*$$

for any $\delta > 0$ and large enough n.

The proof of the direct part, i.e., that of $\mathcal{R}_s^* \subset \mathcal{R}_s$, is also very similar to its single-stage counterpart. Assume $(R_1^{\mathrm{c}}, R_2^{\mathrm{c}}, R_1^{\mathrm{i}}, R_2^{\mathrm{i}}) \in \mathcal{R}_s^*$. Then there exist auxiliary random variables U and V satisfying $Z - X - Y - (U, V)$, $I(Y;U) \leq R_1^{\mathrm{c}}$, $I(Y;U,V) \leq R_2^{\mathrm{c}}$, $I(Z;U) \geq R_1^{\mathrm{i}}$, and $I(Z;U,V) \geq R_2^{\mathrm{i}}$. Generate codevectors $\{U^n(j)\}_{j=1}^{L_1}$ i.i.d. $\sim P_U$, and for each j, generate $\{V^n(k|j)\}_{k=1}^{L_2}$ according to

$$\Pr[V^n(k|j) = v^n | U^n(j) = u^n] = \prod_{t=1}^{n} P_{V|U}(v_t|u_t)$$

where L_1 and L_2 are to be determined later. Also fix $\delta > 0$.

Enrollment: For arbitrary $y^n \in \mathcal{Y}^n$, define $f_1(y^n)$ as the smallest j such that $(y^n, U^n(j)) \in T^n_{[YU]_\delta}$, and if no such j is found, let $f_1(y^n) = 1$. Similarly, define $f_2(y^n)$ as the smallest k such that $(y^n, U^n(f_1(y^n)), V^n(k|f_1(y^n))) \in T^n_{[YUV]_\delta}$, and if no such k is found, let $f_2(y^n) = 1$. Set $J_1(m) = f_1(Y^n(m))$ and $J_2(m) = f_2(Y^n(m))$.

Identification: For arbitrary $z^n \in \mathcal{Z}^n$ and $j_1(1), \ldots, j_1(M), j_2(1), \ldots, j_2(M)$, let $g_1^{(M)}(j_1(1), \ldots, j_1(M), z^n)$ be defined as the smallest m such that $(z^n, U^n(j_1(m))) \in T^n_{[ZU]_\delta}$. If no such m is found, set $g_1^{(M)}$ to 1. Similarly, let $g_2^{(M)}(j_1(1), \ldots, j_1(M), j_2(1), \ldots, j_2(M), z^n)$ be defined as the smallest m such that

$$(z^n, U^n(j_1(m)), V^n(j_2(m)|j_1(m))) \in T^n_{[ZUV]_\delta}.$$

If no such m is found, set $g_2^{(M)}$ to 1. Set

$$\hat{W}_1^{(M)} = g_1^{(M)}(J_1(1), \ldots, J_1(M), Z^n)$$

and

$$\hat{W}_2^{(M)} = g_2^{(M)}(J_1(1), \ldots, J_1(M), J_2(1), \ldots, J_2(M), Z^n).$$

Probability of Error: It is clear that since the first-stage system is the same as in the direct part of the proof of Theorem 1, we have $\Pr[\hat{W}_1^{(M)} \neq W] \leq \delta$ for large enough n whenever

$$I(Y;U) + \frac{\delta}{2} \leq \frac{1}{n}\log L_1$$

and

$$\frac{1}{n}\log M \leq R_1^{\mathrm{i}} - \delta.$$

Recall that with such choice of L_1, we also have

$$\Pr\Big[(Y^n(m), U^n(J_1(m))) \notin T^n_{[YU]_\delta}\Big] \leq \frac{\delta}{3} \qquad (16)$$

for $1 \leq m \leq M$ and large enough n. Set

$$\tilde{R}_1^{\mathrm{c}} = I(Y;U)$$

so that L_1 can be chosen to satisfy

$$\frac{1}{n}\log L_1 \leq \tilde{R}_1^{\mathrm{c}} + \delta.$$

For the second stage, define the error events

$$
\begin{aligned}
\mathcal{F}_1(m) &\triangleq \Big\{\Big(Y^n(m), U^n(J_1(m)),\\
&\qquad V^n(J_2(m)|J_1(m))\Big) \notin T^n_{[YUV]_\delta}\Big\}\\
\mathcal{F}_2(m) &\triangleq \Big\{\Big(Z^n, U^n(J_1(m)),\\
&\qquad V^n(J_2(m)|J_1(m))\Big) \notin T^n_{[ZUV]_\delta}\Big\}.
\end{aligned}
$$

The probability of error computed over the ensemble of codebooks can then be bounded as

$$
\begin{aligned}
\Pr[\hat{W}_2^{(M)} \neq W | W = w] \leq{}& \Pr[\mathcal{F}_1(w)] + \Pr[\mathcal{F}_2(w)|\mathcal{F}_1(w)^c]\\
&+ \sum_{m \neq w} \Pr[\mathcal{F}_2(m)^c].
\end{aligned}
$$

Using standard techniques together with (16), it can be shown that $\Pr[\mathcal{F}_1(w)] \leq \delta/3$ for large enough n if

$$I(Y;V|U) + \frac{\delta}{2} \leq \frac{1}{n}\log L_2.$$

In particular, L_2 can be chosen so that

$$
\begin{aligned}
\frac{1}{n}\log L_2 &\leq I(Y;V|U) + \delta\\
&= I(Y;U,V) - \tilde{R}_1^{\mathrm{c}} + \delta\\
&\leq R_2^{\mathrm{c}} - \tilde{R}_1^{\mathrm{c}} + \delta
\end{aligned}
$$

and thus

$$\frac{1}{n}\log L_1 + \frac{1}{n}\log L_2 \leq R_2^{\mathrm{c}} + 2\delta.$$

It also follows from $Z - Y - (U, V)$ that $\Pr[\mathcal{F}_2(w)|\mathcal{F}_1(w)^c] \leq \delta/3$. Finally

$$\sum_{m \neq w} \Pr[\mathcal{F}_2(m)^c] \leq M 2^{-n[I(Z;U,V) - \delta/2]} \leq \frac{\delta}{3}$$

for large enough n whenever

$$\frac{1}{n}\log M \leq I(Z;U,V) - \delta$$

and thus whenever

$$\frac{1}{n}\log M \leq R_2^{\mathrm{i}} - \delta.$$

Using the same arguments as in the end of the proof of Theorem 1 regarding deterministic codebooks, achievability of $(\tilde{R}_1^{\mathrm{c}}, R_2^{\mathrm{c}}, R_1^{\mathrm{i}}, R_2^{\mathrm{i}})$ follows. But since $\tilde{R}_1^{\mathrm{c}} \leq R_1^{\mathrm{c}} \leq R_2^{\mathrm{c}}$, this implies achievability of $(R_1^{\mathrm{c}}, R_2^{\mathrm{c}}, R_1^{\mathrm{i}}, R_2^{\mathrm{i}})$ and thus $\mathcal{R}_s^* \subset \mathcal{R}_s$. To see this, we follow the argument in [5, Lemma 4] pointing to the fact that one can always transfer compressed bits from the second stage to the first one without changing the system performance except for increased first-stage compression rate. The transferred bits can be simply ignored at the first stage. □

B. Achievability Region for N Stages

We next present a single-letter characterization of the N-stage successive achievability region $\mathcal{R}_s$. We omit the proofs because they are straightforward extensions of their two-stage counterparts.

Definition 5: Let $\mathcal{R}_s^*$ be the set of all $(R_1^{\mathrm{c}}, \ldots, R_N^{\mathrm{c}}, R_1^{\mathrm{i}}, \ldots, R_N^{\mathrm{i}})$ such that there exists auxiliary random variables $U_1, U_2, \ldots, U_N$ taking values in alphabets $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_N$ satisfying

$$Z - X - Y - (U_1, \ldots, U_N)$$

and

$$I(Y; U_1, \ldots, U_i) \leq R_i^{\mathrm{c}}$$
$$I(Z; U_1, \ldots, U_i) \geq R_i^{\mathrm{i}}$$

for all $1 \leq i \leq N$.

Lemma 3: $\mathcal{R}_s^*$ is convex. Moreover, in determining $\mathcal{R}_s^*$, it suffices to focus on $\mathcal{U}_1, \ldots, \mathcal{U}_N$ with

$$|\mathcal{U}_1| = |\mathcal{Y}| + 2N - 1$$

and

$$|\mathcal{U}_i| = |\mathcal{Y}| \cdot \prod_{j=1}^{i-1} |\mathcal{U}_j| + 2(N-i) + 1.$$

Theorem 3: $\mathcal{R}_s = \mathcal{R}_s^*$.

C. Two-Stage Systems Without Capacity Loss

Paralleling the notion of successive refinability without rate loss in the classical rate–distortion sense [3]–[5], we introduce here *successive refinability without capacity loss*.

Definition 6: A source (X, Y, Z) with statistics $P_{XYZ} = P_X P_{Y|X} P_{Z|X}$ is said to be successively refinable without capacity loss if

$$(R_1^{\mathrm{c}}, R_2^{\mathrm{c}}, C(R_1^{\mathrm{c}}), C(R_2^{\mathrm{c}})) \in \mathcal{R}_s. \tag{17}$$

The next lemma presents necessary and sufficient conditions for (17) to hold.

Lemma 4: Successive refinement without capacity loss is possible at $(R_1^{\mathrm{c}}, R_2^{\mathrm{c}})$ if and only if there exists optimal test channels $P_{U_1|Y}^*$ and $P_{U_2|Y}^*$ achieving $C(R_1^{\mathrm{c}})$ and $C(R_2^{\mathrm{c}})$, respectively, such that $Y - U_2^* - U_1^*$ forms a Markov chain.

Proof: It is immediate from Theorems 1 and 2 that a source is successively refinable without capacity loss if and only if there exist (U_1, U_2) such that

$$I(Y; U_1) \leq R_1^{\mathrm{c}} \tag{18}$$
$$I(Y; U_1, U_2) \leq R_2^{\mathrm{c}} \tag{19}$$
$$I(Z; U_1) = C(R_1^{\mathrm{c}}) \tag{20}$$
$$I(Z; U_1, U_2) = C(R_2^{\mathrm{c}}). \tag{21}$$

Now, if there exist optimal test channels satisfying $Y - U_2^* - U_1^*$, then (18)–(21) are automatically satisfied by $U_1 = U_1^*$ and $U_2 = U_2^*$, thus proving the sufficiency of Markovity.

On the other hand, if (18)–(21) are satisfied by some (U_1, U_2), then letting

$$U_2' = (U_1, U_2)$$

we not only obtain a Markov chain $Y - U_2' - U_1$, but it follows from (19) and (21) that $P_{U_2'|Y}$ is an optimal test channel. Similarly, (18) and (20) imply optimality of $P_{U_1|Y}$. This proves necessity of Markovity. □

Remark 2: The necessity part of the proof shows us that, to form the Markov chain, one may have to use a test channel $P_{U_2|Y}^*$ with an alphabet of size much larger than $|\mathcal{U}_2| = |\mathcal{Y}| + 1$. More specifically, in the worst case, it follows from Lemma 2 that $|\mathcal{U}_2|$ may have to be as large as $(|\mathcal{Y} + 3|)^2|\mathcal{Y}| + |\mathcal{Y}| + 3$.

Let us reconsider the examples in Section III-A, and show that in both cases, no capacity loss is incurred on the multistage system. Recall that in both examples, $P_{U_i|Y}^*(R_i^{\mathrm{c}})$ for $i = 1, 2$ are binary symmetric test channels with crossover probabilities $0 \leq q_i \leq 1/2$, where q_i is determined as the solution of

$$1 - \mathcal{H}(q_i) = R_i^{\mathrm{c}}.$$

Now, $R_1^{\mathrm{c}} \leq R_2^{\mathrm{c}}$ implies $q_1 \geq q_2$. Thus, it is possible to find $1/2 \geq \alpha \geq 0$ such that

$$q_1 = q_2 \star \alpha.$$

This, in turn, implies that $Y - U_2^* - U_1^*$ is satisfied.

V. Conclusion

In this paper, extending the work in [14], we investigated the fundamental tradeoff between the rate of the number of data entries that can reliably be identified and the rate of required storage (or equivalently, search complexity). We then introduced a further extension, whereby data is compressed in multiple stages before enrollment, and during identification, the compressed bits are retrieved up to the stage where the achieved capacity is larger than the rate of the number of entries. In both cases, a single-letter characterization for the whole rate region is provided. Paralleling the phenomenon in multiresolution rate–distortion theory known as successive refinement without rate loss, we also introduced the notion of successive refinement without capacity loss, and proved that Markovity of optimal test channels achieving individual capacity/storage tradeoffs is a necessary and sufficient condition.

APPENDIX

Lemma 5: The curve on the $(R^{\rm c}, R^{\rm i})$-plane parameterized by

$$
\begin{aligned}
R^{\rm c}(q) &= 1 - \mathcal{H}(q)\\
R^{\rm i}(q) &= 1 - \mathcal{H}(\gamma \star q)
\end{aligned}
$$

for $0 \leq q \leq 1/2$ is concave.

Proof: We will show $d^2R^{\rm i}/(dR^{\rm c})^2 \leq 0$ for $0 < q < 1/2$. Concavity will then follow from the continuity of both $R^{\rm c}(q)$ and $R^{\rm i}(q)$ at $q = 0$ and $q = 1/2$. Define

$$
f(q) \triangleq \frac{\log \frac{1-\gamma\star q}{\gamma\star q}}{\log \frac{1-q}{q}}
$$

and observe

$$
\frac{dR^{\rm i}}{dR^{\rm c}} = \frac{\frac{dR^{\rm i}}{dq}}{\frac{dR^{\rm c}}{dq}} = (1-2\gamma)f(q),
$$

and thus

$$
\frac{d^2R^{\rm i}}{(dR^{\rm c})^2} = (1-2\gamma)\frac{\frac{df}{dq}}{\frac{dR^{\rm c}}{dq}}.
$$

Since $R^{\rm c}(q)$ is monotonically decreasing in q within the interval $(0, 1/2)$, it suffices to show that $df/dq \geq 0$ everywhere in the same interval.

Now

$$
\frac{df}{dq} = \frac{\log e}{\left(\log\frac{1-q}{q}\right)^2} \cdot \left[\frac{-(1-2\gamma)}{(1-\gamma\star q)(\gamma\star q)}\log\frac{1-q}{q} + \frac{1}{(1-q)q}\log\frac{1-\gamma\star q}{\gamma\star q}\right]
$$

and hence

$$
\frac{df}{dq} \geq 0 \Longleftrightarrow f(q) \geq \frac{(1-2\gamma)(1-q)q}{(1-\gamma\star q)(\gamma\star q)} \triangleq g(q). \tag{22}
$$

In other words, $f(q)$ is monotonically decreasing in an open interval (q_1, q_2) if and only if $f(q) < g(q)$ for all $q \in (q_1, q_2)$. On the other hand

$$
\begin{aligned}
\frac{dg}{dq} &= \frac{1-2\gamma}{(1-\gamma\star q)^2(\gamma\star q)^2}\Big[(1-2q)(1-\gamma\star q)(\gamma\star q)\\
&\qquad - (1-2\gamma)(1-2(\gamma\star q))(1-q)q\Big]\\
&= \frac{\gamma(1-2\gamma)(1-\gamma)(1-2q)}{(1-\gamma\star q)^2(\gamma\star q)^2} \geq 0
\end{aligned} \tag{23}
$$

and, therefore, $g(q)$ is monotonically nondecreasing in $q \in (0, 1/2)$. We next combine these observations, and show that negativity of df/dq at any point results in a contradiction. Towards that end, assume that $df/dq(q_0) < 0$ for some $q_0 \in (0, 1/2)$. Let (q_1, q_2) be the largest open interval containing q_0 such that $df/dq(q) < 0$ for all $q \in (q_1, q_2)$. From (22) and (23), we have

$$
g(q_2) \geq g(q_1) \geq f(q_1) > f(q_2). \tag{24}
$$

Now, if $q_2 = 1/2$, (24) immediately creates a contradiction because $f(1/2) = g(1/2) = (1-2\gamma)$. On the other hand, if $q_2 < 1/2$, (24) is also contradictory since we must then have $df/dq(q_2) \geq 0$, and therefore $f(q_2) \geq g(q_2)$ according to (22). □

ACKNOWLEDGMENT

The author is grateful to Jun Chen for pointing him to the necessity of the Markov condition for successive refinement without capacity loss.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1982.

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

[4] V. N. Koshelev, "Hierarchical coding of discrete sources," *Probl. Pered. Inform.*, vol. 16, no. 3, pp. 31–49, 1980.

[5] B. Rimoldi, "Successive refinement of information: Characterization of the achievable rates," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 253–259, Jan. 1994.

[6] Y. Steinberg and N. Merhav, "On successive refinement for the Wyner–Ziv problem," *IEEE Trans. Inf. Theory*, vol. 50, no. 8, pp. 1636–1654, Aug. 2004.

[7] J. A. O'Sullivan, N. Singla, and M. B. Westover, "Successive refinement for pattern recognition," in *Proc. IEEE Information Theory Workshop*, Punta del Este, Uruguay, Mar. 2006, pp. 141–145.

[8] C. Tian and J. Chen, "Successive refinement for hypothesis testing and lossless one-helper problem," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4666–4681, Oct. 2008.

[9] N. Tishby, F. C. Pereira, and W. Bialeck, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Communication, Control and Computing*, Montecillo, IL, Sep. 1999, pp. 368–377.

[10] E. Tuncel, P. Koulgi, and K. Rose, "Rate-distortion approach to databases: Storage and content-based retrieval," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 953–967, Jun. 2004.

[11] R. Weber, H. J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces," in *Proc. Int. Conf. Very Large Data Bases*, New York, NY, Aug. 1998, pp. 194–205.

[12] M. B. Westover and J. A. O'Sullivan, "Towards an information theoretic framework for object recognition," in *Proc. IEEE Int. Symp. Information Theory*, Chicago, IL, Jun./Jul. 2004, p. 219.

[13] M. B. Westover and J. A. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 299–320, Jan. 2008.

[14] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, "On the capacity of a biometrical identification system," in *Proc. IEEE Int. Symp. Information Theory*, Yokohama, Japan, Jun./Jul. 2003, p. 82.

[15] A. D. Wyner, "A theorem on the entropy of certain binary sequences and application: II," *IEEE Trans. Inf. Theory*, vol. IT-19, no. 6, pp. 772–777, Nov. 1973.

[16] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-22, no. 1, pp. 1–11, Jan. 1976.

Ertem Tuncel (S'99–M'04) received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1995, and the MS degree from Bilkent University, Ankara, Turkey, in 1997, both in electrical engineering. He received the Ph.D. degree in electrical and computer engineering from University of California, Santa Barbara, in 2002.

In July 2003, he joined the Department of Electrical Engineering, University of California, Riverside, as an Assistant Professor. His research interests include rate–distortion theory, multiterminal source coding, joint source–channel coding, zero-error information theory, and content-based retrieval in high dimensional databases.

Prof. Tuncel received the 2007 National Science Foundation CAREER Award.